

LECTURE NOTES IN GEOINFORMATION AND CARTOGRAPHY

LNG&C

Popovich · Schrenk · Korolenko (Eds.)

Information Fusion and Geographic Information Systems



Springer

Lecture Notes in Geoinformation and Cartography

Series Editors: William Cartwright, Georg Gartner, Liqiu Meng,
Michael P. Peterson

Vasily Popovich, Manfred Schrenk,
Kyrill Korolenko,
(Eds.)

Information Fusion and Geographic Information Systems

Proceedings of the Third International Workshop

With 110 Figures



Editors:

Vasily V. Popovich

Dr Sci Tech,
Head of Laboratory of Object Oriented
Geoinformation Systems
39, 14th Linia, V.O.
199178 St. Petersburg, Russia
E-mail: popovich@mail.iias.spb.su

Kyrrill V. Korolenko, P.E.

Chief Scientist / NUWC Code 1543,
B1320
1176 Howell St, Newport, RI 02841-
1708
email: korolenkokv@npt.nuwc.navy.mil

Manfred Schrenk

CEIT ALANOVA gemeinnützige GmbH
Central European Institute of
Technology
Department for Urbanism, Transport,
Environment & Information Society
Am Concorde Park 2, Gebäude F
A-2320 Schwechat
Austria
E-mail: m.schrenk@ceit.at

ISBN 10 3-540-37628-3 Springer Berlin Heidelberg New York

ISBN 13 978-3-540-37628-6 Springer Berlin Heidelberg New York

ISSN 1863-2246

Library of Congress Control Number: 2007925168

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springeronline.com

© Springer-Verlag Berlin Heidelberg 2007

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Cover design: deblik, Berlin

Production: A. Oelschläger

Typesetting: Camera-ready by the Editors

Printed on acid-free paper 30/2132/AO 54321

Preface

This volume contains the papers presented at the International Workshop “Information Fusion and Geographical Information Systems” (IF&GIS’07) held in St. Petersburg, Russia, during May 27-29, 2007. The workshop was organized by the St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS).

Research in the Geosciences field interprets the concept of Information Fusion as a synonym of an approach that permits research into all the problems and issues which program-makers and scientific researchers are faced with. Thus, topics that are to be covered during the workshop relate to issues such as harmonization, integration and information fusion. At the same time the spectrum of problems under discussion exceeds the current bounds of developing GIS applications. This is a significant modern trend since GIS technology is more often used as an interface in support and decision making systems. As a result, it is difficult to consider Geoinformation Science without considering related scientific directions such as Ontology, Artificial Intelligent Systems and Situation Management.

Out of 51 abstracts from 14 countries, 35 papers were selected for submission and from these, the review process of the Program Committee recommended 25 papers from 10 countries for publication, including three invited papers. All the papers that were presented and accepted for publication, were allocated to the following five sessions: Data, Information and Knowledge Harmonization, Integration and Fusion in GIS; Information Assurance and Protection in GIS; GIS as a Basis for Monitoring Systems; Ontologies and Programming Technologies for GIS and GIS Applications; Operations Research and TSMO for GIS Applications. The IF&GIS’07 program was enriched by contributions from three distinguished invited speakers: Gabriel Jakobson, James Llinas and Christophe Claramunt.

The success of the workshop was assured by the team efforts of sponsors, organizers, reviewers and participants. We would like to acknowledge the contribution of the Individual Program Committee members and thank the reviewers for their support and hard work. Our sincere gratitude goes out to all participants and all the authors of submitted papers.

We are grateful to our sponsors, the Russian Academy of Sciences and the US Office of Naval Research Global (ONRGGlobal) for their generous support.

We wish to express our gratitude to Springer’s LNCS team, managed by Dr. Christian Witschel for their help and co-operation.

May 2007

Vasily Popovich
Manfred Schrenk
Kyrill Korolenko

Organization

Workshop Chairmen

General Chairmen

Rafael M. Yusupov

St. Petersburg Institute for Informatics and
Automation, Russia

Program Committee Chairmen

Vasily V. Popovich

St. Petersburg Institute for Informatics and
Automation, Russia

Manfred Schrenk

MULTIMEDIAPLAN.AT, Vienna, Austria

Kyrill Korolenko

NAVSEA, Newport, USA

Program Committee

Adam Rimashevsky	(Adm., Head of the PGS of the Russian Navy).
Alexander Smirnov	(SPIIRAS, St.-Petersburg, Russia)
Anatoly Rodionov	(RAS, St.-Petersburg, Russia)
Bin Jiang	(Division of Geomatics Dept. of Technology and Built Environment University of Gävle, SE-801 76 Gävle, Sweden)
Bo Huang	(Department of Geomatics Engineering University of Calgary, Canada)
Boris Sokolov	(SPIIRAS, St.-Petersburg, Russia)
Christophe Claramunt	(Naval Academy Research Institute Lanveoc-Poulmic, Brest Naval, France)
Gabriel Jacobson	(altusys, Boston, USA)
Gennady Maklakov	(Sevastopol National Technical University, Ukraine)
Grigory Zykov	(Naval Academy, St.-Petersburg, Russia)
Herve Martin	(University, Grenoble, France)
Igor Kotenko	(SPIIRAS, St.-Petersburg, Russia)
Marie-Aude AUFAURE	(Supélec - Plateau du Moulon - Département Informatique, France)
Martin Breunig	(Research Centre for Geoinformatics and Remote Sensing, University of Osnabrueck, Germany)
Michela Bertolotto	(School of Computer Science and Informatics University College Dublin, Ireland)
Mieczyslaw M. Kokar	(Northeastern University, Boston, USA)
Nikolay Moldovyan	(SPECTR, Russia)
Sergei Prisyagnuk	(Institution of Telecommunication, St.-Petersburg, Russia)
Stefan Axberg	(National Defense College, Stockholm, Sweden)
Tao Cheng	(Hong Kong, Polytechnic)
Thierry Badard	(Universite Laval, Quebec, Canada)
Vladimir Vasyukov	(Adm. St. Petersburg, Russia)

Reviewers

Alexander Smirnov	(SPIIRAS, St.-Petersburg, Russia)
Anatoly Rodionov	(RAS, St.-Petersburg, Russia)
Bin Jiang	(Division of Geomatics Dept. of Technology and Built Environment University of Gävle, SE-801 76 Gävle, Sweden)
Bo Huang	(Department of Geomatics Engineering University of Calgary, Canada)
Boris Sokolov	(SPIIRAS, St.-Petersburg, Russia)
Christophe Claramunt	(Naval Academy Research Institute Lanveoc-Poulmic, Brest Naval, France)
Gabriel Jacobson	(altusys, Boston, USA)
Gennady Maklakov	(Sevastopol National Technical University, Ukraine)
Grigory Zykow	(Naval Academy, St.-Petersburg, Russia)
Herve Martin	(University, Grenoble, France)
Igor Kotenko	(SPIIRAS, St.-Petersburg, Russia)
Manfred Schrenk	(MULTIMEDIAPLAN.AT and CEIT ALANOVA)
Marie-Aude AUFAURE	(Supélec - Plateau du Moulon - Département Informatique, France)
Martin Breunig	(Research Centre for Geoinformatics and Remote Sensing, University of Osnabrueck, Germany)
Michela Bertolotto	(School of Computer Science and Informatics University College Dublin, Ireland)
Mieczyslaw M. Kokar	(Northeastern University, Boston, USA)
Nikolay Moldovyan	(SPECTR, Russia)
Sergei Prisyagnuk	(Institution of Telecommunication, St.-Petersburg, Russia)
Stefan Axberg	(National Defense College, Stockholm, Sweden)
Tao Cheng	(Hong Kong, Polytechnic)
Thierry Badard	(Université Laval, Quebec, Canada)

Table of Contents

Invited Papers

- New Challenges for Defining Information Fusion Requirements
Dr. James Llinas 1

- Situation Management: Basic Concepts and Approaches
Gabriel Jakobson, John Buford, Lundy Lewis 18

- Maritime GIS: From Monitoring to Simulation Systems
C. Claramunt, T. Devogele, S. Fournier, V. Noyon, M. Petit, C. Ray .. 34

Data, Information and Knowledge Harmonization, Integration and Fusion in GIS

- Intelligent Images Analysis in GIS
Philipp Galjano and Vasily Popovich 45

- Context-Driven Information Fusion for Operational Decision Making in Humanitarian Logistics
Alexander Smirnov, Alexey Kashevnik, Tatiana Levashova, Nikolay Shilov 69

- From Battle Management Language (BML) to Automatic Information Fusion
Ulrich Schade, Joachim Biermann, Miloslaw Frey, Kellyn Kruger 84

- “Centrope MAP”: Combining X-border Information from Multiple Sources for Planning Purposes
Manfred Schrenk, Walter Pozare 96

Information Assurance and Protection in GIS

- Software Environment for Simulation and Evaluation of a Security Operation Center
Julien Bourgeois, Abdoul Karim Ganame, Igor Kotenko and Alexander Ulanov 111

Security Policy Verification Tool for Geographical Information Systems <i>Igor Kotenko, Artem Tishkov, Olga Chervatuk and Ekaterina Sidelnikova</i>	128
Architecture Types of the Bit Permutation Instruction for General Purpose Processors <i>Moldovyan A. A., Moldovyan N. A., Moldovyanu P. A.</i>	147
New Public Key Cryptosystems Based on Difficulty of Factorization and Discrete Logarithm Problems <i>Moldovyan N. A.</i>	160
GIS as a Basis for Monitoring Systems	
Application of a Dynamic Recurrent Neural Network in Spatio-Temporal Forecasting <i>Tao Cheng and Jiaqiu Wang</i>	173
System of Traffic Control on the Basis of Cartographic Databases and Geoinformation Technologies <i>A.A.Kravtsov, A.N.Kriuchkov, E.E.Sotikova.....</i>	187
Information Fusion of Land Laser Scanning for Geographic Information Systems <i>Ilya S. Tarasov and Nikita A.Pikhtin</i>	194
Using GIS to Analyze Acute Myocardial Infarction in Turkey <i>Mehmet Zeki Coskun, H. Can Ünen, Cevat Kirma , Ercument Yilmaz202</i>	
Ontologies and Programming Technologies for GIS and GIS Applications	
Obtaining Semantic Descriptions Based on Conceptual Schemas Embedded into a Geographic Context <i>Miguel Torres and Serguei Levachkine</i>	209
Representing the Semantic Content of Topological Relations into Spatial Databases <i>Miguel Martinez, Marco Moreno, Miguel Torres and Serguei Levachkine</i>	223

Some Remarks on Topological Abstraction in Multi Representation Databases <i>Andreas Thomsen and Martin Breunig.....</i>	234
Intelligent Simulation of Hydrophysical Fields by Immunocomputing <i>Alexander Tarakanov, Ludmilla Sokolova and Sergey Kvachev.....</i>	252
Visual Simulation of Spatial Processes <i>R.P.Sorokin.....</i>	263
Universal Model of Data for Intellectual GIS <i>Sergey N. Potapychев, Andrey V. Pan'kin.....</i>	272
3D Graphics Applied to Maritime Safety <i>Christopher Gold and Rafal Goralski</i>	286
Operations Research and TSMO for GIS Applications	
Northern Shield US-Russia Maritime Energy Security Cooperation <i>Mr. Gabe Collins</i>	301
Space-Extensive Targets Radar Search in the Presence of Noise <i>I. F. Shishkin, A. G. Sergushev</i>	316
Empirical Bayes Trajectory Estimation on the Base of Bearings from Moving Observer <i>A. Makshanov, A. Prokaev.....</i>	323
Author Index	335

New Challenges for Defining Information Fusion Requirements

Dr. James Llinas

Professor, Executive Director
Center for Multisource Information Fusion
State University of New York at Buffalo
Buffalo, New York, USA
llinas@eng.buffalo.edu

Abstract. The changing geopolitical landscape in the world has been and will continue to be the driving framework within which requirements for new and adaptable capabilities in Information Fusion (IF) technology are defined. For major nation-states of the world, this changing landscape will, it is argued, generate new challenges that significantly broaden both the range and adaptive nature of the capability that future IF systems must have. Another dramatically changing landscape is that of information networking, and the integration and exploitation of such networking in military and defense operations have led to transformations in military thinking and culture, even to the consideration of radically new socio-organizational dynamics for Command and Control (C2). Further, the need to develop deeper insights into agile and creative adversarial behaviors imparts what is called here the need for a “multi-perspective” Information Fusion process that will require new ways to think about exploiting both traditional and novel Intelligence, Surveillance, and Reconnaissance (ISR) sources. Further, there is the impact of informational dimensionality via the need, similarly motivated, to fuse and exploit the “PMESII” (Political, Military, Economic, Social, Infrastructure, Information) spectrum of information. Lastly, there is the desire on the part of the military to focus on “Effects-based” operations; here too there is an impact of new requirements onto the Data Fusion process.

These extensive changes in both the application context and the technological foundations for IF have far-reaching implications for the both the architectural design of IF processes as well as the foundational algorithms employed in IF systems. Significant challenges exist toward achieving robustness and scalability of IF capabilities, the role of and support to human involvement on the IF process, and the ability of IF systems to estimate not only states in the physical domain but also in the informational and cognitive domains. This paper and presentation will survey this extensive new and changing landscape as regards the impacts on IF requirements, with some thoughts on new strategies for IF process design.

1 Introduction—Conflict in the World: Brief Thoughts

Not being a student of international affairs and politics, I hesitate to go deeply into the nature of conflict in the world, but as nation-state and irregular group policies drive international defense policies, and thus requirements for defense-related technologies, these matters become the prime factors that ultimately influence the direction of research and development of international defense technology thrusts. Conflict in the world seems to be the norm rather than the exception; in [1] it is said that “Researchers have calculated that since 3600 BC, there have been 14,531 wars resulting in 3 billion deaths over the years (current world population: 5 billion). Peace prevailed for a total of only 292 years on earth, about 5% of the time”. An assessment from [2] shows the following summary of ongoing armed conflicts in the world:

Significant Ongoing Armed Conflicts, 2006

Main warring parties	Year began ¹
Middle East	
U.S. and UK vs. Iraq	2003
Israel vs. Palestinians	1948
Asia	
Afghanistan: U.S., UK, and Coalition Forces vs. al-Qaeda and Taliban	2001
India vs. Kashmiri separatist groups/Pakistan	1948
India vs. Assam insurgents (various)	1979
Indonesia vs. Christians and Muslims in Molucca Islands	1977
Indonesia vs. Papua (Irian Jaya) separatists	1969
Nepal vs. Maoist rebels	1995
Philippines vs. Mindanaoan separatists (MILF/ASG)	1971

Sri Lanka vs. Tamil Eelam2	1978
Africa	
Algeria vs. Armed Islamic Group (GIA)	1991
Côte d'Ivoire vs. rebels	2002
Democratic Republic of Congo and allies vs. Rwanda, Uganda, and indigenous rebels	1997
Somalia vs. rival clans and Islamist groups	1991
Sudan vs. Darfur rebel groups	2003
Uganda vs. Lord's Resistance Army (LRA)	1986
Europe	
Russia vs. Chechen separatists	1994
Latin America	
Colombia vs. National Liberation Army (ELN)	1978
Colombia vs. Revolutionary Armed Forces of Colombia (FARC)	1978
Colombia vs. Autodefensas Unidas de Colombia (AUC)	1990

Many of the problems listed above can be labeled as insurgencies (or “low-intensity warfare”, and other terms), and although several of the major powers of the world have been focused on counter-insurgency and the international counter-terrorism conflict, there is still a significant potential for large-scale and/or regional conventional warfare. Michael Mazaar, who has written extensively on world conflict says, in [3], “To say, moreover, that all war is now small war, that state-to-state conflict has given way to Fourth Generation Warfare, generates an obvious blind spot for the traditional, state-on-state wars that without doubt remain possible. If China were to attack Taiwan,.....”. He argues further that “No serious observer of world politics denies, despite all the trends toward free trade, democracy, and interdependence, that major states could still go to war; but ac-

cepting that truism is to place a very large neon asterisk next to theories that claim our future is nothing but counterinsurgency”. In [3] Mazaar also offers some definitional distinctions, about the “character of battle”, this being the engagement “mechanics” and largely argued as unchanging, then he mentions the “form of warfare”, involving “the tactics and operational art governing units in battle — infantry war versus *blitzkrieg*, insurgency versus classical force-on-force duels. Whereas the character of battle may be eternal, the form of warfare constantly evolves, responding to new technologies, new tactics, and new social organizations”. He then goes on to discuss what he argues is the real fundamental root of conflict which is the *nature of conflict*—“This is the highest strategic level of analysis and deals with the causes and character of severe political-military-socioeconomic disputes in the international system. International conflict generates the context for warfare, but also much else — Schellingesque bargaining games, coercive diplomacy, deception and artful dodges short of warfare and battle.” So then, if we take these assertions as valid, it is really the “form of warfare” that drives the need for new technologies, and those technologies must fit properly into both the evolving tactics and social dynamics of new organizations. But as Mazaar points out, it is the underlying changes in the nature of conflict, and the “why” of conflict that push the changes in the forms of warfare and the consequent technological evolution. Since the forms of warfare can change for any given nature of conflict, depending on why that conflict exists, there will be a need for considerable flexibility in providing the technologies that support varying forms of warfare—as Information Fusion is a part of that technological inventory, it too must provide considerable flexibility across a significantly-wide “problem space” of conflict.

It appears that the United States defense community is aware of and trying to deal with these perspectives. In the latest “Quadrennial Defense Review (QDR)” [4], defense planners have partitioned the space of conflicts into four categories of “challenges” as shown in Figure 1.

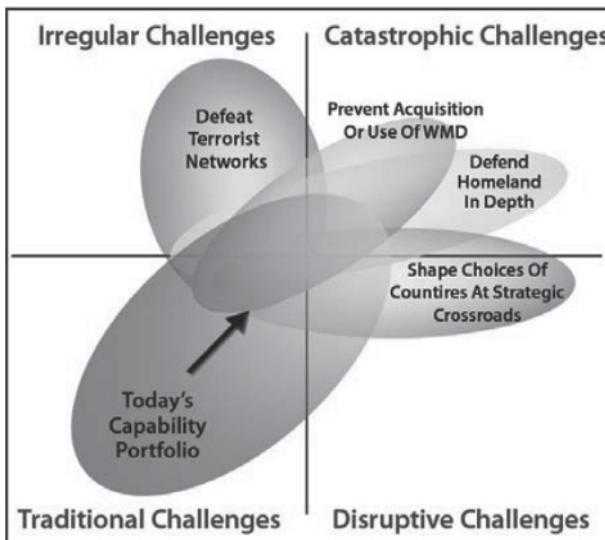


Fig. 1. Challenge categories addressed in the US 2006 QDR [4]

2 Challenges to the Information Fusion Community

These challenges or categories of conflict require different “forms of warfare” as defined by Mazaar, meaning different tactics and social organizations, and in turn new technological capabilities for each category. In the “Irregular” category, the QDR discusses:

- HUMINT to aid in Intent estimation
 - Persistent surveillance
 - Locate, tag, track terrorists including cyberspace
 - Fusion frameworks for combining intelligence and operations
 - Cultural adaptability of the semantic foundations of the IF process
 - Support to non-lethal weaponry
 - Support to urban warfare operations
 - Interoperability between military and local/regional civilian C2/IF systems
 - Support to joint C2 centers as Emergency Operations Centers
- In the “Catastrophic” category, the QDR cites:

- Joint command and control for homeland defense and civil support missions, including communications and command and control systems that are interoperable with other agencies and state and local governments.
- Air and maritime domain awareness capabilities to provide increased situational awareness and shared information on potential threats through rapid collection, fusion and analysis.
- Capabilities to manage the consequences of major catastrophic events.

In the “Disruptive” category we have:

- Considerably improved language and cultural awareness to develop a greater understanding of emerging powers and how they may approach strategic choices.
- Persistent surveillance, including systems that can penetrate and loiter in denied or contested areas.
- The capability to deploy rapidly, assemble, command, project, reconstitute, and re-employ joint combat power from all domains to facilitate assured access.
- Prompt and high-volume global strike to deter aggression or coercion, and if deterrence fails, to provide a broader range of conventional response options to the President. This will require broader authorities from the Congress.
- Secure broadband communications into denied or contested areas to support penetrating surveillance and strike systems.
- Integrated defenses against short-, intermediate-, and intercontinental-range ballistic and cruise missile systems.
- Air dominance capabilities to defeat advanced threats.
- Undersea warfare capabilities to exploit stealth and enhance deterrence.
- Capabilities to shape and defend cyberspace.
- Joint command and control capabilities that are survivable in the face of WMD-, electronic-, or cyber-attacks.

This is of course a huge list of recommended adaptations, and the ability to invoke these changes involves a complex planning process that needs to prioritize these lists and decide what can be done in a timeframe over which the underlying nature of international conflict remains reasonably steady, else the pace of development will be out of synchronization with the pace of change in type of conflict.

In this paper, we address five categories of new requirements for the Information Fusion community that can be correlated to many of the items on the above lists. These categories are:

1. Challenges of Robustness and Scalability to extend the range of operation of IF processes
2. IF process designs that will operate effectively and efficiently in the frameworks of Network-centric Warfare/Network-enabled Capability
3. IF processes that support Effects-based Operations
4. IF processes that are “multi-perspective” and support the Operational Net Assessment concept
5. IF processes that can absorb a much broader range of the information spectrum

3 Discussing the New Requirements

3.1 Robustness and Scalability

The natural evolution of capabilities for a given algorithmic paradigm typically extends these capabilities within a bounded genre of problem classes. This notion of extensibility within a problem genre we call “Scalability”. A good example of such an evolutionary development process is the growth of the Kalman Filter paradigm that, over time, has been extended to allow for a broad range of special cases within the object-tracking problem class. Some of these extensions have been difficult however, as in the extension from the tracking of air and space objects to the tracking of ground-moving objects. However, no organized research has occurred that has attempted to understand the parametric boundaries (i.e., the boundaries of various independent variables) of each sub-class or, equivalently, the similarities in sub-class parameters, and the associated types of performance that modified algorithms have yielded. Said otherwise, no characterizations of the “performance space” of algorithm-sets designed for problem-space sub-classes have been collected and analyzed in an orderly way. The notional analytic procedure for doing this was analyzed by Chong in 2001 [5] for tracking algorithms, where he employed the notion of “context metrics” as an efficient framework to assess algorithm performance. Context metrics represent the complexity of the problem that affects algorithm performance, and can be viewed conceptually as equivalent classes on the problem parameters. The strategy involves partitioning the feasible values of the problem-affecting parameters (independ-

ent variables) into equivalence classes such that all problem parameter values in the same equivalence class produce the same values for the performance metrics. The context metrics basically represent the equivalent classes, i.e., if the problem parameters for a set of problems map to the same values of the context metrics, then their performance metrics will have the same values. Another way to describe this is that we seek a mapping of sets of independent variables such that for that mapped set the range of performance of the algorithm is fixed.

In contrast, our definition of “Robustness” is an ability to understand how an algorithmic technique for a given problem domain can be extended to an entirely different problem domain. Our definitions may not pass rigorous ontological tests etc but are considered rational enough for discussion purposes. The notion is only slightly different than that offered for Scalability but here the independent variables are those that distinguish *semantically-different* problem domains but for which the subject algorithm’s functionality is appropriate. (We focus on the semantic aspects to avoid the circularity if the mathematical definition of domain is used, where “The term domain is most commonly used to describe the set of values D for which a function (map, transformation, etc.) is defined”, [6]). Roughly, our definition bears similarity to the notion of robustness in genetics where an organism’s capability remains constant in spite of mutations in its operating space; it is a sense of resilience in the face of radical semantic changes. Understanding the potential for Robustness of an algorithm requires understanding the semantic differences in the domain labels and the component models or sub-functions employed in the algorithm. If the semantic differences can be understood and a mapping made between the domain semantics, and if the sub-models in each can be understood, then the viability of applying that algorithm to the new domain can be properly assessed. The major challenge here is negotiating the semantic differences.

One way in which various communities have sought Scalability is to employ and manage multiple algorithms over their individual operating ranges, i.e., to employ an “intelligently-managed” set of multiple algorithms. This requires first knowing which algorithms are truly distinct in their performance as characterized by the domain’s set of context metrics (since multiple algorithms exhibit similar performance in a given space of context metrics). Selecting one algorithm from this set could be based on other than context metrics such as staff familiarity or computing speed, etc. Secondly, the nature of the performance overlap needs to be understood and compared against the overall performance specifications. The algorithm manager logic for switching the algorithms (i.e., terminating one algorithm and invoking the other) will be based on the nature of the overlap.

A diagrammatic characterization of this switching process is shown in Figure 2.

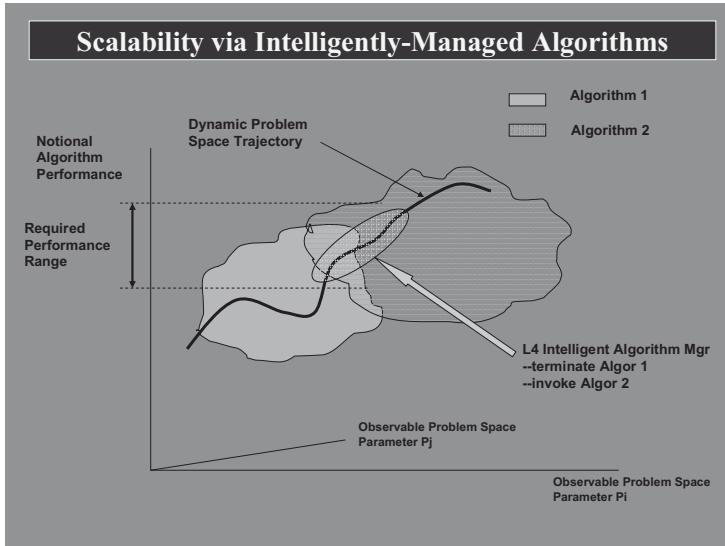


Fig. 2. Notional characterization of intelligently-managed multiple algorithms

Here we see that there are two algorithms whose joint performance spans the range of the specified or desired performance (or, alternately, that neither algorithm alone can satisfy the needed performance). The algorithms however have a region of performance overlap; the detailed nature of that overlap will govern how the algorithm manager switches the algorithms as the problem enters the “switching space” of context metrics.

In addition to the switching logic, Scalable systems will need a host architecture that enables such multi-algorithm operations to take place. Some research has been done that has offered architectural ideas that are hospitable to such ideas. One of these is the “Extensible Data Architecture” or XDA architecture studied in the USA at the Air Force Laboratory’s Information Directorate [7]. XDA was an extension to a prior program called Adaptive Data Fusion (ASF) that extended the ASF capability to include a simple ontology mechanism that enables the definition and maintenance of high-level object models to capture the shared semantics necessary for interoperability.

A future goal in this area as regards achieving Scalability is to eventually design and build autonomic Information Fusion systems that have an

ability to self-configure; research has been going on in autonomic systems, e.g., [8], whose ideas can be imported for IF process design.

3.2 Fusion Process Designs and Network-centric Warfare/Network-enabled Capability

Achieving persistent surveillance, combining intelligence and operations, achieving interoperability between military and local/regional civilian C2/IF systems, wide-area communications and in general better force structure integration are among several of the QDR recommendations that can be addressed in whole or in part by the realization of a highly networked information infrastructure as characterized by the Network-Centric Warfare (NCW) vision.

In a very top-level sense, the NCW vision involves a few major components—the most important of course being the people involved, structured in what NCW calls “Communities of Interest (COI’s)”, the middleware core that supports so-called “Enterprise Services”¹, and the functional software Services, of which Information Fusion Services are a part. This top-level concept is shown in Figure 3:

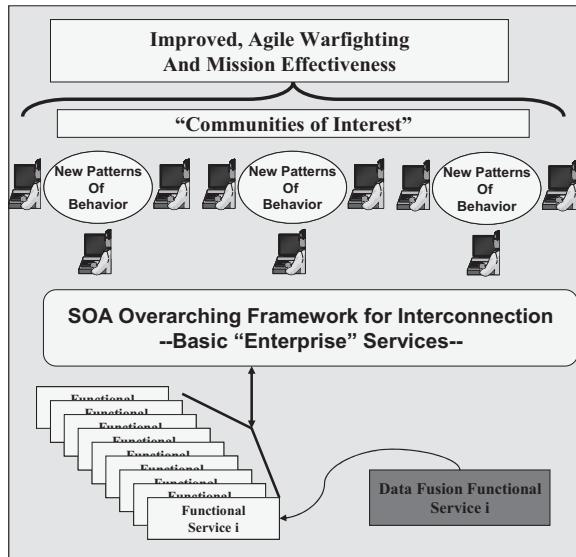


Fig. 3. Notional Top-level Network Centric Warfare Architecture (IT Architecture)

¹ In the USA these are called “Net-Centric Enterprise Services” or NCES

The “payoff” argument for NCW revolves around what NCW architects call the “Value Chain” of NCW (see [9]), in which the ultimate payoff of NCW is in enabling much more rapid “Sensemaking” (situational awareness and understanding), which when combined with new ways of collaborative decision-making allow for rapid, creative problem-solving for the new conflict challenges discussed in Section 1. In Fig. 3, we see these ideas, shown as new patterns of behavior among and within the COI’s, enabled by the NCES middleware (and of course the communications/data-linking backbones not shown in Fig 3) which, from a computer-science point of view is designed as a “Service-Oriented Architecture” or SOA as noted in Fig.3. All functional services such as Information Fusion (IF) tie into the NCES/SOA as are “discovered” by the users in the COI and employed for their purposes.

At this very high level, the key challenges for IF process designers are two-fold:

1. to understand the new information requirements of those users in the COI’s who will be employing new paradigms for both understanding and decision-making, and
2. to understand the technical details of designing IF processes which can function efficiently and effectively with an SOA-based middleware of Enterprise Services.

Number (1) above will not be easy to do. The paradigms being discussed within the NCW community for dealing with the most complex conflicts are touted as “Complex Adaptive Systems (CAS)”, wherein the COI’s, as social organizations, will be exhibiting revolutionary new dynamics involving self-organization, self-synchronization, and semi-autonomous decision-making and action. This is consistent with the “Power to the Edge” principles of NCW (see [10]) wherein power or decision-making authority is distributed to the lower or “edge” elements of the organization. These new organizational dynamics will almost certainly require new types of fused information products from the IF processes or services; e.g., various writings on CAS describe that such organizational dynamics do not seek optimality but work in a fast-paced exploration mode, looking not for optimal solutions but creative, satisficing solutions (e.g., [11]). This type of analytical dynamic will mean that new IF algorithms will have to be researched and developed that service this new user paradigm.

It is believed that Number (2) will be somewhat easier to achieve but not easy. Designing IF processes to function in an SOA architecture will require insight into both the computer-science details affecting such designs but also into (again) the workflow processes of the people in the COI’s. For example, one challenging aspect in this regard will be to de-

sign “dynamically-composable fusion services”. This is the idea of allowing users to “compose” the IF service in a fashion that best satisfies their local needs. Achieving this, as might be expected, requires good insight into the range of flexibility (range of compositions) that users will require or desire.

3.3 IF Processes that Support Effects-based Operations (EBO)

Effects-based operations (EBO) are defined in [12] as “operations conceived and planned in a systems framework that considers the full range of direct, indirect, and cascading effects—effects that may, with different degrees of probability, be achieved by the application of military, diplomatic, psychological, and economic instruments.” Said otherwise, EBO is a kind of systems thinking at the mission planning and execution level, not at the physical target level. A healthy and objective view of EBO is as an amplification of the scope of operations, not as “different” than traditional operations. Davis, in [12], has offered the following taxonomy of “effects”:

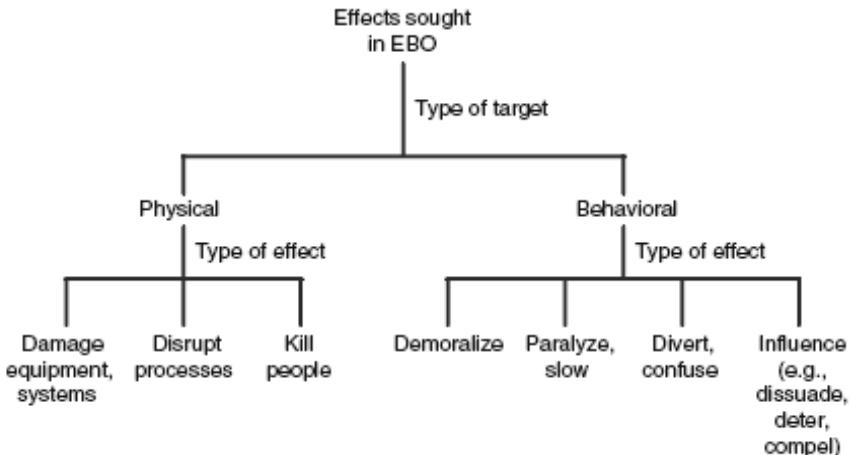


Fig. 4. Taxonomy of Effects (after [12])

This taxonomy distinguishes EBO by the addition of behavioral effects to the more traditional physical effects. Davis points out that it is important to understand that in seeking to cause certain a priori defined effects, that there can be unintended consequences (he cites a number of cases where this has happened) due largely to the assertion that most types of warfare behave as a Complex Adaptive System and have unpredictable aspects in spite of the best modeling of what certain actions may yield as ef-

fects. Therefore, the notion of modeling “cause and effect” as an aid to planning of EBO activities must be viewed carefully, and in that analysis it needs to be realized that the heart of CAS is nonlinearity, and that the action-effect coupling at some low level in a hierarchy can yield destabilizing consequences. Notionally, one seeks a dependency chain as shown in Figure 5. Figure 5 shows an example dependency tree that links the execution of certain tasks T with effects E via “causal links”—this type analysis needs to be interpreted and used very carefully.

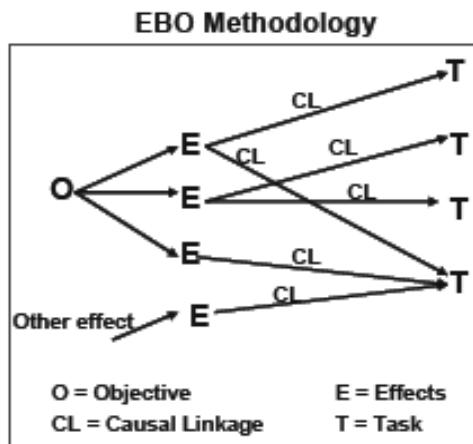


Fig. 5. Notional Dependency Tree for EBO (after [12])

One impact to IF processes from EBO is to emphasize the importance of the multidimensional aspect of the constructs of the “higher-level” fusion estimates, i.e., the estimates of situations and impacts. The fusion community has always considered situational states to comprise physical entities, events and behaviors, so the “behavioral” side of EBO is not an entirely new influence to IF process design. However, note from Fig. 4 that many of the behavioral effects shown have a psychological side to them (demoralize, influence, confuse, etc), and estimating the degree to which these effects have been achieved is an extension of the notions of situational and impact estimation as studied so far in the IF community. Ideally, IF estimates would also enable the estimates of “damage assessment” to the effects created, to close the loop back to planning of the next phase of tasking.

3.4 “Multi-perspective IF” and Support the Operational Net Assessment Concept

It has been argued that the failure of intelligence analysis in Iraq was the result of being locked in a mindset. Given that the purpose of intelligence analysis is to “know your enemy”, such analyses involve considerable insight and, at the same time, open mindedness. In these days of terrorism, with terrorists necessarily operating in an asymmetric mode, trying to perceive and understand their cleverness, deceptiveness, and subtleties requires deeper than normal levels of analysis. In the USA, the Joint Forces Command (JFCOM) has characterized a process called “Operational Net Assessment” (ONA; see [13]), that defines a need for an IF process that we would call “multi-perspective”. In the ONA process, the model calls for an ability to develop estimates not only of the traditional “Blue’s view of Red” (i.e., Friendlies view of the Adversary), but also:

- Blue’s view of Blue—meaning Blue’s self-awareness, e.g., of friendly-force dispositions, etc
- Blue’s view of Red’s view of Blue—ie the friendly estimate of the adversary’s view of us—what we think he knows about us
- Blue’s view of Red’s estimate of Red—the friendly view of Red’s self-awareness

All of this is to be developed from information sources and sensors available only to the friendly or Blue force. A diagrammatic depiction of this process is shown below:

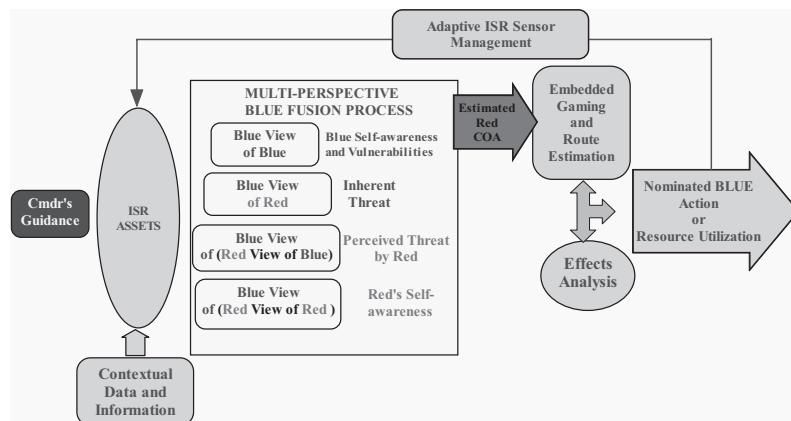


Fig. 6. Notional Characterization of the Operational Net Assessment Process

Another important component of ONA is an internal, embedded, fast-executing wargaming capability that takes the initially-estimated Red Course of Action (COA), generates viable Blue optional COA's, and wargames them to iterate on the preferred Blue option. This wargaming capability of course requires deep insight to the adversary but exploits the multiple views generated by the “multi-perspective” data fusion process. Note that the ONA process, in its wargaming and general approach, could also be Effects-oriented. The culmination of this process is a Blue COA or the employment of a Blue resource. As for all IF processes, there is a feedback to the ISR process to enable an ongoing verification of the correctness of the fusion estimates and of the recommended actions.

The notion of multiple perspective IF can be further extended. Vane and others [14] have described a need to not only focus the IF estimation process on physical objects but also on informational targets and cognitive targets. Of course cognitive targets have always been part of military philosophy (e.g., the principle of overcoming the adversary’s will to fight) but here the principle is overtly integrated into the IF process as a means to provide the friendly commander the information needed to do this efficiently. These ideas considerably extend the dimensionality of the fused state vector and no doubt have significant impacts on computational complexity as well. These state elements are interdependent, as shown in Figure 7.

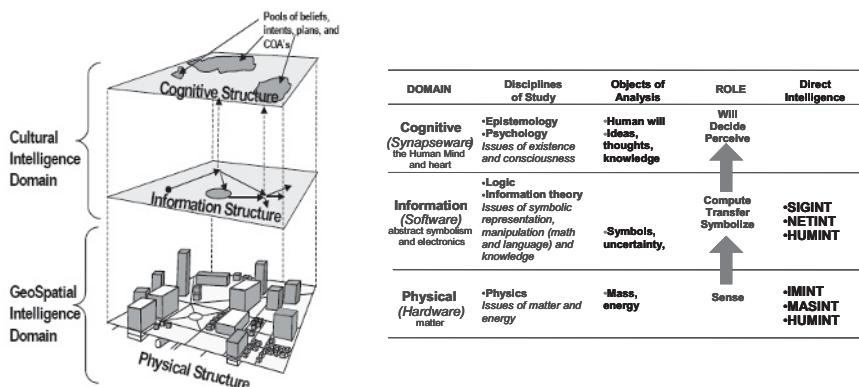


Fig. 7. Three State Layers for Information Fusion (after [14])

If we combine these two ideas, we see that we have a 12-perspective state vector to be produced by the IF process:

	Physical	Informational	Cognitive
Blue view of Red	The traditional view		
Blue view of Blue			
Blue view of (Red view of Blue)			
Blue view of (Red view of Red)			

The overall depth of insight that could be developed from these combined ideas is extraordinary and will not be achieved soon, but the ideas are worth pursuing and can provide one framework for IF technology planning and development.

3.5 The Information Challenge: PMESII

And finally, we come to last challenge discussed here—the Information Challenge. This challenge can be considered an overlay that could be added to any of the above ideas, especially the multi-perspective ideas just discussed. This PMESII term is just a shorthand for the several categories of information that would be part of the input stream of an idealized IF process: Political, Military, Economic, Infrastructure, Information. Incorporation of all these views and information types when developing an IF process, if done from a deductively-based approach, leads to a requirement to model societal interactions, a grand challenge in itself. Whether the defense and military research communities go down that path or not remains to be seen. All of these views and arguments about depth of awareness and understanding drawn from a highly-quantitative approach (as typically employed in IF process designs) will likely have to be balanced with the insights that can be gained from more subjective and social-theoretic approaches.

References

1. <http://faculty.ncwc.edu/toconnor/areas/worldconflicts.htm>
2. www.infoplease.com/ipa/A0904550.html
3. Mazaar, M.J., “Extremism, Terror, and the Future of Conflict”, Policy Review March 6, 2006

4. Quadrennial Defense Review Report, Feb 6, 2006
<http://www.globalsecurity.org/military/library/policy/dod/qdr-2006-report.htm>
5. Chong, C., "Problem characterization in tracking/fusion algorithm evaluation" IEEE Aerospace and Electronic Systems Magazine, Volume 16, Issue 7, July 2001 Page(s):12 – 17
6. Wolfram Math World, <http://mathworld.wolfram.com/Domain.html>
7. Chao, A.I., et al, "An extensible, ontology-based, distributed information system architecture", Information Fusion, 2003. Proceedings of the Sixth International Conference, Volume 1, 2003
8. Sterritt, R. and Bustard, D., "Towards an Autonomic Computing Environment", 14th International Workshop on Database and Expert Systems Applications, 2003
9. Evidence Based Research, Inc., "Network Centric Operations Conceptual Framework Version 1.0", Prepared for Office of Force Transformation, November 2003
10. Alberts, D. S. and Hayes, R. E. "Power to the Edge: Command and Control in the Information Age." Command and Control Research Program publications, (June 2003)
11. Ilachinski, A., "Land Warfare and Complexity, Part I: Mathematical Background and Technical Sourcebook", US Center for Naval Analyses, July 1996
12. Davis, Paul K. "Effects-Based Operations: A Grand Challenge for the Analytical Community", RAND Corporation, Santa Monica, Ca., 2001.
13. Biggie, J., Operational Net Assessment Concept Description, Nov 2003
http://www.mors.org/meetings/decision_aids/da_pres/Biggie.pdf
14. Vane, R., et al, "Urban Sunrise", US Air Force Research Lab report AFRL-IF-RS-TR-2004-22, Final Technical Report, February 2004

Situation Management: Basic Concepts and Approaches

Gabriel Jakobson¹, John Buford², Lundy Lewis³

¹Altusys Corp., jakobson@altusystems.com, ²Avaya, buford@avaya.com,

³Southern New Hampshire U. , l.lewis@snhu.edu

Abstract. This paper scopes the issues of Situation Management in dynamic systems, defines the basic concepts of Situation Management, and identifies several key enabling technologies. Particular focus of the paper is given to situation modeling. The paper reviews major aspects of situation modeling and discusses associated technologies, including Situation Calculus, Situation Semantics, Situation Control, Situation Awareness and others. In more detail we discuss an approach to situation management based on multi-agent systems, event correlation and case-based reasoning.

1 Introduction

The term Situation Management has been sporadically used in different connotations, e. g, in inter-personal influence and power management, processing abnormal alarms in industrial systems, in evaluating business/financial situations, and most recently, we have used this term in the context of managing complex dynamic operations and systems [27, 28]. The objective of this paper is to scope the issues associated with Situation Management (SM), define main concepts and identify technologies enabling SM. We have to mention immediately, that the field of SM is still in the rapid development and flux, and several concepts and solutions have not yet reached ultimate acceptance.

Research in dynamic, temporal, semantic and logical aspects of situations, and situational behaviour of humans, systems, and organizations have attracted diverse groups of scientists and institutions producing significant and well-articulated results in disciplines such as Situation Awareness [1], Situation Calculus [2], Situation Semantics [13] and Situational Control [3]. Despite individual merits, not enough synergy has been explored between these disciplines. This paper, while benefiting from the results obtained by these contributing to SM disciplines, aims defining a more general understanding of situations as objects of management. We see SM as a framework of concepts, models and enabling technologies for

recognizing, reasoning about, affecting on, and predicting situations that are happening or might happen in dynamic systems during pre-defined operational time.

In the overall “big picture” of SM (see Figure 1) we see a core theory of Situation Management, Contributing Disciplines, Associated Disciplines and Situation Management Applications. The core theory of SM includes Situation Modelling, Situation Recognition and Situational Reasoning. While describing the domain of SM, we will refer to several associated to SM disciplines, including Artificial Intelligence (AI), Information Fusion, Distributed Multi-Agent Systems, Semantic Web, Sensor Networks, Self-Organizing Systems, and Human Factors.

Our interest in this paper will be mostly on cognitive (high-level) aspects of SM, i.e. on aspects related to the meaning of situations, the intelligent methods of reasoning about the situations, and action planning. In order to exhibit such intelligent capabilities, the systems should possess fairly elaborate conceptual knowledge about the domain, i.e. domain ontology.

Let’s give some examples of SM applications: real-time fault, performance and configuration management in telecom, mobile ad hoc and sensor networks, tactical and asymmetric battlespace operations management; post-disaster emergency, and rescue and relief operations coordination.

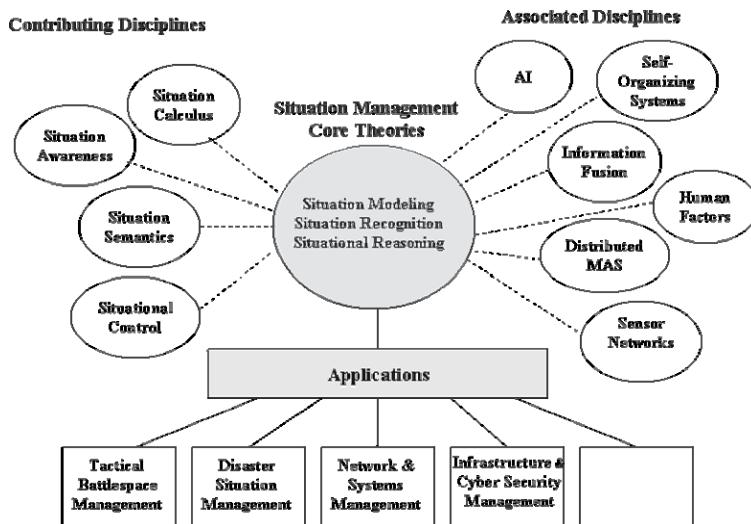


Fig. 1. Situation Management Domain

As an example of an application area of situation management consider a generalized scenario of coordination of post-disaster medical relief operations [4]. It is well known that natural, human-made and terrorist-caused disasters are unparalleled in the complexity and uncertainty of the challenges to response, relief and recovery operations. The medical relief operations are characterized by significant distribution of data across teams of people, systems, information sources, and environments.

At a high-level the generalized scenario of post-disaster medical relief operations involves the construction of a Disaster Situation Model in real time, which is constantly refreshed as the medical relief situation changes on ground. The Disaster Situation Model contains a knowledge-level view of the disaster from the medical relief perspective using an ontology designed for that domain. The model is created and updated by a constant flow of events and reports collected from the operational space. These events include both human intelligence and signal intelligence. Because of the large amount of raw data being collected, the event stream needs to be processed and correlated to produce “situational events”, i.e. events at the domain level. This reduction and inference step is performed during an information correlation stage.

Integrated with the real-time Situation Model are decision support systems (DSS) for medical relief operations. The DSSs rely on the Situation Model and operations staff oversight to manage the scheduling, dispatching, routing, deployment, coordination and reporting tasks. A chain of distributed communication and control systems leads from the DSS to the medical relief personnel in the field to direct the execution of these tasks.

The rest of the paper is organized as follows. In the next section we discuss the basic framework of SM, including the notion and components of SM. Section 3 describes the elements of situation modelling. Section 4 gives an overview of various disciplines, which are contributing to SM. The paper ends with the Conclusion section and References.

2 Situation Management: A General View

Situation Management is a synergistic goal-directed process of (a) sensing and information collection, (b) perceiving and recognizing situations, (c) analyzing past situations and predicting future situations, and (d) reasoning, planning and implementing actions so that desired goal situation is reached within some pre-defined constraints [27] (Figure 2).

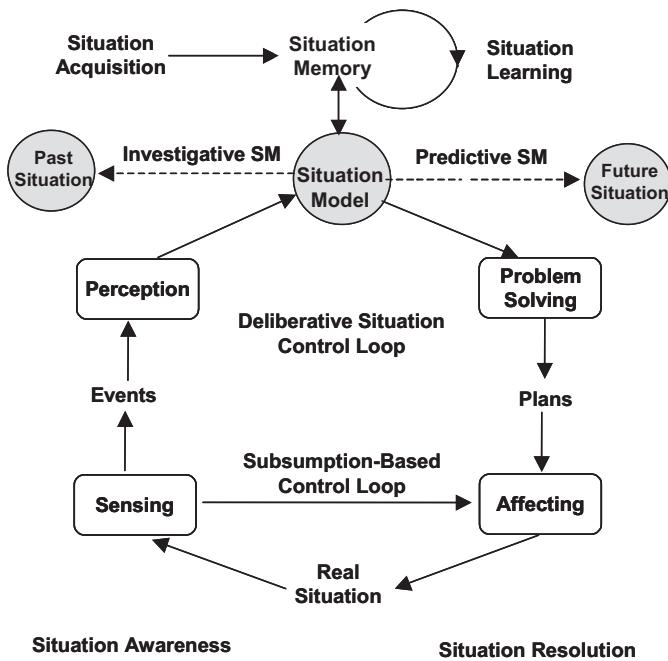


Fig. 2. Situation Management – A General Process Loop

Although the SM definition is rather lengthy, we feel that it covers the essential components of SM. First, depending on the particular goals of SM, one can entertain three aspects of it -- the investigative, control, and predictive aspects. It is useful to see these aspects mapped to a time axis. The investigative aspect of SM is concerned with a retrospective analysis of causal situations which determine why a certain situation happened. The control aspect of SM aims to change or keep the current situation, while the predictive aspect of SM aims to project possible future situations. For example, finding a root of a packet transmission failure in a telecommunication network is an example of an investigative SM; moving a tank unit from the area of direct hostile fire is a control type SM; and a projection of a potential terrorist attack on a critical infrastructure element is an example of a predictive SM.

In consideration of the control aspect of SM, Figure 2 depicts a deliberative situation control loop involving the four major steps of deliberation, sensing, perception, planning and effecting. The situation deliberation process echoes the architecture of intelligent control systems discussed in [5]. While the deliberative situation control architecture might be in some cases less effective than subsumption-based control [6]

where effectors are controlled directly by changes in sensing, the deliberative situation control model is generally superior in managing complex dynamic systems.

There are several other dimensions of SM that we should mention, e.g. situation awareness, resolution, acquisition, and learning. Situation awareness is based upon the steps of sensing and perception, aimed towards building an understanding of a current operational situation, very often in the ergonomics and human factor-based context of an operational room (see more on situation awareness in the Related Work section). Situation resolution is based on the steps of action planning and implementing the actions to close the situation control loop. Situation acquisition and learning, being off the timeline of direct SM processes, are the main sources of building knowledge structures required for the SM processes.

3 Situation Modeling

3.1 General Framework

In order to scope the issues of situation modeling, it is beneficial to see the following components of situation modeling: the structural, dynamic and representational components of situation modeling [27].

The structural component identifies the world, systems and individual objects, where we observe the happenings of events and evolving situations as situational transitions. Particularly, we identify the following structural objects:

- Entities: things, which have independent meaning in our area of interest;
- Attributes of the entities with associated attribute value domains and constraints;
- Classes of entities: the abstractions of the sets of entities that share common features like attributes, relations, operations, and behavior;
- Relations between the entities, including class, component, containment, location, and other domain-specific relations.

The dynamic component defines the behavior of the entities and entity systems in time. Our interest will be mostly on situations, events, and time.

The representational component is orthogonal to the two above-mentioned components and its utility is in the description of them. Essentially, the representational component is set of languages, and associated interpreters and environments. We can refer here to the following types of situation modelling languages:

- Primary concept specification languages, e.g. set-theoretical and finite state machines;
- Structural specification languages, e.g. OWL [7] and DAML [8];
- User-oriented graphical modelling languages, e.g. UML [9];
- Specialized languages, e.g. Situation Definition Language SDL [10] and GOLOG [11].

3.2 Structural Component of Situation Modeling

Entities and Attributes

Let's assume that there exists a world, real or imaginary that could be sensed, perceived, reasoned about and affected. The world is populated with entities, which are engaged in different class, structural, causal, spatial, temporal and other relations. The entities possess distinguishable attributes with certain attribute values. Some entities are active, they change their state in space and time, and some of the entities can interact with other entities forming multi-entity systems.

An entity is a thing of significance, either real or abstract that has distinctive existence. A set of entities with certain common attributes defines an abstract entity class.

Let e be an entity, $e \in E_i$, where $E_i \subseteq U$ is a subclass of all entities of the universe U . Entity e is represented by its set of attributes $\{a_1, a_2, \dots, a_p\}$.

Each attribute is a collection of attribute properties, such as attribute name, type, value, default value, get_value function, assign_value function, and other application-specific properties like attribute creation date, attribute security level, etc.

Attribute value (or default value) is a triplet containing an actual attribute value, certainty estimation, and time, either a point or interval time during which the attribute holds the actual value. Attribute a value at time t will be denoted as $a(t)$. We will consider entities as dynamic time-dependent objects with their time of creation t' , time of clear t'' , and corresponding lifespan $\delta = (t', t'')$. Any attribute value of an entity is defined only during the existence of the entity, i.e. $a(t), t \in \delta$.

The value of an attribute at a particular time can be determined by evoking the `get_value` function. A new value to the attribute is created by the `assign_value` function, whenever an external event arrives, or by system clock as a scheduled new value assignment, or by any other rule/action specific to the application domain. The latter case allows to model different natural phenomena associated, e.g. to the growth of biological systems or movement of physical objects. It also permits the use of the results of predictive modeling to project the potential future attribute values.

Among all entity attributes we will define a subset called situational attributes. The semantics of situational attributes is declared or computed depending on the particular operational context of the application. For example, a situational attribute could be one, which changes its value during the existence of an entity, or is in a significant relation with another situational argument.

Relations

Relation is a mental abstraction of linking a certain number, very often two, entities together. Mathematically, relation

$$R \subseteq E_1 \times \dots \times E_m = \{(e_1, \dots, e_m) / e_1 \in E_1, \dots, e_m \in E_m\},$$

where $E_1, \dots, E_m \subseteq U$.

In most practical applications it is enough to consider only binary relations $R \subseteq E_i \times E_j$.

A relation R could be considered as a set of instant relationships $R = \{r_1, \dots, r_q\}$. In case of binary relations the commonly used notation for $r \in R$ is $r = e_i R e_j$, where $r = (e_i, e_j) \in R$.

In several practical applications it is required to consider relations as entities, in sense that they are characterized by set of attributes $\{b_1, b_2, \dots, b_h\}$, and all the features that were attached to the attributes of entities.

In the same way as entities, we will consider relations as dynamic time dependent objects with their time of creation t' , time of clear t'' , and corresponding lifespan $\delta = (t', t'')$. The following time dependency should hold for a relationship: if $e_i R e_j$ and δ_i, δ_j are lifespans of e_i, e_j , accordingly, then for the relationship $e_i R e_j$ the lifespan $\delta \subseteq \delta_i \cap \delta_j$. Any attribute value $b(t)$ of a relation is defined only during the existence of the relation, i.e. $b(t), t \in \delta$.

For our further discussion, it is important to consider the following types of relations between entities: class, structural, special and domain-specific relations. Class relation establishes a link between an entity and

abstract entity class or between entity classes. Class relation is the major tool of conceptualization of entities and building conceptual frameworks of abstract concepts (ontology).

Structural relations Part-Of, Overlaps-With and Similar-With are the basic construction primitives of the universe. Spatial relations Inside, Near, Above, etc. are used to express topological (spatial) links between the entities. There is large number of various domain specific relations, which semantics depends on the particular domain. For example, Service x Supported-by Network y, Unit x Under-Fire-of unit y, Element x Connected-Trunk-T1-to Element y.

3.3 Dynamic Component of Situation Modelling

In this section we take a closer look at what situations are. In an everyday usage a situation is understood as a state of affairs or combination of circumstances at a given time moment. While this description of the situation captures the main characteristics of the notion, it is far from useful in situation management. Below we provide a formal definition of a situation

Base Situations

1. *Entity-based situations.* Let $\{a_1, \dots, a_p\}$ be set of situational attributes of entity e. Situation $S_e(d)$ on entity e during a time interval d, $d \subseteq \delta$, where δ is the lifespan of entity e is defined as

$$S_e(d) = \langle a_1(t), \dots, a_p(t) \rangle \in v_1 \times \dots \times v_p / \forall (t, t') \in d [\langle a_1(t), \dots, a_p(t) \rangle = \langle a_1(t'), \dots, a_p(t') \rangle]$$

As we see from the above-given formal definition a base situation $S_e(d)$ is a collection of entity e states (the time-stamped attribute vectors) that have the same value during a time interval d. Consequently, a situation has a duration, i.e. a start-time and an end-time.

2. *Relational entity-based situations.* Let $\{b_1, \dots, b_q\}$ be set of situational attributes of relation R. Situation $S_R(d)$ on relation R during a time interval d, $d \subseteq \delta$, where δ is the lifespan of relation R is defined as

$$S_R(d) = \langle b_1(t), \dots, b_q(t) \rangle \in v_1 \times \dots \times v_q / \forall (t, t') \in d [\langle b_1(t), \dots, b_q(t) \rangle = \langle b_1(t'), \dots, b_q(t') \rangle]$$

It is justified in several applications to consider relations as entities with their own attributes. Such interpretation of relations is also useful for aggregating multiple relations and handling such aggregation as a single entity. Following such formalism, used, e.g. in entity-relation modeling, we introduced the notion of a relational-entity-based situation as was defined above. Similarly to the entity-based situation $S_e(d)$, the relational entity-based situation $S_R(d)$ has its own duration, start-time and end-time.

3. Relational situations. Let $R \subseteq E_i \times E_j$, where $E_i, E_j \subseteq U$, $(e_i, e_j) \in R$, and δ_i, δ_j are lifespans of e_i, e_j , accordingly, then

$$S_{(e_i, e_j)}(d) = e_i R e_j$$

is a situation, where $d \subseteq \delta$, $\delta = \delta_i \cap \delta_j$, δ is the lifespan of the relation R and.

The relational situation considers only “relational” aspects of a situation, other words it is not concerned with the attribute values of the corresponding relation.

We should note that there is a difference between the entity (relation) lifespan and duration of a situation happening on that entity (relation).

Definitions 1-3 define the basic situational components that could be used for constructing more complex situations. During the SM process these basic situational components could be recognized using different methods, e.g. event correlation and case-based reasoning, as will be discussed in the Section 4.

Compound Situations

Complex compound situations could be constructed from other situations using set-theoretical union and intersection operations.

If $S_{B_1}(d_1)$ and $S_{B_2}(d_2)$ are two situations, where $B_1, B_2 \subseteq U$ and d_1, d_2 are subsets of common lifespans of all entities in B_1, B_2 , correspondingly, then,

$$S_B(d) = S_{B_1}(d) \cup S_{B_2}(d) \text{ and } S_B(d) = S_{B_1}(d) \cap S_{B_2}(d)$$

are situations, where, correspondingly

$$d = d_1 \cap d_2 \text{ and } B = B_1 \cup B_2 \quad \text{and} \quad d = d_1 \cap d_2 \text{ and } B = B_1 \cap B_2$$

While considering situations, we look on a subset of all entities, called active entities. Like situational attributes, the semantics of active entities is declared or computed depending on the specific operational context. Consequently, not all entities and relations of a large system are considered when observing situations happening in these systems. Due to our notions of active entities and situational attributes, multiple different situations can be defined on the same set of entities and relations.

Events

In a broad term, an event is an act of transition of a system from a state to state, or in our area of interest – from a situation to a situation. The external manifestation of such event is an informational event and as such it is an artifact created for human interpretation and practical utility. In more practical connotation, we consider an informational event as a time-stamped piece of information, which represents a change in the state of an object or manifests an action. Events could be considered either in a point or interval time. Several temporal relations can be defined between events, such as x AFTER y , x ENDS_BEFORE y , and x SIMULANEOUS_TO y , play a critical role in event correlation applied to dynamic situation analysis. Some classes of temporal event correlation in an interval time were discussed in [25].

4 Overview of Approaches to Situation Management

Research on understanding situations and reasoning about them have been the focus of various scientific communities, including logic, psychology, linguistics, human factors, artificial intelligence, and others. Not always those activities have been synergistic. In the previous section we gave a quite broad picture of SM including the steps of sensing, perception, problem solving and affecting the domain. We also talked about the investigative, control and predictive aspects of SM, the role of situation modeling, and learning and situation knowledge acquisition. It is obvious that such a rich and complex “ecosystem” of SM is based on multiple paradigms and enabling technologies prompting various approaches.

4.1 Situation Calculus

As it was mentioned earlier, in an everyday usage a situation is understood as a state of affairs or combination of circumstances at a given time moment. The first formal specification of a situation was given by McCarthy and Hayes in their Situation Calculus [3], where they used first order logic (FOL) expressions to define a situation as a snapshot of a complete world state at a particular time. Since it was computationally inefficient to consider a situation as a complete state of the world, Reiter and Pirri [12] in their approach to situation calculus defined a situation as a sequence of actions enabling calculation of the current state knowing the initial state and the sequence of actions transforming the initial state. For example, if S_0 is an initial state and $\text{do}(a, s)$ is a situation resulting from applying action a in situation s , then $\text{do}(\text{put}(A, B), \text{do}(\text{put}(B, C), S_0))$ is a situation resulting in putting block B on block C , and in an intermediate situation putting block A on block B .

Along with “fixed” actions Situation Calculus defines fluents, i.e. such functions and predicates whose value depends on situations, where they are applied. In Situation Calculus actions may have attached preconditions and effects (e.g. affecting the fluents). Situation Calculus has several well-known associated problems, namely, the Frame Problem (how to describe those aspects of a state, which are not changed by an action), the Ramification Problem (what are the ramifications and side-effects of performing of an action), and the Qualification Problem (what preconditions are required for performing an action). As the basis of the situation calculus a programming language GOLOG (alGOL LOGic) was developed and applied for several planning tasks, e.g. robot planning [12].

4.2 Situation Semantics

A deviation from a complete world state specification was also argued by Barwise [13], who looked on situations from the viewpoint of understanding speech acts by “intelligent situated agents”. Barwise and his colleges developed situation semantics theory based on FOL. The emphasis of the Barwise theory was not so much in exploring under what circumstances an utterance is true, but rather what is the semantics (i.e. meaning) of speech acts. In his later work Barwise made an important comment stating that in understanding language, thought and inference is crucial to handle situations as first class objects that can have properties and stand in relations.

4.3 Situation Control

Another school of understanding situations and the use of them in control of large engineering systems was proposed by Pospelov in Russia. Known as situational control theory [3, 14] it was based on semiotic models of the domain developed in linguistics and language psychology. Semiotics as a science of signs explores the syntactic, semantic and pragmatic aspects of signs. Pospelov considered situations as states of the relations between objects referred to some point in time. Pospelov's situation formalism was based initially on graph theory and finite state machines, and later on formal relational expressions close to FOL.

4.4 Situation Awareness

A closely related discipline to situational control is situation awareness. In the early nineties the term “situation awareness” was almost indistinguishable from industrial ergonomics and human factors studies of human operator safety and effectiveness. Several situation awareness models were proposed, most notably the models developed by Endsley and Garland [1]. Situation awareness has found an important place in data/information fusion research and engineering, initially related to military applications of signal fusion for target identification and tracking. The abstraction to a more general model of fusion prompted the development of the JDL fusion model, where level 2+ was directly associated with operational situation awareness and threat prediction [15]. Several studies have targeted the development of a synergistic model of situation awareness based on the JDL 2+ model and the situation awareness model of Endsley [16]. While research on situation control and situation awareness have been conducted to large extent independently, it is important to mentioned that the interest in semiotic models has been in both communities, e. g. the work by Hugo on the use of semiotics for describing a control room situation awareness [17].

4.5 Ontology-Based Approach

An ontology-based approach to situation awareness was developed by M. Kokar and his colleagues [18]. The approach uses formal ontology to describe events, domain objects and the relations between them, and logical rules to define the process of recognition of situations and situation transitions. A set of typical situations, e.g. an “under-the-fire” situation, were examined to define a core library of situations.

4.6 Multi-Agent Approach to Situation Management

Multi-agent systems are widely used for modeling complex distributed systems due to such features as a capability to act independently in a persistent manner, rational reasoning, interaction with the world, and mobility [19]. One can identify three layers of rational reasoning: the reactive, deliberative, and reflective layers. In the context of SM we see the reactive layer performing the following two major functions: (a) sensing and perceiving world events and (b) detecting multi-event patterns and recognizing situations happening in the world. The deliberative layer performs different reasoning functions related to various cognitive acts such as planning, control, prediction, and explanation. The reflective layer contains processes for reasoning about the events happening in the reactive and the deliberative layers, and how well they proceed with the goals set for the agent.

One of the most popular formal models of MAS is the Belief-Desire-Intension (BDI) model. It was conceived as a relatively simple rational model of human cognition [20]. It operates with three main mental attitudes: beliefs, desires and intentions. Rao and Georgeff [21] replaced the declarative notion of intentions with a procedural specification of instantiated and executable plans. Among many BDI agent models, the dMARS formalism serves as a well-recognized reference model for BDI agents [22].

An agent's beliefs are the facts about the World that the agent possesses and believes to be true. Desires are an agent's motivations for actions. Plans are operational specifications for an agent to act. An agent's plan is invoked by a trigger event, e.g. the acquisition of a new belief, removal of a belief, receipt of a message, or acquisition of a new goal. An agent's intention is understood as a sequence of instantiated plans that an agent is committed to execute. In response to a triggering external event, an agent invokes a plan from the plan library, instantiates it, and pushes it onto a newly created stack of intentions. In many applications found in telecommunications, military battlefields, homeland security, and other domains, decisions are often made on the basis of multiple events. Thus, the classical BDI agent model was extended with a situation awareness capability [23].

4.7 Event Correlation and CBR Approaches to Situation Management

Changes in the entities and relations can create fairly complex situational pictures of the world. Per our definition of situations, where we defined a situation as a collection of time dependent states of active entities and relations, we are able to consider multiple concurrent local situations happening in the world. Such an interpretation of the SM process allows an identification of two major steps of dynamic situation recognition: (a) recognition of local component situations using some technology and (b) constructing a global situational picture based on combining the component situations using the operations of union and intersection of component situations'

In this section we briefly review two technologies that have been found useful for handling complex situation: Event Correlation (EC) and Case-Based Reasoning (CBR). EC is an effective technology for recognizing situational patterns that involves complex real-time temporal analysis of multiple events [25, 26]. The EC technology has proved its effectiveness in very dynamic event-situation environments; however, its effectiveness decreases if very complex multi-component situations need to be recognized. Such deficiency of the EC approach could be addressed with building multi-layer EC systems using inter-connected EC nodes.

The CBR approach to SM has several benefits: it could be useful for recognizing very complex situational patterns [24]. In addition, previous applications of CBR have demonstrated the potentials of CBR for resolving the situation prediction and situation learning tasks. One the weak-points of CBR are a relatively static behaviour an inability to act in response to complex relations among the events. We found very promising an approach that combines the strong points of the EC and CBR approaches to SM. Such approach embedded into MAS environment was described in [23].

5 Conclusions

Situation Management field as described in this paper is in the stage of formation. There are several important driving forces behind the advancement of the field of SM, including the need for an integrated management of complex dynamic systems, progress in several important enabling technologies, and emergence of new critical applications such as operational battlefield management, disaster situation management, homeland security applications, and the management of complex dynamic sys-

tems and networks. At the same time SM faces several challenges, such as integration of computational and symbolic reasoning processes, development of situation modeling languages, increasing the effectiveness of SM models, tools and platforms, and the development of effective methods of situational learning.

References

1. M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human Factors*, 37(1), 1995, pp. 32-64.
2. J. McCarthy and P. Hayes, "Some Philosophical Problems from the Standpoint of Artificial Intelligence." In D. Michie, editor, *Machine Intelligence 4*, American Elsevier, New York, NY, 1969.
3. D. A. Pospelov, *Situation-Driven Control: Theory and Practice*. Moscow: Nauka, 1986.
4. G. Jakobson, N. Parameshwaran, J. Buford, L. Lewis, P. Ray, "Situation-Aware Multi-Agent System for Disaster Relief Operations Management," in Proceedings of the 3rd Information Systems for Crisis Response and Management Conference ISCRAM 2006, Newark, NJ, USA, May 2006.
5. J. S. Albus and A. M. Meystel, "A Reference Model Architecture for Design and Implementation of Intelligent Control in Large and Complex Systems," *International Journal of Intelligent Control and Systems*, Vol.1, No. 1, 1996, pp15-30
6. R. A. Brooks, "A Robust Layered Control System for a Mobile Robot," *IEEE Journal of Robotics and Automation*, RA-2(1), 1986, pp. 14-23.
7. M. K. Smith, C. Welty, and D. L. McGuinness (Editors), *OWL Web Ontology Language Guide*, W3C Recommendation, 2004.
8. J. Hendler and D. McGuiness, "The DARPA Agent Markup Language," *Intelligent Systems*, 15, No. 6, 2000, pp. 67-73.
9. S. Cranefield, and M. Purvis, "UML as Ontology Modeling Language," in Proceedings of the Workshop on Intelligent Information Integration, 16th International Joint Conference on AI (IJCAI-99), Germany, 1999.
10. S. Greenhill, S. Venkaresh, A. Pearce, T. C. Ly, *Situation Description Language (SDL) Implementation*, Defense Science and Technology Organization, Department of Defense, Australian Government, Report DSTO-GD-0342, 2002.
11. H. J. Levesque, R. Reiter, Y. Lesp  rance, F. Lin, R. B. Scherl, "GOLOG: A Logic Programming Language for Dynamic Domains," *Journal of Programming*, 31, 1997.
12. F. Pirri and R. Reiter, Some Contributions to the Situation Calculus. *J. ACM*, 46(3): 325-364, 1999.
13. J Barwise, "The Situation in Logic," CSLI Lecture Notes, Number 17, Leland Stanford Junior University, 1989.

14. A. I. Ehrich, V. F. Khoroshevsky, D. A. Pospelov, G. S. Osipov, "Semiotic Modeling and Situation Control", in Proceedings of the 1995 ISIC Workshop, 10th IEEE International Symposium on Intelligent Control, Monterey, California, USA, 1995.
15. A. N. Steinberg, C. L. Bowman, and F. E. White, "Revisions to the JDL data fusion model," in Proceedings of the NATO IRIS Conference, Quebec, Canada, October 1998.
16. J. Salerno, M. Hinman, D. Boulware, "Building a framework for situation awareness", in Proceedings of The 7th International Conference on Information Fusion, Stockholm, Sweden, 2004, pp. 219-226.
17. Hugo, J. (2005), "The Semiotics of Control Room Situation Awareness", Fourth International Cyberspace Conference on Ergonomics, Virtual Conference, 2005.
18. C. J. Matheus, M. M. Kokar, and K. Baclawski, "A Core Ontology for Situation Awareness," in proceedings of the 6th International Conference on Information Fusion, 2003, PP. 5454-552.
19. M. Wooldridge. An Introduction to Multi-Agent Systems. John Wiley and Sons, 2002.
20. M. Bratman, Intension, Plans, and Practical Reason. Harvard University Press, 1987.
21. A. Rao and M. Georgeff, "BDI Agents: From Theory to Practice." in Proceedings of the First International Conference on Multiagent Systems (ICMAS'95), 1995.
22. M. d'Inverno, M. Luck, M. Georgeff, D. Kinny, and M. Wooldridge, "The dMARS Architecture: A Specification of the Distributed Multi-Agent Reasoning System," Journal of Autonomous Agents and Multi-Agent Systems, 9(1-2), 2004, pp.5-53.
23. J. Buford, G. Jakobson, L. Lewis. "Multi-Agent Situation Management for Large-Scale Disaster Relief Operations Management," Special Issue on "Emergency Management Systems", International Journal of Intelligent Control and Systems, 2007 (forthcoming)
24. L. Lewis. Managing Computer Networks: A Case-Based Reasoning Approach. Artech House, 1995.
25. G. Jakobson and M. Weissman, Real-Time Telecommunication Network Management: Extending Event Correlation with Temporal Constraints. Integrated Network Management IV, IEEE Press, 1995.
26. G. Jakobson, M. Weissman, L. Brenner, C. Lafond and C. Matheus, "GRACE: Building Next Generation Event Correlation Services," in Proceedings of the IEEE Network Operations and Management Symposium, April 2000.
27. G. Jakobson, L. Lewis, C. Matheus, M. Kokar, J. Buford. Overview of Situation Management at SIMA 2005. SIMA Workshop at MILCOM 2005.
28. G. Jakobson, J. Buford, and L. Lewis. "Towards an Architecture for Reasoning about Complex Event-Based Dynamic Situations," International Workshop on Distributed Event-Based Systems DEBS '04, Edinburgh, UK.

Maritime GIS: From Monitoring to Simulation Systems

C. Claramunt, T. Devogelet, S. Fournier, V. Noyon, M. Petit, C. Ray

Naval Academy Research Institute, Lanveoc-Poulmic, BP 600, 29240 Brest Naval,
France
{name}@ecole-navale.fr

Abstract. Combined research in the fields of Geographical Information Systems (GIS) and maritime systems has finally reached the point where paths should overlap and continue in better unison. This paper introduces methodological and experimental results of several marine-related GIS projects whose objectives are to develop spatial data models and computing architectures that favour the development of monitoring and decision-aid systems. The computing architectures developed integrate agent-based reasoning and distributed systems for the real-time monitoring, manipulation and simulation of maritime transportation systems.

1 Introduction

Recent advances in telecommunication and positioning systems, client-server and distributed architectures, and mobile devices offer new perspectives and challenges to maritime GIS and transportation research. However, there is still a need for data management and communication protocols, graphic and exploration interfaces that support the development of integrated maritime GIS. Such systems will be of great interest for many maritime applications oriented to the monitoring and analysis of maritime traffic and transportation.

One of the current limitation to the development on integrated maritime and GIS systems relies in the fact that current GIS models, software and interfaces do not yet provide the functionalities to make this technology compatible with maritime transportation, particularly when considering the telecommunication and navigation-based systems available in maritime transportation. This poor level of integration is often the result of the different paradigms used within GIS and maritime systems, and the resulting fact that the development of integrated solutions implies the re-design of existing software solutions. Moreover, current GISs are not adapted to the management of dynamic phenomena due to the lack of modelling and

processing interoperability with real-time navigation systems. The development of real-time GISs, characterized by a high frequency of changes, implies a reconsideration of the storage, modelling, manipulation, analysis and visualization functions whereas current GIS models and architectures have not been preliminarily designed to handle such dynamic phenomena.

Safety and security are constant concerns of maritime navigation, especially when considering the constant growth of maritime traffic around the world, and constant decrease of crews on decks. This has favoured and lead to the development of automated monitoring systems such as the AIS (Automatic Identification System) and the ECDIS (Electronic Chart Display and Information System) as a support of electronic mapping services. However, officers on the watch and monitoring authorities still require the development of additional and advanced decision-aid solutions that will take advantage of these communication and cartographical systems and thus improve their benefits.

Amongst several technological solutions that might contribute to the emergence of maritime-based decision-aid systems, integration of Geographical Information Systems (GIS) with maritime navigation systems appears as one of the promising directions to explore. The projects presented in this paper present several contributions to such a field of maritime GIS: from the real-time monitoring of navigations for a local authority and maritime clients (Section 2), to the diffusion of maritime data to mobile interfaces (Section 3), and the development of a relative-based model and visualisation system for maritime trajectories (Section 4). Finally Section 5 draws some conclusions and perspectives.

2 Real-time Traffic Monitoring

Nowadays, Automatic Identification Systems (AIS) used to detect and warn about possible maritime navigation collisions are a suitable solution for well equipped ships, but are unfortunately relatively expensive and difficult to maintain for small ships and pleasure boats. The Share-Loc project (fig. 1) purpose is to design and implement a flexible maritime navigation system for small ships and boats [1][3].

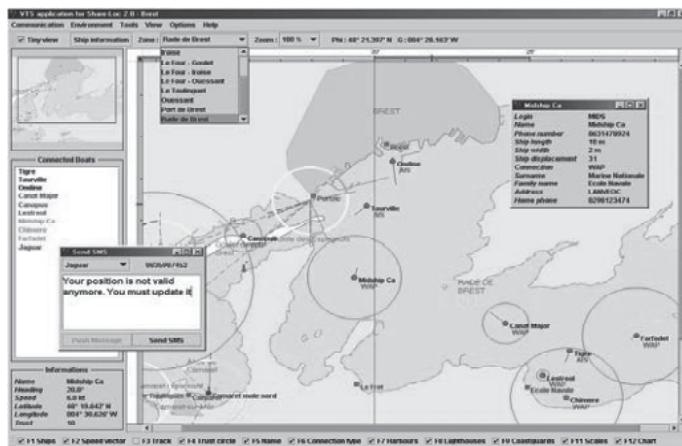


Fig. 1. VTS application

The Share-Loc system is based on a distributed architecture and real-time services for the diffusion of maritime geographical information, at different levels from the global monitoring of the maritime traffic of a given area, to individual services on request. The Share-Loc system is made of a navigation database server and mobile navigation clients. It is based on a client-server architecture that maintains a global view of a given navigation area on the server side, and a WAP-based solution that provides location-based data on the clients side. The server is a web-based program running on an Internet connected computer. Each client is a WAP-enabled mobile phone device that accesses the WEB server data through a WAP gateway program running on the server computer. Mobile clients are connected to Internet service providers according to their choice of telephone handset and mobile telephone network. Subsequently each authorized client is able to request the appropriate address on the server and to start the navigation-based application. The project provides a solution for the monitoring of small ships, and is oriented towards Vessel Traffic Services (VTS). A VTS is a marine traffic monitoring system established by harbour or port authorities, similar to solutions used in air traffic control. VTS are based on radar systems and AIS to keep track of vessel movements, and improve navigation safety in a given maritime area.

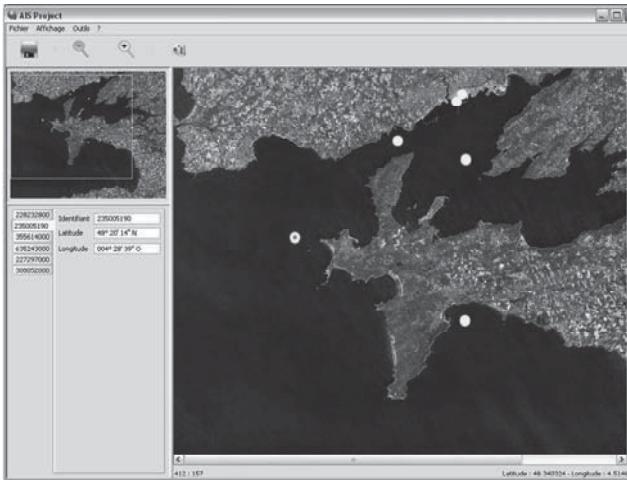


Fig. 2. Real-time AIS monitoring

This system is completed by a monitoring-based project whose purpose is a management and visualization system for positional data coming from an Automatic Identification System (AIS) that uses a VHF ad-hoc network to broadcast data over large distances. The aim is to provide a generated view of a real-time maritime situation of the Brest maritime neighbourhood. Thanks to a three-tiers application, AIS messages are received from ships and stored on the server side. Clients access information using a Web interface written in Java. This constitutes a sort of abstraction of the appliance level. The application is designed to run on every device that can browse the Web and execute a Java runtime. AIS messages provide additional information on the ships manipulated (fig. 2). Figure 2 shows a maritime configuration where each spot represents a vessel location. Attribute values of the selected vessel are displayed on the left part of the interface.

A second development that extends the functional objectives of the Share-Loc prototype relies in a real-time monitoring system developed in the context of an international sailing race. This event has a large audience, and requires appropriate solutions to diffuse real-time information, from the coastal maritime area to the users located in the ground. This generates different needs in term of geographical information usage and appliance. The experimental prototype is composed of two parts: a wireless network and an experimental adaptive GIS. Ships locations during the race are acquired through a real-time infrastructure. The implementation of the geographical context into different views and an appliance context

divided into several classes of devices are considered by the adaptive process.

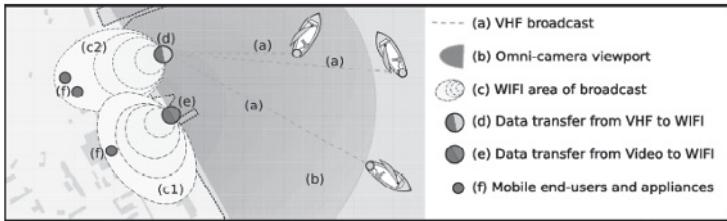


Fig. 3. Real-time communication infrastructure

The user context is modelled by a generic concept of user group that aggregate user behaviours. An important aspect of the adaptation process is its ability to integrate real-time geolocalisation information that delivers GIS data and influences the geographical context and related services. A localization system has been developed and allows for real-time reception of ship's positions and a continuous video stream of the race (fig. 3). Locations are provided by an embedded system available on ships. This system includes a GPS, a configurable modem, a VHF transmitter and fulfills several constraints: light weight (less than a kg), long range (5 to 10 km), high autonomy (8 to 10 hours). This module collects and diffuses the real-time locations of the ships to the ground station (fig. 3-(d)). The transmission to the ground station is a VHF communication (fig. 3-(a)) based on APRS frames (Automatic Position Reporting System). The ground station is composed by a VHF receiver and a VHF-to-WiFi bridge that broadcasts real-time data to a given area (fig. 3-(c2)). Mobile end-users (fig. 3-(f)) located in this broadcast area can access ships data, whatever the form of their appliance.

A general drawback of coastal sailing races is the lack of visibility on the ground, because of the distance to the coast. As real-time positions are crucial, video streams are provided and offer a concrete service of the geographical data. The installed video system presented in fig. 3-(e) broadcasts a video stream of the race (fig. 3-(b)) to a given WiFi deserved region (fig. 3-(c1)).

Complementary geographical data views, each associated to a particular service, are presented to the user. The “2D mapping” service delivers ships location information. Different levels of zoom are automatically computed several times per minute to provide detailed views on the race activity. The “3D mapping” service displays similar data, but in a 3D view of the region of interest. Basic displacement and zooming functions in the

scene are available using keyboard combinations. The “Video” service provides a real-time view of the race region. Zooming and camera movements are also allowed.

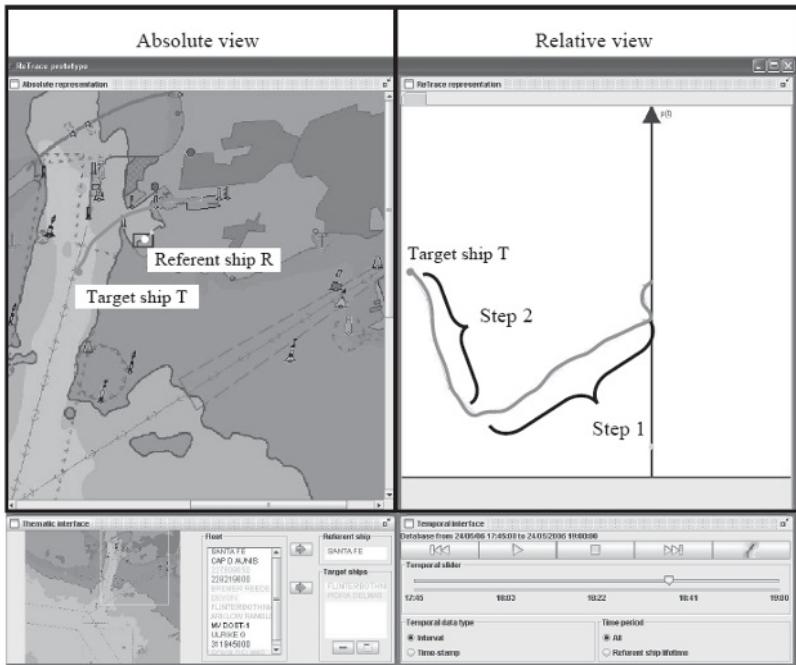


Fig. 4. ReTrace prototype

Despite the interest of modelling and visualisation functionalities, there is still a need for additional prediction mechanisms that should improve the monitoring and planning of navigation decisions. Cognitive studies have showed that the conventional absolute vision of space does not completely reflect the human perception of the environment [10][11]. Observer-based properties such as the relative position and relative velocity of an object with respect to an observer are difficult to evaluate visually when using an absolute frame of trajectory representation. The relative view of space provides a human-oriented vision of space to model closer to the way it is perceived from a mobile observer acting in the environment [12]. We made the assumption that it might be possible with a relative frame of reference to offer a different view of the way a mobile object behaves in space and time. This should offer a direction to explore for the study of maritime trajectory interactions, particularly when the objective is to analyse the behaviour of one or several moving objects (i.e. ships) with re-

spect to an observer also acting in the environment (i.e. either a ship or an observer acting in the environment).

This leads us to explore and design a relative-based trajectory data model: the ReTrace data model (Relative-based Trajectory data model) [13], where relative position and relative velocity of a mobile object with respect to an observer acting in the environment are modelled over time. One of the research objectives is to explore to which degree such a model completes the conventional absolute view of a spatial trajectory. The model is supported by formally-defined process primitives that are also qualified at the cognitive level [13]. A prototype implements the ReTrace data model with a dual visualisation frame that integrates the absolute and relative representations (fig. 4). The figure shows a target object T that gets closer the referent object R and accelerates during the time interval Step 1, and moves away from the referent object R and accelerates during the time interval Step 2. This dual representation provides complementary views that favour the understanding of trajectory behaviours in space and time, thus favouring the analysis of the emerging processes and patterns.

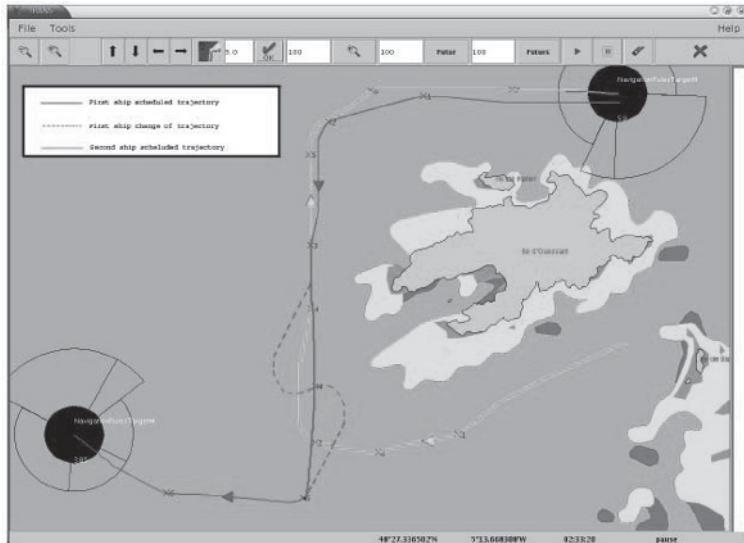


Fig. 5. TRANS scenario example

Simulation capabilities are also required to predict future vessel positions. The TRANS prototype (Tractable Role-based Agent Prototype for Concurrent Navigation Systems) develops a multi-agent spatial decision support system that supports micro-simulation capabilities, and where ships are modelled as autonomous agents acting in their environment

(fig. 5) [5]. Several modelling concepts enhance the semantics of this meta-model. In particular, they allow an agent to be part of a group and to act according to the roles defined at the group level. Role priorities and constraints give additional flexibility to the meta-model. Simulation objectives of the TRANS prototype are to model and anticipate ship behaviors and trajectories in order to avoid collisions and running aground. The figure 5 illustrates the TRANS interface and displays two ship trajectories and show a case where one ship avoids the other by the application of navigation rules. This ship then returns to its initial trajectory. These changes of behaviours are supported by the role mechanisms. The aim is to develop a realistic simulation environment that can be used either for monitoring purposes or as educational software for training purpose. Further work concerns progressive interoperability of the TRANS prototype with simulation systems that integrate the modelling of continuous phenomena such as ship trajectories, tides, streams and winds.

3 Towards Adaptive Interaction Techniques

These navigation-based systems illustrate the variety of methods and research issues that support exchange of geographical data between a centralised system and mobile users. Although they make a basic usage of the geographical context taking into account ships location and characteristics, no flexible automated adaptation is provided to the users at the interface level. These elements are not new when studied individually, but less considered as a whole. For instance, previous work in the field of adaptive GIS introduces a technology-driven approach for an hardware-based interaction medium [8]. Adaptation of an open GIS layer descriptor to specific user needs and contexts has been also studied in [7]. A context-sensitive model for mobile cartography that emphasizes different levels of data adaptation and presentation has been proposed in [9].

Another on-going project introduces an adaptive GIS defined as a generic and context-aware GIS that can be automatically adapted according to several contexts defined by (1) the properties and location of the geographical data manipulated, (2) the underlying categories that reflect different user profiles and (3) the characteristics of the computing systems, supporting web and wireless techniques [4]. This classification has been inspired by a previous work done by Calvary et al. [2]. These contexts cover the components of the diffusion of geographical data in wireless environments. The dimensions identified are of different nature as they involve data, computing processes and interfaces, and categories of users.

In order to consider the problem from a global point of view, this project develops an integrated contextual-based architecture that considers these different factors and interrelationships. The framework is developed and applied to maritime navigation (fig. 6) and combines mobility and distributed services. From a maritime point of view, heavy and even increasing maritime traffic need very low human response time in order to prevent accident. The use of adaptive GIS as a decision-aid system for end-users appears as a useful approach for maritime transportation systems.



Fig. 6. Adaptive interaction

4 Conclusion

The development of integrated maritime and GIS systems still requires the integration of different geographical information sources to be combined, adapted and shared in real-time between different levels of users acting in the maritime environment. The development of information and telecommunication technologies brings new and often unexpected possibilities for integrating, analyzing and delivering maritime traffic data within GIS. Integrating GIS information architectures and services with maritime information systems should improve the economical and technological benefits of transportation information by allowing the diffusion of traffic information to a larger community of decision-makers, engineers and final end-users.

Research challenges are varied: development of cross-domain protocols and exchange standards for the transmission and interoperability of traffic data. Conventional statistical, geographical data analysis and visualization methods should also be adapted to the specific nature of traffic information

often associated with large volumes of data. At the implementation level, there is a need for the development of GIS-based distributed computing environment, computational and processing capabilities as traffic data and applications are usually physically allocated in different geographical locations and computationally expensive in terms of the data volumes generated. The diversity of projects presented in this paper illustrates the range of opportunities of the integration of GIS and Intelligent Transportation Systems (ITS) for maritime navigation. We believe that all these application domains should benefit for this information integration and those methodological findings should be shared and cross-fertilized amongst the research communities active in these fields.

References

1. F. Barbe, F. Gelebart, T. Devogelete and C. Claramunt, A knowledge-based GIS for concurrent navigation monitoring, *GIS in the Environment*, P. Hall (ed.), Taylor and Francis, pp. 135–146, 2001
2. G. Calvary, J. Coutaz, D. Thevenin, Q. Limbourg, L. Bouillon, and J. Vanderdonckt, A unifying reference framework for multi-target user interfaces, *Interacting with Computers*, vol. 15(3), pp. 289–308, 2003.
3. G. Desvignes, G. Lucas de Couville, E. Peytchev, T. Devogelete, S. Fournier and C. Claramunt, The Share-Loc project: A WAP-based maritime location system, In *Proceedings of the 3rd International Conference on Web Information Systems Engineering*, B. Huang et al. (eds.), IEEE press, pp. 88–94, 2002
4. M. Petit, C. Ray and C. Claramunt, A contextual approach for the development of GIS: Application to maritime navigation, In *Proceedings of the 6th International Symposium on Web and Wireless Geographical Information Systems (W2GIS 2006)*, J. Carswell and T. Tezuka (eds.), Springer-Verlag, LNCS 4295, 2006, to appear
5. S. Fournier, T. Devogelete and C. Claramunt, A role-based multi-agent model for concurrent navigation systems, In *Proceedings of the 6th AGILE Conference on Geographic Information Science*, Gould, M. et al. (eds.), Presses Polytechniques et Universitaires Romandes, pp. 623–632, 2003
6. A. Zipf, Using styled layer descriptor (SLD) for the dynamic generation of user and context-adaptative mobile maps - a technical framework, In *Proceedings of the 5th International Workshop on Web and Wireless GIS (W2GIS 2005)*, K.J. Li and C. Vangenot (eds.), Springer-Verlag, LNCS 3833, pp. 183–193, 2005
7. H. Hampe and V. Paelke, Adaptive maps for mobile applications, In *Proceedings of Mobile Maps, Interactivity and Usability of Map-based Mobile Services*, 2005
8. T. Reichenbacher, Adaptive methods for mobile cartography, In *Proceedings of the 21th International Cartographic Conference*, pp. 1311–1322, 2003

9. B. Tversky, Cognitive maps, cognitive collages and spatial models, Spatial Information Theory: Theoretical Basis for GIS, A. Franck and I. Campari (eds). LNCS 716, Springer- Verlag, pp. 14-24, 1993
10. W. Maass, A cognitive model for the process of multimodal, incremental route descriptions, Spatial Information Theory: Theoretical Basis for GIS, A. Franck and I. Campari (eds.). LNCS 716, Springer-Verlag, pp. 1-13, 1993
11. S. Imfeld, Time, points and space – Towards a better analysis of wildlife data in GIS, Technical Report, University of Zurich, 2000
12. V. Noyon, T. Devogele and C. Claramunt, A relative-based model for the representation of trajectories, 2006, submitted

Intelligent Images Analysis in GIS

Philipp Galjano and Vasily Popovich

St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences
14-line 39, St. Petersburg, 199178, Russia, popovich@iias.spb.su

Abstract. The paper proposes an analysis of conventional pattern recognition methods and their use for images' analysis in GIS applications. The methods like cluster analysis, neural networks, and immunocomputing are studied in detail. Classic methods of raster segmentation, including lines detection, levels' difference, connecting of contours, threshold processing, and other are considered as applied to snapshots' analysis. Special attention is paid to immunocomputing method implementation. The above method efficiency as well as its application's domain is demonstrated with specific examples. Methods of isomorphism and metaclasses are developed to automatically recognize complex objects. The methods' efficiency as well as advantages and disadvantages are manifested based on a real snapshot.

1 Introduction

Methods of Earth's Remote Sensing Data (ERSD) analysis are considered in the paper under the angle of their integration (fusion) into GIS. At that an intelligent GIS (IGIS) [7] realizing methods and techniques of Artificial Intelligence (AI) at representing geo-spatial data and modeling geo-spatial phenomena and processes is considered. At growth of a demand for ERSD also grows the importance of their automatic analysis methods.

Analysis of ERSD aimed at classification of terrestrial areas and search of given type objects is one of the problems being solved by IGIS. Both above described tasks are studied within the AI framework and represent particular cases of pattern recognition problem; as a rule classic mathematical statistics' methods and artificial neural networks (ANN) are used to solve them. Now there exists appeared a comparatively new AI trend called immunocomputing (IC), that surpasses the most advanced leads of computing intelligence: ANN and genetic algorithms (GA) in performance (forty times at the least) and in error-free recognition (half as much again at the least). The given paper considers IC application to both tasks solv-

ing in a learning mode with a teacher. Two methods based upon a sought object's representation as a metaclass and pattern-matching search were additionally developed and used to solve task 2. Here the problem of pattern recognition served as a Body of Interest and recognition of raster images by pattern-matching search techniques ("isomorphism" method), expedient geometric relations analysis ("metaclass" method) and immunocomputing technologies served as a Subject. Objects of "aircraft" type were used as a sought object.

2 Analysis of Pattern Recognition Conventional Methods

2.1 Cluster Analysis

The term of cluster analysis, pioneered by Tryon in 1939, comprises over 100 various objects' classification algorithms. To be analyzed all objects are described against similar characteristics' scale. Objects are united in clusters depending on their similarity. Characteristics' values form an *attribute space*.

Cluster analysis methods could be divided into two groups: hierarchical and nonhierarchical .

The hierarchical clustering main point is either a sequential integration of smaller clusters in bigger ones or splitting of bigger clusters into smaller ones. The major advantage of hierarchical clustering methods is their evidence.

Agglomerative Nesting, AGNES

At the algorithm start all objects (each raster's pixel in case of ERSD analysis) are separate clusters. At the first stage the most alike objects are integrated in a cluster. At the following stages the integration continuous until all elements compose one cluster.

Divisive ANALYSIS, DIANA

These methods appear to be a logical antagonist to agglomerative methods. At the algorithm starting all objects belong to one cluster being divided into smaller clusters at the succeeding steps, and in the result a sequence of splitting groups is formed.

Nonhierarchical Clustering consists in a data set splitting into a definite number of separate clusters. There exist two approaches: the first one concludes in clusters' delimitation as the densest zones within multi-dimensional basic data space, i.e. cluster's defining at the high "points"

densening”; the second approach concludes in minimization of the objects’ distinction measure.

The k-means algorithm, also called a rapid cluster analysis, prevails among nonhierarchical methods.

Algorithm description

- Initial objects’ cluster distribution. The number k is selected, and at the first step certain points are considered to be clusters’ centers. For every cluster there exists one center, so each object is designated to a definite cluster.
- Iterative process. The clusters’ centers are calculated, that further are considered to be clusters’ coordinate-wise means. Objects are being re-distributed again.

The process of the clusters’ centers calculation and objects’ redistribution continuous until one of the following conditions is met: either the clusters’ centers got stabilized, i.e. all observations belong the cluster where they belonged prior the current iteration, or the iterations number equals the maximal iteration number.

The k-means algorithm’s simplicity and usage rate, its intelligibility and clarity are its undoubted virtues. On the other hand its sensitivity to outliers, that can distort means; slow functioning in case of great databases can be regarded as disadvantages. The use of data sample leads to a possible solution of the given problem.

Methods’ Comparative Analysis

Nonhierarchical methods demonstrate the higher stability regarding noises and outliers, incorrect metrics selection, inclusion of insignificant variables in a set engaged in clustering. However, at their use an analyst should in advance determine a number of clusters, number of iterations or a stopping rule, as well as other clustering parameters.

Limitations on data set level, difficulties of proximity measure selection, inflexibility of received classifications comprise complexities to clustering’s **hierarchical methods**. On the other hand the advantage of these methods in comparison with nonhierarchical ones lies in their clearness and a possibility to receive an idea of data structure in detailed.

Currently new requirements that should be met by clustering algorithm have appeared caused by an emergence of hyper-big data bases. The major new requirement is the algorithm scalability. So, lately new clustering algorithms capable of hyper-big data bases processing are being actively developed; among them are generalized clusters’ representation as well as selecting and use of data structures supported by underlying data bases. Algorithms integrating hierarchical clustering methods with other ones

(BIRCH, CURE, CHAMELEON, ROCK) have been developed. It is worth noting that an effective clustering algorithm is supposed to reveal results' independence of input data order and algorithm's parameters independence of input data.

2.2 Neural Networks

Neural Networks – are the models of brain biological neural networks, where neurons are imitated by relatively simple elements (artificial neurons) often of the same kind. The neural network can be represented by a directed graph with weighted connections, where artificial neurons are graph nodes and synaptic connections are arcs. In the best known configuration the input signals are being processed by an adaptive adder, then the adder's output signal enters nonlinear converter and is converted by an activation function, after that the result enters an output (branch point). A neuron is characterized by its current state and possesses a group of synapses – unilateral input connections linked with the outputs of other neurons. Neuron possesses an axon – an output connection of a given neuron where from an excitation or stopping signal enters the synapses of the following (next) neurons. One or several neurons whose inputs are given the same common signal are called a layer. The current neuron's state is determined as a weighted sum of its inputs; the neuron's output is its state function.

Neural Network Type	Brief Description	Advantages	Disadvantages
Rosenblatt's Perceptron	A separate neuron calculates a weighted sum of an input signal's elements, subtracts a shift value and passes the result through a stiff threshold function, whose output equals +1 or -1. Depending upon an output signal's value, a decision is made what class an input signal belongs of the two ones.	Software and hardware realizations of a model are rather simple. Learning algorithm is fast and simple.	Primitive separating surfaces (hyper planes) only allow solving the simplest recognition problems.
Neuronet of Back Propagation	As a rule, it has an input layer, output layer and at least one hidden	Back Propagation is an efficient and popular learning al-	The method of back propagation considers a multi-criterion optimiza-

Neural Network Type	Brief Description	Advantages	Disadvantages
	<p>layer. Neurons are organized in layer-by-layer structure with a direct signal transmission. Each neuron generates a weighed sum of its inputs, passes this value through a transfer function and generates an output value.</p>	<p>gorithm of multi-layered neuronets allowing to solve many practical tasks.</p>	<p>tion problem as a set of mono-criterion problems, at that, at each iteration the values of net parameters that only improve operation with one example of a learning sample are changed. Such an approach considerably reduces the learning rate.</p>
Delta Bar Delta	<p>Is a formal modification of Back Propagation net. Realizes “informal” approach to artificial nets learning when each weight has its own adapted learning factor, and past error’s values are used to calculate the future values.</p>	<p>Delta Bar Delta paradigm is an attempt to speed up the process of back propagation algorithm convergence at the expense of using an additional information about parameters’ and weights’ changes during the learning process.</p>	<p>1) Even minor coefficient’s linear increase can lead to a significant increase of learning rate, that causes sudden changes in weights’ space.</p> <p>2) Now and then a coefficient’s decrease can be not sufficiently fast.</p>
Directed random search	<p>Feed Forward standard architecture based on Back Proragation algorithm and correcting weights at random is used. To secure an order in such a process a direction component is added to a random step, that guaranties weights’ assignment to preliminary successful direction of search.</p>	<p>For small and medium neuronets a random guided search provides for good results within a short time.</p>	<p>A number of connections’ weights imposes a practical limitation on the problem dimension.</p>
Kohonen Nets	<p>Has just two layers: input and output, and is called self-organized map.</p> <p>Successive approximations method supports Kohonen Net learning. Starting from ran-</p>	<p>Kohonen Net is capable of functioning under obstacles’ conditions, as a number of clusters is fixed, the weights are being modified slowly, the weights</p>	<p>The Net can be used for cluster analysis only when the number of clusters is a priory known.</p>

Neural Network Type	Brief Description	Advantages	Disadvantages
	domly selected centers' output arrangement, the algorithm gradually improves for learning data clustering. их данных.	adjustment ends after learning.	
Learning Vector Quantization	Comprises an input layer, self organized Kohonen map and output layer. Kohonen layer is learning classification using a learning set. The Net uses rules of controllable learning.	Exactly specifies a bound between domains with a possibility of incorrect classification.	For complex classification of similar input examples the Net requires a big Kohonen map with a great number of neurons per a class. The above can be mastered by an appropriate selection of learning examples or by an extension of input layer.

2.3 Immunocomputing

Immunocomputing (IC) studies information processing principles used by immune system for solving certain complex problems including pattern recognition. A concept of formal protein (FP) forms a basis for IC. FP is an abstraction describing biophysical principles that determine dependence between protein's free energy (see below) and space configuration. In respect to the method of computer realization FP is described by four real numbers that are the analogues for biological protein aminoacid bonds. At that control actions are considered the FP input and its state the FP output. Apparently, FP is the simplest mathematical model, bearing such important biological protein features as its ability to self-assembly and initial (source) form non-communicative dependence of the bonds' number and order.

FP *self-assembly* is a process by whose means FP changes its state in conformity with the differential equations systems given in [1]. In respect to computer realization of the method the forming FP stable state is of interest rather than the self-assembly process. The indicated states may conform to information storing and output generating as well as to recognition by means of FP.

Free energy is the most important biophysical characteristic of the bond between free proteins. At that the free energy level is inversely propor-

tional to proteins' bonding force. FP stable states are steady states that local minimum of the free energy corresponds to. It could be shown that formal protein can change to another steady state through binding with other FP (a so called *allosteric* effect). Allosteric effect allows FP binding with other FP, binding with was originally impossible. The resulting FP can enter the process of sequential binding. At that network of binding is being formed that is a sequence of bonds between FP comprising allosteric effect.

Cluster analysis based on immunocomputing is somewhat different of other methods. The major approach peculiarity is that an arbitrary cluster is considered as an information source to change the binding energy. Mathematically the given approach is based on the features of an arbitrary matrix singular decomposition over a field of real numbers. Used in IC mapping provides for a strict mathematical method for representing all clusters disregarding the attribute space dimension in 2D binding energy space. This plane could be called IC shape space. Clusters' representation in IC shape space allows their natural grouping into groups of close points. Classification can be made by experts in a mode of learning by instruction as well as by IC in a mode of learning without an instruction.

IC model of molecular recognition has two utterly important specific features, distinguishing it from the cluster analysis known methods. In the first place, the features' sets are not directly coded; they only assign binding energy between definite FP (FP-samples). In the second place, a similarity between these indicators' sets is determined by FP-samples recognition rather than by their comparison. Two formal proteins mutually recognize each other when they interact with the binding energy is less or equal to threshold value. It is possible to say that the less is the binding energy level the better is recognition. In a result the class being recognized is determined by FP-samples with minimal binding energy (the best recognition). The detailed substantiation of the used algorithm is given in [1].

3 Methods of Space Snapshots' Analysis

For a general case, the process of space snapshots analysis consists in three stages: raster *segmentation*, i.e. image subdividing into comprising parts or objects; *description* of the parts (objects) separated on the raster; raster *classification* of the parts (objects) based on their descriptions, i.e. immediate recognition.

3.1 Classic Methods of Raster Segmentation

3.1.1 Detection of Lines and Brightness Jumps

Let us define a mask as dimension matrix $n \times n$ over a field of nonnegative integer numbers. At use of such a mask response in every image point is given by the following formula:

$$R = \sum_{i,j=1}^n w_{i,j} \times z_{i,j}$$

where $z_{i,j}$ – brightness value of a pixel, corresponding the mask $w_{i,j}$ coefficient. The mask's response is assigned to a position of its central element. The difference in coordinates of adjacent mask's weights equals one. The above mask is used to detect the lines, and the line is considered detected if $|R| \geq T$, where T – is nonnegative threshold.

Though line detection plays a significant role in certain segmentation problem cases, the detection of brightness jumps presents a much more general approach to detection of interpretable discontinuities on the brightness picture.

Ideal contour jump is a set of connected pixels, where each one is located next to a rectangular brightness jump (i.e. next to such a jump that is not entailed by a gradual change of adjacent pixels brightness, and brightness at once changes from maximal to minimal value). In practice due to optical constraints, digitization, and imperfection of other elements in images registration system the fuzzy brightness jumps are arrived at.

Value of the first derivative can be used to detect the brightness jump in each point. The second derivative sign allows determining whether a pixel located at the jump lies at its light or dark side. Often a *property of zero level crossing* is used: an imaginable straight line that connects maximal positive and negative values of the second derivative close to the jump crosses the zero level approximately in the middle of brightness jump.

The image point is a *jump point* if its first derivative exceeds some given threshold.

Extended brightness jump is called *a contour*. The term *contour section* is usually used when the contour size is small as compared with the image size. One of the tasks emerging at segmentation is to assemble the contour's sections into longer contours.

In the image the first order derivatives are calculated via gradient. Laplacian can be used to obtain the second order derivatives.

3.1.2 *Contours Binding*

Ideally the above given methods in the image are only supposed to isolate the pixels located on the contours. Practically, due to various noises, contours discontinuities caused by lighting heterogeneity and other effects distorting the brightness picture continuity, this pixels' set rarely defines the contour sharply. This is why contour detection algorithms are as a rule completed by binding procedures in order to form substantial contours out of contour's points set. Contours binding algorithms are divided into local and global ones. Algorithms of the second type when binding the contours' sections use information about the whole raster; particularly, they incorporate the methods based on Hough's transform and graph theory. These algorithms are notable for multi-stage analysis and realization complexity.

At the usage of local processing one of the simplest approaches to point binding consists in analyzing pixels characteristics in a limited vicinity (say, 3×3 or 5×5) of each image point (x, y) that was marked as a contour point by one of the above considered methods. All points that are similar according to some preset criteria are bound and form a contour, composed of pixels matching meeting these criteria.

At such an analysis the following two major parameters are used to set the contour's pixels similarity: operator's gradient response value and gradient vector direction. The pixel in a given vicinity is united with a central pixel (x, y) if the value and direction similarity criteria are met. This process is being repeated at each image point and concurrent memorizing of the binding pixels found at the vicinity's center moving takes place.

3.1.3 *Threshold processing*

Suppose that there exists a histogram of monochromic image brightness $f(x, y)$. The obvious technique of pixels separation comprises a selection of threshold T value distinguishing brightness' distribution. The threshold transform can be considered as an operation in whose presence a comparison with the following function T occurs:

$$T = T(x, y, p(x, y), f)$$

where f – image, and $p(x, y)$ – some local characteristics of image point (x, y) , for instance, a mean brightness in a vicinity having its center in the above point. Image $g(x, y)$, obtained as a threshold transform result is determined as follows:

$$g(x, y) = \begin{cases} 1 & T < f(x, y) \\ 0 & f(x, y) \leq T \end{cases}$$

If value of T only depends on f , i.e. is the same for all image points, the threshold is called *global*.

Let us to study some other segmentation methods.

3.1.4 Areas Raising

As follows from the title *areas raising* is a procedure that groups pixels or raster subareas into bigger areas based on preset criteria (on the basis of *segmentation predicate*). Major approach consists in taking at the beginning a certain number of points standing as “crystallization centers”, and then building areas up them by joining to each center those neighboring pixels whose properties are close to crystallization center (e.g., their brightness fall within certain range).

Morphological watersheds segmentation

Watershed notion is based upon image representation in a form of 3D surface that is defined by two space coordinates and brightness level standing for terrain (relief) height. Let us assume that in every local minimum a hole is punched, thus, causing filling the terrain by water uniformly arriving from the below punched holes, so the water level is the same all over the terrain. When the rising water in two neighboring pools is close to junction, the junction preventing partition is placed between pools. Eventually, the water rising reaches the level when only the partitions’ edges are seen above the water. These partitions corresponding to watersheds lines form the continuous boundaries, singled out by watersheds segmentation algorithm.

3.2 Classic Segmentation Methods' Comparative Analysis

Method	Advantages	Disadvantages
Brightness jumps detection	Explicit use of binding notion, threshold and distance measures for objects' bounds integration.	Multistage processing model (significant realization complexity)
Threshold processing	Comparative simplicity of realization The method is often used in tasks of image processing.	A pronounced unimodality of most satellite snapshots' histograms is revealed that negatively affects the method's function. Binding notion is not used in the method.
Areas' raising	Uses binding notion. The method is well appropriate for solving problems of images with multimodal histograms classification.	Strong dependence of a learning sample is revealed.
Watersheds segmentation	High robustness. It is allowed to introduce additional constraints form a data base. Explicit use of binding components.	High realization complexity. Direct application usually causes a redundant segmentation.

4 VI Use of Immunocomputing Method to Analyze Satellites' Snapshots

4.1 IC Classification Algorithms

Pattern recognition

Let *pattern* be defined as n -dimensional vector column $X = [x_1, \dots, x_n]^T$, where x_1, \dots, x_n are real numbers and " T " – is a matrix transpose symbol. *Pattern recognition* is defined as representation $f(X) \rightarrow \{1, \dots, c\}$ of any pattern X , where integers $1, \dots, c$ represent *classes*.

Pattern recognition problem can be defined as follows:

It is given:

- Number of classes c ;
- Set of m learning patterns: X_1, \dots, X_m ;
- Class of any learning pattern: $f(X_1) = c_1, \dots, f(X_m) = c_m$;
- Arbitrary n -dimensional vector P .

Arrive to:

Vector class $P: f(P) = ?$

Learning

1. Form a learning matrix $A = [X_1 \dots X_m]^T$ of $m \times n$ dimension.
2. Calculate a maximal singular number s , as well as left and right singular vectors L and R of the learning matrix based on the following iterative (*evolution*) scheme:

$$L_{(0)} = [1 \dots 1]^T, \quad R^T = L^T A, \quad R_{(k)} = R / |R|,$$

where $|R| = \sqrt{r_1^2 + \dots r_n^2}$,

$$L = AR_{(k)}, \quad L_{(k)} = L / |L|, \text{ where } |L| = \sqrt{l_1^2 + \dots l_m^2},$$

$$S_{(k)} = L_{(k)}^T A R_{(k)}, \quad k = 1, 2, \dots,$$

$$\text{Until } |S_{(k)} - S_{(k-1)}| < \varepsilon, \quad s = S_{(k)}, \quad L = L_{(k)}, \quad R = R_{(k)}.$$

3. Store a singular number s .
4. Store a right singular vector R (as “*antibody-sample*”).
5. For each $i = 1, \dots, m$ store a component l_i of a left singular vector L (as a *cell of FIS*) and class c_i corresponding to learning image X_i .

Basic recognition algorithm

6. For any n -dimensional pattern calculate its binding energy with

$$R : w(P) = P^T R / s$$

7. Select l_i that has a minimal distance (maximal relation) to w :

$$\min_i |w - l_i|, \quad i = 1, \dots, m.$$

Items 6 and 7 coincide with a basic algorithm

8. Consider class c_i a sought class of pattern P .

Modified recognition algorithm.

9. Calculate P_i :

$$P_i = \frac{|-l_i + w|}{M}$$

10. Calculate values t_i for all $i = 1..m$ and find their minimal one.

Here $0 \leq \beta_i \leq 1$. - is a class weight set by an expert¹.

$$t_i = P_i(1 - \beta_i)$$

11. Consider class c_i a sought class of pattern P , and $1 - P_i$ – probability of object's belonging to class c_i . Compare a probability of object's belonging to a class with statistical threshold set by an expert. Object is considered recognized if value $1 - P_i$ is *not less* than threshold.

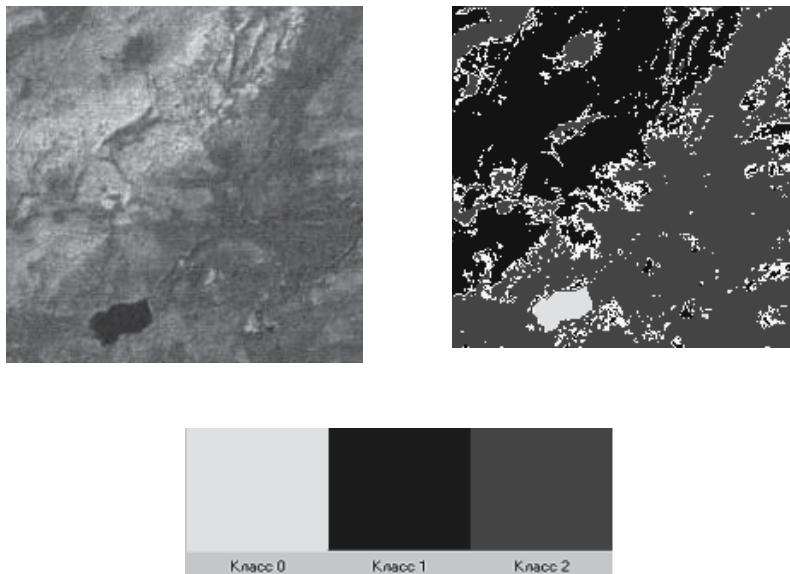
	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
0	1	1	1	1	1	1	1	1	1	4	4
0,1	1	1	1	1	1	1	1	1	1	4	4
0,2	1	1	1	1	1	1	1	1	1	4	4
0,3	1	1	1	1	1	1	1	1	1	2	4
0,4	1	1	1	1	1	1	1	1	1	2	4
0,5	1	1	1	1	1	1	1	1	1	1	4
0,6	1	1	1	1	1	1	1	1	1	1	4
0,7	0	0	0	1	1	1	1	1	1	1	4
0,8	0	0	0	0	0	1	1	1	1	1	4
0,9	0	0	0	0	0	0	0	0	1	1	4
1	0	0	0	0	0	0	0	0	0	0	4

¹ Here a weight equal to 0 means that the corresponding class is as much important as it used to be important under the absence of weight coefficients, and setting a feature weight equal to 1 will result in all objects belonging to this class (or to one of such classes when several classes exist).

4.2 Algorithm Implementation

Earth's surface areas classification.

An aggregate of Landsat satellite snapshots having resolution equal to 200 by 200 points (pixels) was used, and three classes were separated in it. The aggregate comprised six snapshots belonging to different spectral ranges², though to the same Earth surface area. Then a four pixels learning sample was formed, and experts identified belonging of each pixel to a certain class.



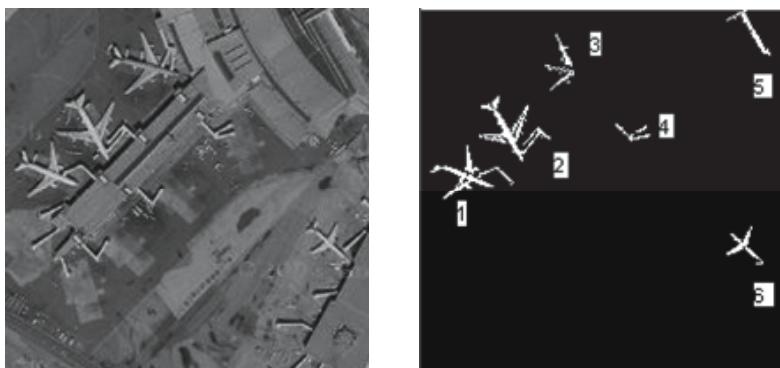
The color correspondence to a cluster was set by the given scheme. At that a number of points in a sample uniformly distributed along all fragmentation classes. The procedure of a learning sample forming was repeated for each image four times. The obtained sequence of four recognition results was analyzed by a program-analyzer. If pixel class was identical on all four images it was considered correctly recognized, otherwise the pixel was ascribed to ambiguously recognized ones and highlighted yellow. The basic recognition algorithm was used in the analysis. An original image and recognition result are given here. Three fragmenta-

² Visible blue ($0,45 \div 0,52$ micron), visible green ($0,52 \div 0,60$ micron), visible red ($0,63 \div 0,69$ micron), neighboring IC range ($0,76 \div 0,90$ micron), mean IC range ($1,55 \div 1,75$ micron), heat IC ($10,4 \div 12,5$ micron).

tion classes were considered (Class 0 –water surface, class 1- rocks, class 3 – vegetation).

4.3 Object's Recognition by IC Means

The image being segmented into areas, the obtained pixels aggregates are as a rule described and represented suitably for further computer processing. In the given case a signature based description was used. Signature is a description of the object's bound by 1D function that could be differently arrived at. The method used in the current work comprises building dependency of a distance between the area's centroid and the object's bound in a form of an angle function.



Original monochromic image within pixels' brightness range $0 \div 255$ (200 by 200 pixels) and image obtained as a result of segmentation performed by threshold processing with a global threshold equal to 150^3 . The objects are numbered for sake of description convenience.

It should be here noted that one of the airplanes disappeared from the image (in the left upper corner). The above took place because the vanished airplane was *not composed of one binding component*. It is possible to keep the airplane contour on the raster (e.g., by decreasing threshold at threshold processing), however, this may cause significant changes in other areas contours and squares, thus, making an effective classification of obtained areas impossible⁴.

³ Connected components including less than 40 pixels are removed from the image. This does not affect the recognition result and increases the method functioning speed.

⁴ It should be accounted for that the sought object is considered composed of one bound area. If the sought object were represented as an aggregate of bound areas (and this is introduced by "Meta-classes" method, see below) the recognition quality would be improved.

Center coordinates (as a simple mean of all area points' coordinates) were calculated and signature having a center in the obtained point was built for all remaining *binding components*. The built vectors were composed of 36 elements – values of signature radius. In case of detecting several intersections between the signature radius and area bound, accounted for was only the one that maximal radius angle corresponded to. Two objects were selected as a learning sample (and IC was accordingly seeking for two classes on the image): the most left airplane on the raster and a line segment shaped area in the raster upper part. The modified algorithm was chosen for the recognition process, and the recognition results could be seen in the here represented table. In the table lines all weights of “airplane” class are being changed, in its columns - of “all the rest” class; thresholds in all cases are equal 0. Totally after segmentation six areas are separated on the raster and three of them are airplanes. One of the airplanes is a part of the learning sample and, consequently, is not subjected to recognition. So, IC is supposed to detect on the raster two airplanes markedly differing in sizes and shapes (*but not in space orientation*). On the other hand the “all the rest” class is introduced artificially, and this class areas are not much alike. This is why the faultless recognition is reached at a significant increase of “all the rest” class weight, up to 0,9 and decrease of “airplane” class weight, 0,3 ÷ 0,4.

5 Isomorphism and Meta-classes Methods

5.1 Major notions

Monochromatic digital image can be represented by a matrix whose each element is an image element or a pixel. Numerical value of the element within a certain range determines the pixel's brightness level. .

Definition *Monochromatic raster* – is $m \times m$ brightness square matrix over a field of positive integers whose each element is a numeric value of the brightness level of the digital image's corresponding pixel. Matrix elements values fall within 0 to M range. Pixels are denoted by lowercase Roman letters: p, r, q . Each pixel has two coordinates: x and y . Notations « p » and « (x_p, y_p) » - are equivalent. Denote a set of all raster matrix elements as S .

Definition. Cardinality of set S – is a number of contained elements that is denoted by $|S|$.

Definition. $\forall p(x_p, y_p), q(x_q, y_q) \in S$ distance is determined as:

$$\|p, q\| = \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2}$$

Definition. $\forall p(x_p, y_p) \in S$ four closest neighbors p are determined as:

$$N_4(p) = \{q(x_q, y_q) \mid \|q, p\| = 1\}$$

Definition. $\forall p(x_p, y_p) \in S$ four diagonal neighbors p is determined as:

$$N_D(p) = \{q(x_q, y_q) \mid \|q, p\| = \sqrt{2}\}$$

Definition. Eight neighbors $N_8(p)$ are determined as:

$$N_8(p) = N_4(p) \cup N_D(p)$$

If the point lies on the image bound then some of its neighbors are beyond the image.

Definition. Let V – be a certain brightness values set from the range of image pixels' brightness values. *Contiguity* can be determined: Two pixels $p, q \in V$ called contiguous if $q \in N_8(p)$.

Definition. Two pixels' subsets S_1 and S_2 are called contiguous if $\exists q \in S_1 \wedge \exists p \in S_2, q$ and p are contiguous.

Definition. *Discontinuous track (curved)* from pixel p with coordinates (x, y) to pixel q with coordinates (s, t) is a non-recurrent sequence of pixels with coordinates:

$$K(p, q) = (x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$$

Where $(x_0, y_0) = (x, y), (x_n, y_n) = (s, t)$ and pixels $r_i(x_i, y_i)$ and $r_{i+1}(x_{i+1}, y_{i+1})$ are contiguous when $1 \leq i \leq n$. In such a case n is called the *track length K*.

Definition. Let S – be a certain image elements subset. Its two elements p and q are called *bound in S*, if there exists a track between them only composed of subset S elements:

$$p \in S \wedge q \in S \wedge \exists(p_1, \dots, p_n) \in S, \text{ such that}$$

$$\forall i = 1, \dots, n, p_i \in K(p, q)$$

For any pixel p from S a set of all pixels bound with it in S is called *binding component* (or *component of bound*) S :

$$R(p) = \{q \mid \exists K(p, q)\}$$

Definition. $C_p(x_c, y_c)$ – center $R(p)$:

$$x_c = \frac{\sum_{i=1}^{|R(p)|} x_i}{|R(p)|}, \quad y_c = \frac{\sum_{i=1}^{|R(p)|} y_i}{|R(p)|}.$$

Definition. If set S contains only one component it is called a *connected set* $S \in S' \Leftrightarrow \forall p \in S \exists! R(p)$

5.2 “Metaclasses” Method

5.2.1 Pilot Analysis

Performed over a segmented raster.

1. Define set V as $p(x, y) \in V \Leftrightarrow g(x, y) = 1$. All pixels to be further considered belong to V – set of “white” pixels of bicolored raster.

2. Define set $A = \{R(p) | |R(p)| \geq Z_{\max} \wedge |R(p)| \leq Z_{\min}\}$. Values Z_{\min}, Z_{\max} are set by an expert.
3. $\forall R(p) \in A$ calculate value $C_p(x_c, y_c)$.
4. Define set $Q(p)$ as

$$q(x, y) \in Q(p) \Leftrightarrow (|x_q - x_c| \leq Size/2) \wedge (|y_q - y_c| \leq Size/2)$$
5. Define set $E(p) = \{R(q) | q \in Q(p)\}$ and its subset $E'(p) = \{E(p) | |E(p)| \leq MaxComp\}$. $E(p)$ elements is a finite number of connected domains, located in immediate vicinity.
6. For all elements of $E'(p)$ set – building a signature for all areas included in aggregation.

5.2.2. Expedient Geometric Relations Analysis

Set of signature vectors S_i composed of St dimension signature's vectors built for each set $E'(p)$ element is the result of the previous stage. Further the following is implemented for each signature:

1. Normalization: $Sign_i = Sign_i / \max(Sign_1, \dots, Sign_{St}), i = 1..St$. Here $Sign_i$ – i -th signature element.
2. For each signature – calculate a number of maxima⁵:

$$Maxs_{Sign_i} = \{(i, Sign_i) | Sign_i > Sign_{i-1} \wedge Sign_i > Sign_{i+1}\}$$
3. If $\exists i, j \in Maxs \wedge (i - j) < MaxAngle$, then in the $Maxs$ set substitute elements $(i, Sign_i)$ and $(j, Sign_j)$ for element $(k, Sign_k)$, at that $k = (i + j)/2$, $Sign_k = (Sign_i + Sign_j)$. The indicated elements correspond to an airplane's tail unit; substitution operation is performed once.
4. Define sets:

$$Ansv = \{Sign | |Maxs_{Sign}| = U, \exists i, j \in Maxs_{Sign}, |i - (360 - j)| < F\}$$

⁵ Here a case when maximal $Sign_i$ value falls on $i = 0$ or $i = St$ was not considered, however, the program realization was accounting for such an option.

For an airplane $U = 4$, F – an allowable deviation of hull elements from one axis (degrees).

5.

$$\begin{aligned} Ansv' = \{Sign \mid Sign \in Ansv \wedge \exists k, l \in Maxs_{Sign}, \\ Z_1 < (k+l)/2 < Z_2 \wedge k \neq l \pm 1\} \end{aligned}$$

Here Z_1 and Z_2 – are minimal and maximal values of mean wings length against a hull. Set $Ansv'$ elements are considered the recognized exemplars of “airplane” class.

5.3 “Isomorphism” Method

Within a given approach a pattern of a sought object (airplane) is set as brightness matrix (*object of DB patterns*), and detection is performed by comparison of a monochromatic raster area with a pattern’s brightness matrix. Brightness matrix contains elements of only two types – 0 and 1 (“black” and “white”). Matrices of DB object brightness (MO) and of recognized object are always square. A pilot analysis in the here considered method is performed based on the algorithm used in “Metaclasses” method.

5.3.1 Direct Recognition

1. Define set $MS_{Sign} = \max(Sign_1, \dots, Sign_{St}), i = 1..n$ for each signature $Sign$.
2. Forming set $R_{Sign} = \{(x_c, y_c, MS_{Sign})\}$, consisting of ordered triples - 2 coordinates of cluster center on raster and maximal signature element for a given cluster.
3. In DB a sought object is transformed to a size of studied object. At that each coordinate of object matrix element is multiplied by MS/MP and matrix dimension is reduced by a factor of MS/MP, where MP - MO size (side length). The multiplication result is rounded off to integers. $MO' = \{(x, y) \in Z \mid x = x * MS / MP\}$.

4. The following subset is separated from raster pixels
 $L = \{(x, y) | |x - x_c| \leq MS, |y - y_c| \leq MS\}$
5. Comparison function f_c is calculated. If $f_c > L_v$, the object is considered found, otherwise –unrecognized. Here L_v – threshold set by an expert. f_c and L_v belong to $[0,1]$.

Value returned by a comparison is calculated by the following algorithm:

Let R – be DB object raster pixels. Then
 $A_\alpha = \{p | p = q \wedge p \in L, q \in R\}$

Here the pixels' equality is understood as the equality of their brightness numerical values.

$$f_c = \max_{\alpha} (|A_\alpha|) / |R|, \alpha = 1, \dots, 360.$$

At that the matrix turn relatively to center⁶: is performed:

$$\begin{aligned} x_q &= x_0 + x \times MS / MP / 2 \times \cos(\alpha) - y \times MS / 2 \times \sin(\alpha), x_q = \\ &= x + MS / 2 \times \sin(\alpha) + y \times MS / MP / 2 \times \cos(\alpha) \end{aligned}$$

6 Methods Implementation Results

Parameters used in a pilot analysis, “Metaclasses” method as well as results of “Isomorphism” method implementation are given in the table below. At implementation of “Metaclasses” method the objects 1, 2, 3 and 6 were recognized as airplanes, i.e. percentage of errors for a given raster equals 0.

In “Isomorphism” method same parameters as at implementation of “Metaclasses” method were used for segmentation and pilot analysis. As a pattern of a sought object was selected a priori enlarged object #2 cleared of elements not presenting airplane parts. Recognition results are given in the table ($L_v = 0,6$). Percentage of errors for a given raster also equals 0.

⁶ Thus, a maximal value of a comparison function cannot accede an area of a circle ratio to circumadjacent square, that is equal to one fourth of π , or approximately 0.78539816.

Pilot analysis parameters	
Parameter	Value
Segmentation threshold	150
Zmin	40
Zmax	40000
Size	7
MaxComp	3

Results of “Isomorphism” method implementation				“Metaclasses” method parameters	
area#	Value f_c	Object is recognized as airplane	Object is an airplane	Parameter	Value
1	0,7356	Yes	Yes	Step of extremum ⁷ search algorithm	0,1
2	0,73469	Yes	Yes	Z_1	0,3
3	0,76620	Yes	Yes	Z_2	0,9
5 ⁸	0	No	No	F	20
6	0,64201	Yes	Yes	St ⁹	36
4	0,58884	No	No		

7 Conclusion

All above represented and considered methods reveal their applicability to ERSD problems solving.

IC technologies when implemented to classification of terrain types allow natural representing of monochromic raster aggregate in a form of features' vectors set. However, at their use a complexity of learning sample forming is revealed. The method's advantage is its deeply substantiated theory [1].

When applied to the objects' search IC technologies effectively differentiate objects against their shapes. At that no necessity arises to describe a sought object explicitly. Although, the algorithm used to describe area on

⁷ A search method with a constant pitch was used to find an extremum.

⁸ For area 5 a comparison function value equals 0, for in the adopted algorithm at object's projection to a point (181, 11) and following scaling a part of circle describing object appears to be beyond the analyzed raster.

⁹ A number of signature elements was reduced to correctly recognize one of the objects.

the segmented raster strongly depends on the engaged segmentation methods efficiency. The above disadvantage can be subdued through object's representation as a metaclass (an aggregate of connected domains) and this was done hereby at this method application. A possibility to set class weights and receive for each object a numeric value for a degree of recognition reliability allow using the method even when certain class objects are characterized by weak similarity or in the presence of alike objects in different classes.

The results' strong dependence on the typical representative selected for sought objects' class characterizes "Isomorphism" method. At that the search of objects looking like definite airplanes are sought for rather than search of a "generalized airplane" is performed. The method possesses no capacity for a complex abstracting. Object is being described as an area (in our case it is a circle), whose shape in a general case does not coincide with the shape of original object, and this is not always appropriate. On the other hand, a possibility to define a search object directly is an advantage for some cases, say, when one definite object is a search target. Besides, it is easy to obtain additional information about the sought object: space orientation and dimensions against the selected typical class representative.

"Metaclasses" method in its current state at the analysis only accounts for the outer bound shape of the analyzed areas cluster on the segmented raster, that is insufficient for certain cases. More over, the sought object is to be analytically described and this may be too complex. However, the method is really seeking for definite type objects independently of a definite class representative, thus, allowing avoid casual errors due to wrong selection of a typical object. Additionally a possibility exists to set explicitly comparison tolerances, thus, adjusting the method to specific of the recognized raster.

References

1. Tarakanov, A. O., Skormin, V. A., Sokolova, S. P. Immunocomputing: Principles and Applications. Springer, New York, 2003.
2. Стюарт Рассел, Питер Норвиг. Искусственный интеллект: современный подход. «Вильямс», Москва, 2006.
3. Р. Гонсалес, Р.Вудс. Цифровая обработка изображений. Техносфера, Москва, 2006.
4. В.Е. Гершензон. Дистанционное зондирование Земли: общие проблемы и российская специфика.
<http://www.scanex.ru/ru/publications/pdf/publication1.pdf>.

5. Ч. Монделло, Дж. Ф. Хепнер, Р.А. Вильямсон Рынок данных ДЗЗ в мире.
<http://www.scanex.ru/ru/publications/pdf/publication5.pdf>.
6. Чубукова И. А. Data Mining.
<http://www.intuit.ru/department/database/datamining/>.
7. Popovich, V.V., Potapichev, S.N., Sorokin, R.P., Pankin, A.V. Intelligent GIS for Monitoring Systems Development.// Proceedings of CORP2005, February 22-25, 2005, University of Technology Vienna.

Context-Driven Information Fusion for Operational Decision Making in Humanitarian Logistics

Alexander Smirnov, Alexey Kashevnik, Tatiana Levashova, Nikolay Shilov

St.Petersburg Institute for Informatics and Automation of the
Russian Academy of Sciences (SPIIRAS),
39, 14 linia, St.Petersburg, 199178, Russia
{smir, alexey, oleg, nick}@iias.spb.su

Abstract. The paper describes an approach to information fusion based on the context management aimed at supporting decision making for humanitarian logistics operations. Application of GIS as one of the major information sources is discussed in detail. The paper describes an application of knowledge logistics as an intelligent information fusion service for creation efficient routing plans (as one of the major logistics tasks in virtual supply network management) under given constraints and preferences. The implementation is based on the concept of open services in a distributed environment of a networked organization. This application is illustrated via a case study of delivering hospital supplies to the disaster site.

Keywords: decision making support, ontology, context, context versioning, humanitarian logistics.

1 Introduction

User-centric decision support is of high importance for disaster relief & evacuation operations where multiple participants have to collaborate most efficiently, and the coalition is often unstable (participants may leave and come continuously). The quality of decision making depends upon the quality of information at hand. Problems with information (outdated, incomplete, unreliable, etc.) form a major constraint in decision making. The practice shows that one of the most difficult steps in responding to such situations is providing for the right relief supplies to the people in need at the right time. At the same time delivering of too much supplies or wrong supplies means losing time and money [1]. Therefore, humanitarian logistics standing for *processes and systems involved in mobilizing people, resources, skills and knowledge to help vulnerable people affected by natural disasters and complex emergencies*, is a central issue for disaster relief [2].

Logistics systems play an important role in humanitarian operations. An intelligent decision support based on the technology of knowledge manage-

ment may significantly enhance the logistics system capabilities (e.g., reduce costs and time of supplies delivery). Evidently, one of the key information sources for such a system is a geographical information system (GIS) providing for such information as region geography, available transportation means (roads, their characteristics and availability), existing infrastructure and facilities, navigation-related information, etc.

The paper describes an approach to information fusion based on context management supporting decision making for humanitarian logistics operations. GIS is used in the approach as one of the major information sources. Based on the information provided by GIS, creation of efficient routing plans (as one of the major logistics tasks in disaster relief and evacuation operations) under given constraints and preferences is performed. The implementation is based on the concept of open services in a distributed environment of a networked operation. This application is illustrated via a case study from the above mentioned domain of disaster relief and evacuation.

The methodology presented proposes integration of environmental information and domain knowledge in a context of current situation. It is done through linkage of representation of this knowledge with semantic models of information sources providing for information about the environment. The methodology (Fig. 1) considers context as a problem model based on the knowledge extracted from the application domain and formalized within an application ontology by a set of constraints. The set of constraints, additionally to the constraints describing domain knowledge, includes information about the environment and various preferences (user defined constraints) of the user concerning the problem solving. Humanitarian logistics as a coalition operation assumes different user roles. The methodology considers different user's roles as different levels of user's responsibility. The problem is suggested to be modeled by two types of contexts: abstract and operational. *Abstract context* is an ontology-based model integrating the problem relevant information and knowledge. *Operational context* is an instantiation of the abstract context with data provided by the information sources.

The paper is structured as follows. The following sections 2-4 describe the scenario of the proposed fusion-based decision support; consider all the stages of the user request processing as well as the context versioning as one of the core technologies in fusion-based decision support systems (DSS). The reason for this is that changing environmental conditions should be taken into account at the moment of decision making. For example, availability of the transportation routes, possibilities to use certain transport, etc. should be taken into account by GIS involved into the decision support. Structure of the problem can also be altered by the changes in the environment. In this case it is necessary to maintain consistency of the contexts and information sources for their reuse and definition of missing information and knowledge.

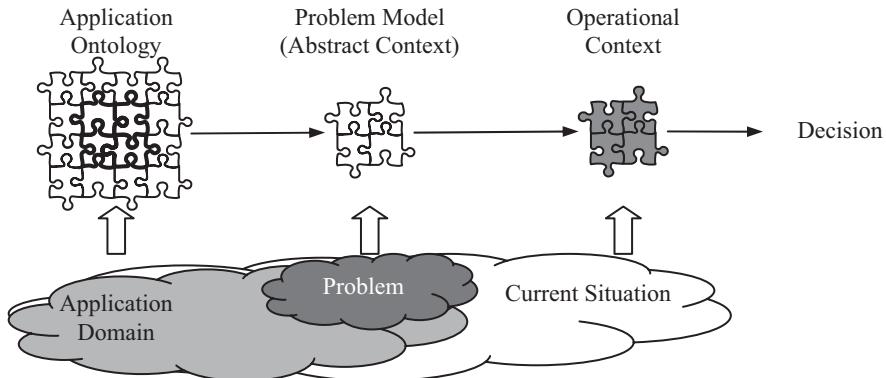


Fig. 1. Context-based decision support.

Finally an illustrative case study of disaster relief and evacuation is presented. It involves GIS as one of the major sources of information for decision making support. Section 5 incorporates major conclusions the authors arrived at.

2 General Scenario of the Fusion-Based Decision Support

The overall scenario of context-driven information fusion in the presented decision support system (DSS) is shown in Fig. 2. Below, a detailed description of the scenario is presented.

Before a request can be processed an application ontology describing the problem the request is related as one to have been built (4). The application ontology combines the domain knowledge described in domain ontology (3) and problem solving knowledge described in the task and methods ontology (2). The domain ontology, in turn, is built based on existing ontologies describing related domains (1). At these stages fusion of structural knowledge and problem solving knowledge is performed. At the moment this is proposed to be done by experts supported by previously developed ontology management environment [3].

Decision making in dynamic domains is characterized by a necessity to dynamically process and integrate information from heterogeneous sources and to provide the user with context-driven assistance for the analysis of the current situation. Systems of context-driven decision making support are based on usage of information / data, documents, knowledge and models for problem identification and solving. This requires an access to electronic documents complementing the domain knowledge.

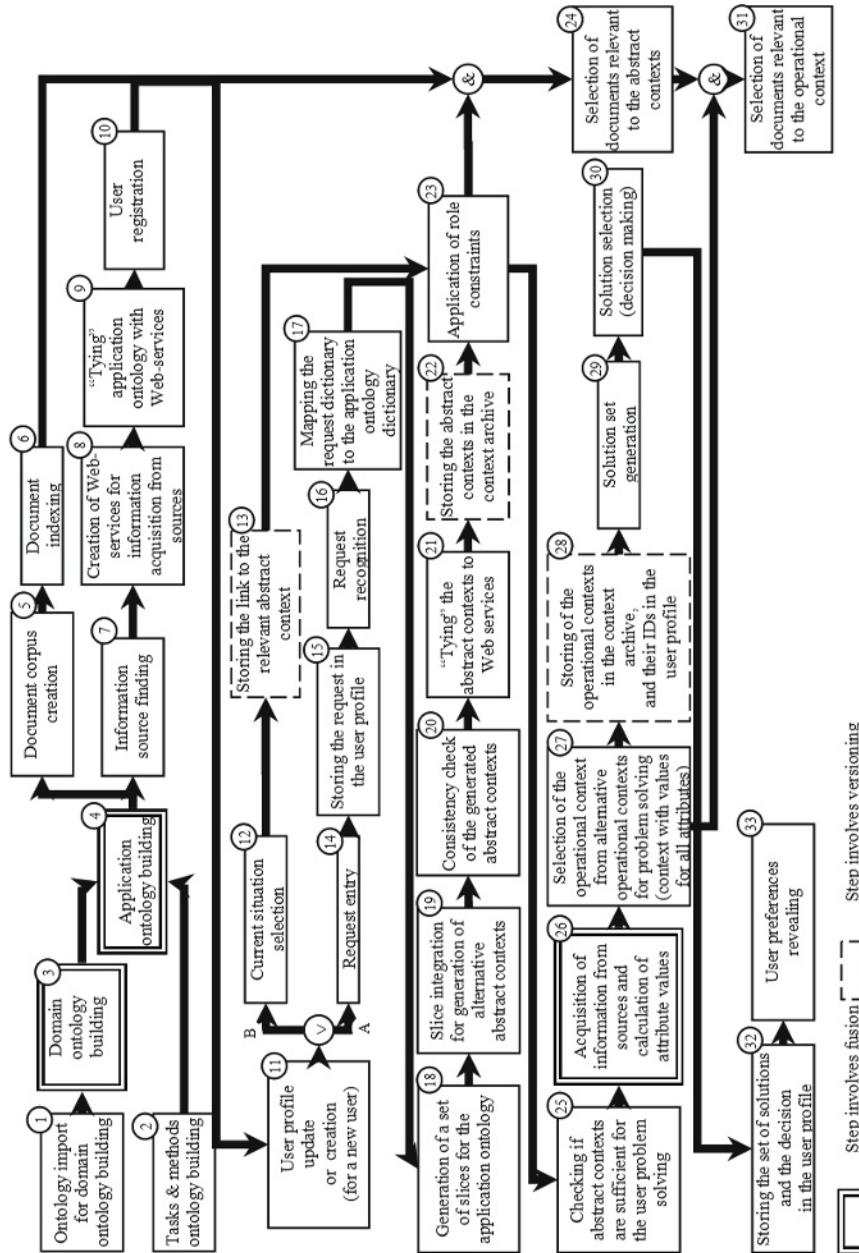


Fig. 2. System operation scenario.

In accordance with the developed methodology [4] a search for documents relevant to the user's request requires preparation of an index of the formed corpus of documents (5). The indexing (6) is based on the application ontology and takes into account its classes, attributes and relationships between them.

Another necessary prerequisite for the system operation is its connection to information sources. It consists of three operations: finding information sources (7), connection of the found sources to the system using Web-services (8), and assigning information source references to attributes of the application ontology, in other words, defining what sources the application ontology attributes take values from (9).

Once the above operations are completed the DSS is considered ready for request's processing. Personalization is one of the important features of operational decision making support. For this reason users are described in the system via user profiles and associated with different roles that set certain limitations to provide only information that is useful for a particular user.

After registration in the system (10) the user profile is updated or created if it did not exist before (11). After that depending on the user role and available information the user either inputs a request (case "A") or selects a situation (case "B"). In case "B" the abstract context is assumed to exist and is being activated (12, 13). In order to ensure usability of the previously created abstract context the context versioning technique described in the next section is used. In case "A" few more steps are required. The request (14) stored in the profile for further analysis (15) is recognized by the system (16), and recognized meaningful terms are mapped to the vocabulary of the application ontology; the vocabulary includes class names, attribute names, string domains, etc. (17). Based on the identified application ontology elements and developed algorithm a set of slices relevant to the request is built (18). Based on these slices a set of abstract contexts is generated (19). Since some of the problem solving methods imported from the task & methods ontology can be alternative, integration of such slices leads to a set of alternative abstract contexts. The generated abstract contexts are checked for consistency (20), their attributes are assigned information sources based on the information from the application ontology (21), and it is saved in the context archive (22). The context archive together with the context versioning enables reuse of the previously created context to facilitate the information fusion process.

Starting this point both cases "A" and "B" are treated the same way with the only difference that in the case "B" there is only one abstract context and there is no selection from the alternative contexts (27).

First, role constraints are applied to the generated abstract context to remove information that is not interesting or not supposed to be seen by the user (23). After this a search for documents relevant to the abstract context can be done (24).

The abstract contexts are checked if it is enough for solving the current user problem (25). Then the required information is acquired from the appropriate sources and calculations are performed (26). At this step the fusion of parametric knowledge is performed. Due to the chosen notation of object-oriented constraint networks a compatible constraint solver can be used for this purpose. The constraint solver analyses information acquired from sources and produces a set of feasible solutions eliminating contradictory information.

The above operations result in forming operational contexts for each abstract context. From the operational contexts one with values for all attributes is selected (27). This operational context is to be used at the later stages of the request processing. It is also used for selection of documents relevant to the particular user request or current situation (31). All the operational contexts are stored in the context archive, and references to them (their IDs) are stored in the user profile (28). Application of context versioning techniques enables the reuse of operational contexts if the problem structure remains the same (the abstract context is the same), but parameters change.

Based on the identified operational context the set of solutions is generated (29). From the generated solutions the user selects the most appropriate one – makes a decision (30). The solutions and the final decision are stored in the user profile for further analysis (32). Stored in the profile information can be used for various purposes including audit of user activities, estimation of user skills in certain areas, etc. In the presented research this information is used for revealing tacit user preferences (33).

As it can be seen the context versioning is one of the main enablers of the knowledge fusion. The following section describes the versioning technique used in detail.

3 Context Versioning Model

In the approach a context is produced based on the application ontology, that is, the approach deals with ontology-based context modeling. Hence both ontology versioning and context versioning techniques are proposed for consideration in the developed context versioning model. The model aims at tracking changes in the application ontology, contexts, and information sources and at keeping overall consistency. Relations considered in the model are: (1) between succeeding revisions of the application ontology; (2) between versions of the application ontology, abstract contexts, and the information sources linked to them; and (3) between versions of abstract and operational contexts.

The model of context versioning includes the following components of the decision support system (DSS): application ontology, abstract context, opera-

tional context, and representations of information sources. These components are interrelated as shown in Fig. 3. Versions of the application ontology are stored in the ontology library, versions of abstract contexts, operational contexts, and conditions of information sources at the moment of decision making are stored in the context archive.

Since the context versioning model is developed for decision support purposes, relations between context versions, a set of solutions generated based on these contexts, and decisions made by the user based on these solutions are proposed to be taken into consideration in the versioning model. Due to this alternative solutions proposed by DSS for the environmental conditions at the moment of decision making will be accessible for analysis [5, 6]. E.g., if the problem has no solutions in the current situation, examining versions of the contexts of this problem would help to clarify what information source conditions were for the moment the problem had been solved at [7, 8]. The sets of solutions and decisions are not components of versioning, that is, no versions of them are created and therefore, operations included in the context versioning model (e.g., change propagation) do not affect them.

Versions of the application ontology are created when an ontology engineer modifies it. Since each abstract context is related to the application ontology, creation of a new version of the application ontology causes creation of a new version of the abstract contexts. Versions of the operational context related to the abstract context are created when (1) modifications of the abstract context concern addition or removal of links to information sources; (2) these modifications involve changes of attribute domains that do not necessitate disconnection of the linked information sources; and (3) each time DSS generates a set of solutions. Modifications of the application ontology, and respectively abstract contexts, dealing with any other changes are considered as resulting in a new operational context that is not a version of the operational context related to the modified abstract context. Information source conditions are saved when DSS generates a set of solutions.

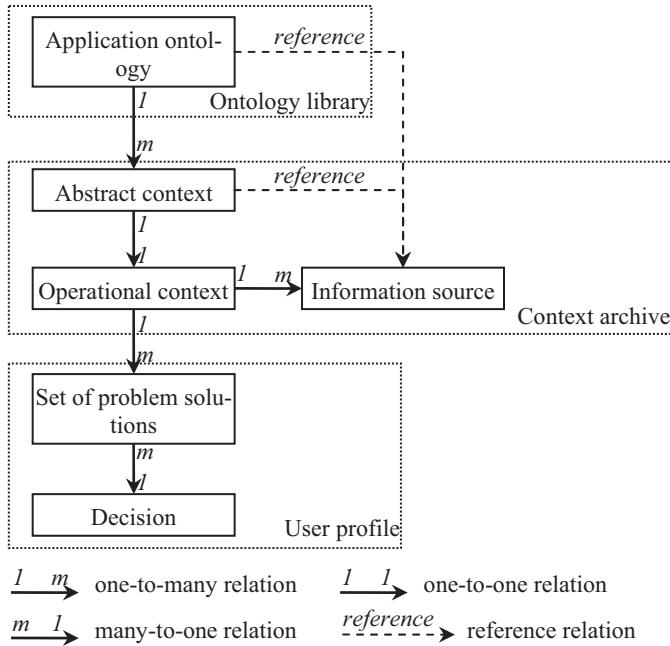


Fig. 3. Structure of the DSS repository.

Change propagation aims at determination of what is to be changed if something has changed and keeping consistency between the DSS components. The general schema of change propagation is as follows: (1) create a new version of the DSS component that has changed; (2) create new versions of the DSS components the changed component relates to; (3) propagate changes to the new component versions; and (4) check overall consistency.

Within this research, changes to be propagated can be caused by (1) modifications of the application ontology in cases of its revision and (2) replacement / addition / removal of information sources related to the application ontology and contexts. Although abstract and operational contexts are reusable they cannot be modified independently on the application ontology modifications. The abstract contexts are modified through propagation of changes from the application ontology. The operational contexts are modified through propagation of changes from abstract contexts they are produced from.

Propagation of changes depends on the kind of modifications (Table 1). The modifications of the application ontology not concerning the information sources cause propagation of the changes to the abstract contexts related to it. Any changes concerning information sources are propagated to all components of DSS the information sources are related to. All kinds of modifica-

tions concern operational contexts since they contain the application ontology components and deal with values provided by information sources.

Table 1. Change propagation for DSS' components.

No.	Application ontology modifications	Information source	Abstract / operational context
1.	Removal / addition of links to information sources	+	+
2.	Modifications of attribute domains	+	+
3.	Removal / addition of classes and attributes	+	+
4.	Modifications of functional constraints that are responsible for retrieval and conversion of values from the information sources	+	+
5.	Removal / addition of “part-of”, associative constraints, and constraints on class compatibilities	-	+
6.	Modifications of functional constraints that are responsible for correspondence between class attributes of the domain ontology constituent and input / output arguments of methods in the constituent of tasks & methods ontology ¹	-	+

Table 1 uses elements of the notation of object-oriented constraint networks used in the here presented research [9]. According to this notation an ontology is represented by a set of classes; a set of class attributes; a set of attribute domains; and a set of constraints. The set of constraints comprises (1) taxonomical (“is-a”) relationships, (2) hierarchical (“part-of”) relationships, (3) class cardinality restriction, (4) class compatibilities, (5) associative relationships, and (6) functional relations.

Changes are propagated to a new version of the operational context only if the application ontology modifications concern items 1 and 2 in Table 1. In other cases a new operational context accounting for all the modifications is created so that the new version of abstract context relates to this new operational context. This allows profiling techniques to differentiate the solutions generated for the same problem with different data and environmental conditions from the solutions generated for a modified problem or a different problem.

Consistency checking determines whether a version of a DSS component can be used for problem solving and in what way.

¹ Application ontology integrates knowledge from domain ontology and tasks & methods ontology. Parts of the application ontology corresponding to knowledge from these ontologies are referred as constituents of domain ontology and tasks & methods ontology.

Consistency between versions of the application ontology and abstract context, and the information sources linked to them determines if a new version of the application ontology and abstract context can use the information sources used previously. Links between the new versions and information sources that these versions can use on are kept. To distinguish that the information sources previously used are out of use in this version these links are indicated. In case of a sensor failure this sensor can be replaced by any sensor consistent with the current version of the abstract context. That is, data values in the operational context can be provided by a sensor previously used.

Consistency between versions of the application ontology and abstract contexts shows if the versions of abstract context specify instances in the same manner as the new version. Versions of the abstract contexts that specify instances differently than the new version can be reused for tasks that do not require the renewed specification.

Consistency between versions of abstract and operational contexts determines if the attribute values in a version of the operational context are consistent with their specification in the abstract context. Inconsistency means that the version of operational context holding inconsistent data may not be used for inference of missing information (knowledge). The values that are inconsistent with the new specification are indicated. In case of missing data they cannot be derived from the indicated data.

If modifications of abstract context result in creation of a new operational context (not a version of operational context) then consistency between the existing versions of the operational context related to this abstract context and the new operational context is not checked.

Since state-of-art operational decision making faces problems of management and sharing of huge amount of knowledge a personalization for decision maker support is required. This will allow DSS to reduce the amount of information to be analyzed by the decision maker and to present this information in accordance with the decision maker's preferences. To reduce the amount of information user roles are introduced within the research. One of the purposes of this is to divide a problem so that the users involved in solving it would be provided for information that is relevant to a particular user.

For each role its level of responsibility is specified. The levels show for what part of the set of tasks the user filling the role is allowed to make decisions. The levels are represented as attributes of the classes corresponding to the roles. Domains of the attributes are specified as enumerable. A list of values specified for the domains defines possible values for levels of responsibility.

An ontology for roles has been imported from The Component Library developed by Knowledge Systems Group, UT-Austin² and adapted for the re-

² <http://www.cs.utexas.edu/users/mfkb/RKF/clib.html>

search purposes. Roles concerning participation of persons in an event are included in class “Job-Role” of the ontology. As a result of the adaptation each role dealing with decision making is assigned a set of tasks pertinent to it (a set of decisions usually made by one filling this role). The set of tasks corresponds to the set of tasks formalized within the application ontology.

The level of responsibility for the user accessing DSS is instantiated in the operational context and in that way constrains the set of tasks formalized in this context. The constraint solver generates a set of solutions only for the tasks that the user filling a certain role deals with.

Operational context in some cases is supposed to be lack of information & knowledge. This issue is known as missing information & knowledge. In this research the issue is thought of as a lack of attribute values. Missing knowledge here can be caused by an unavailability or failure of information sources. In this case versions of the operational context are queried for the purpose of revealing values that were used in the versions. This method is more or less reliable only for context-insensitive attributes (e.g., databases). For context-sensitive attributes (time, location ...) only probable values can be derived. If operational context is a new one (i.e. there are no versions of this context) versions of information sources that are responsible for the attributes with missing data are queried. This is implemented as queries to the Web-services responsible for the interaction with information sources.

4 Information Fusion for Humanitarian Logistics: Case Study

In the presented case study the application ontology is assumed to exist. In the implemented case study three user roles are considered: dispatcher, decision maker and medical / firefighting brigade leader (brigade leader). The dispatcher submits a request (Fig. 2, case “A”) to generate an abstract context (Fig. 2, steps 10-23). The decision maker is working with the generated abstract context (Fig. 2, case “B”, steps 10-31). User preference revealing is not shown.

At the first stage the dispatcher enters a request about a disaster (in the case study fire and accident are considered as possible disasters). He / she chooses the situation type, enters its location, potential number of victims and additional description. These parameters identify the abstract context and set constraints on some of the variables. The abstract context based on this request is built and saved in the context archive. An appropriate task is created for the decision maker.

When the decision maker logs in the system he / she can see current tasks and their statuses. When an unsolved task is activated the abstract context is

filled with information from the sources thus an operational context is built. The information includes: (i) road network of the region provided by a GIS, (ii) current weather conditions provided by a weather service, (iii) available medical and firefighting brigades, (iv) their locations provided by the brigades themselves (e.g., using GPS), as well as (v) hospitals and their current capacities. Based on this information the following calculations are preformed: (i) calculation of the number of required firefighting brigades, (ii) availability of roads (some of them can be flooded in case of rain), and (iii) limitations on usage of helicopters (e.g., in case of a strong wind). These calculations are a part of the fusion process. At this stage some of the solutions that are not feasible are eliminated (e.g., solutions with helicopters are eliminated in case of strong wind).

The current situation is presented to the decision maker (Fig. 4). He / she can add additional constraints (e.g., optimization by time) and launch the solution generation. At this stage the following calculations are performed: (i) search for the shortest route for each brigade, (ii) search for the shortest routes to hospitals, (iii) creation of the evacuation schedule meeting the requirements set by the decision maker. This is the second part of the knowledge fusion process when feasible solutions are generated based on information from different sources and available problem-solving methods (e.g., shortest route calculation).

The solutions are presented to the decision maker (an example solution is shown in Fig. 5). Based on the final decision the system forms tasks for the brigades. Brigade leaders can either accept or decline the tasks. In the latter case the task is considered to be unresolved and is passed to the decision maker for making another decision.

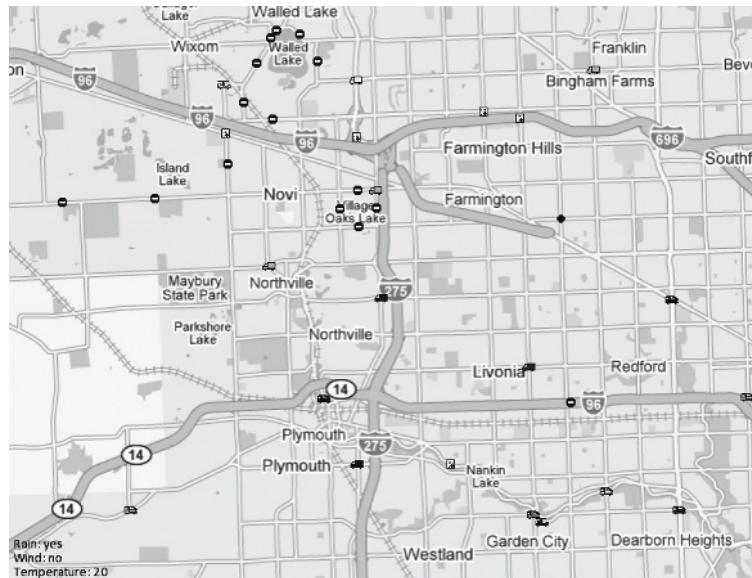


Fig. 4. Current situation presented to the decision maker.

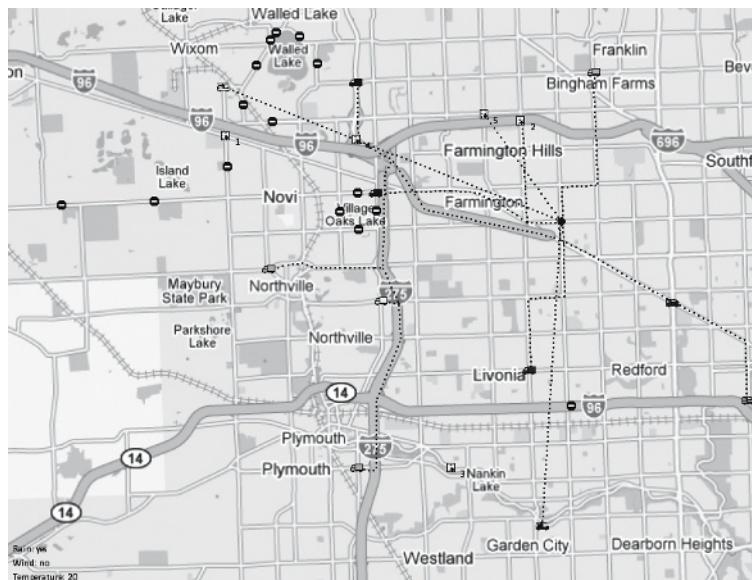


Fig. 5. Solution presented to the decision maker.

5 Conclusion

The paper describes an approach to information fusion based on usage of ontologies and contexts for knowledge representation. The problem is described via application ontology, built in a result of fusion of structural (from domain ontologies) and problem solving (from task & methods ontologies) knowledge. Application of contexts allows reducing the problem domain in accordance with the task being considered, user role and current situation. Due to usage of the notation of object-oriented constraint networks, fusion of parametric knowledge / information from different sources is achieved by applying constraint satisfaction and propagation technologies implemented in various solver engines. It is also shown that GIS is one of the key information sources in the area of humanitarian logistics. Application of the developed approach is illustrated via a case study of disaster relief and evacuation.

5 Acknowledgements

The paper is due to the research carried out as a part of CRDF partner project # RUM2-1554-ST-05 with US ONR and US AFRL, projects funded by grants # 05-01-00151 and # 06-07-89242 of the Russian Foundation for Basic Research, as well as projects # 16.2.35 of the research program "Mathematical Modelling and Intelligent Systems", # 1.9 of the research program "Fundamental Basics of Information Technologies and Computer Systems" of the Russian Academy of Sciences, # 78 of the scientific program of St.Petersburg Scientific Center, and # 26/06 of the Science and High School Committee of St.Petersburg.

References

1. Humanitarian Logistics: Getting the Right Relief to the Right People at the Right Time, Fact Sheets, Fritz Institute, (2005). URL: http://www.fritzinstitute.org/-fact_sheets/f_s-hls.html.
2. Scott, P., Rogova, G.: Crisis management in a Data Fusion Synthetic Task Environment. Proc. of the 7th Conf. on Multisource Information Fusion / Fusion 2004 (2004).
3. Smirnov A., Pashkin M., Chilov N., Levashova T.: "Web-DESO": Ontology Management System. Int. Conf. on Intelligent Information Technology / ICIIT'2002 (2002) 485-494.
4. Smirnov A. V., Levashova T. V., Pashkin M. P., Shilov N. G., Krizhanovsky A. A., Kashevnik A. M., Komarova A. S.: Context-sensitive access to e-document corpus. Proc. of Int. Conf. Corpus Linguistics 2006 (2006).

5. Brézillon P.: Context in Problem Solving: a Survey. *The Knowledge Engineering Review*, Vol. 14, No. 1 (1999) 1-34.
6. Robinson R.: Context Management in Mobile Environments. PhD. Honours Thesis, School of Information Technology, University of Queen Island, Australia, October 2000.
URL: <http://www.rickyrobinson.id.au/university/honours/thesis.doc>
7. Fox M. S, Lin J., Gupta L.: Knowledge Network: An Information Repository with Services for Managing Concurrent Engineering Design; Technical Report, Enterprise Integration Laboratory, Department of Industrial Engineering, University of Toronto, Toronto, Canada M5S1A4 (1996)
URL: <http://www.eil.utoronto.ca/kbd/papers/lin-wetice96.pdf>.
8. Griffiths J., Millard D.E., Davis H., Michaelides D.T., Weal M.J.: Reconciling Versioning and Context in Hypermedia Structure Servers. Proc. of Metainformatics International Symposium / MIS 2002 (2002) 118-131. URL: <http://eprints.ecs.soton.ac.uk/6829/01/mis02.pdf>.
9. Smirnov A., Pashkin M., Levashova T., Chilov N.: Ontology-based support for semantic interoperability between SCM and PLM. *Int. J. of Product Lifecycle Management*, Vol. 1, No. 3, Inderscience Enterprises Ltd. (2006) 289-302.

From Battle Management Language (BML) to Automatic Information Fusion

Ulrich Schade, Joachim Biermann, Miłosław Frey, Kellyn Kruger

FGAN-FKIE, Neuenahrer Straße 20, 53340 Wachtberg-Werthhoven, Germany

Abstract. Current operations, military as well as disaster relief operations, are executed by coalition forces which have to exchange information effectively. This exchange must be supported by a smart architecture of the underlying network and its systems and by a specific language format used to formulate military communications, i.e., orders, requests and especially reports. Such a language format has to support automatic processing of the communication as well as automatic information fusion. Otherwise, the forces' headquarters would only accumulate huge piles of data without any chance to analyze them in time, i.e., quickly enough to exploit those fleeting opportunities permitted by the enemy or by circumstance.

The language format, we propose is Battle Management Language (BML). A Battle Management Language is an unambiguous language to be used for communications among C2-systems – systems to support the military process of command and control –, their users, simulation systems and robotic forces. BML expressions can be processed automatically by parsers as defined in the field of computational linguistics. The output of the parsers are so-called “feature-value matrices” in which the information is represented in a XML-like structure. This paper will point out and illustrate by example how the feature-value representation of BML reports can be exploited for automated information fusion.

1 Introduction/ Motivation

“[Information is] the foundation of all our ideas and actions.” ([6] Clausewitz 1832/1968, p. 162)

Current military research aims at the exploitation of Information Age concepts and technologies under the key term “Network Centric Warfare (NCW)” [1], [2]. It is assumed that deployed forces are connected and that they exchange information automatically without effort. Besides, the exploitation of information has to be supported. In order to prevent information overload and in order to ensure that commanders get the important pieces of information quickly enough to exploit those promising but fleeting opportunities permitted by enemy’s fault or by pure chance, informa-

tion has to be pre-processed automatically by the system. This includes automatic information fusion because the most interesting pieces of information often can only be appreciated fully in connection with other pieces of information.

2 Information Fusion

A system component for automatic information fusion has to perform parts of the tasks to be done in intelligence production. In military organizations intelligence is obtained during a structured and systematic series of operations called “Intelligence Cycle” [3]. It includes the four steps Direction, Collection, Processing, and Dissemination which are defined by the NATO Glossary of Terms and Definitions (AAP-6). Processing is where the information that has been collected in response to the direction of the commander is converted into intelligence products. It is a structural series of operations, defined as the production of intelligence through Collation, Evaluation, Analysis, Integration, and Interpretation of information and/or other intelligence. The system described in this paper focuses on support of information exploitation and fusion within the Collation, Analysis, and Integration steps. The military understanding of these processing steps is kept in the Allied Joint Intelligence Counter Intelligence and Security Doctrine (AJP 2.0) and in the NATO Glossary of Terms and Definitions (AAP6).

In general, the incoming information consists of all kinds of reports (from reconnaissance assets as well as from human sources) about observations of activities, military sites, equipment, and others. It is the task of the intelligence processes to aggregate the categorized and correlated individual situation elements given by the incoming reports and to fuse them into more complex elements, thereby creating new objects in the domain of concern. Sensor data fusion identifies objects and tracks by applying mathematical methods to underlying physical models. In contrast to this, information fusion is based on models that describe actions and behaviors. In principle, we can differentiate between two kind of models, behavioral models and normalcy models. Behavioral models describe (complex) actions other factions of interest might undertake conducting their aggressive or hostile activities. In our case, these actions are military tasks executed by the faction in question. Incoming sequences of reports about the faction are compared to the models in order to identify the faction’s current behavior and to predict its future behavior. In addition to the behavioral models, there are normalcy models that are needed to detect deviations from the

normal and changes in a faction's behavior, as well as errors in our assumptions about the behavior of this faction. The role of such models can be illustrated by the following example. Eight hours after initiating the ground campaign "Desert Storm", General Schwarzkopf received the information that the Iraqis had destroyed the desalination plant of Kuwait City. He then, correctly, noticed that this is not the normal behavior of occupation troops, but the beginning of a major withdrawal operation: The Iraqis were pulling out of the city [17 p. 453].

In order to exploit behavior models and to fuse information, the information first has to be available. With respect to automatic information fusion, this also means that the information must exist in a format which allows automatic processing. Reports made by humans normally do not exist in such a format. To circumvent this, BML should be used.

3 Battle Management Language (BML)

Battle Management Language (BML) is being developed as an open standard in cooperation between a study group initiated by SISO (Simulation Interoperability Standards Organization) and a NATO RTO modeling and simulation group. Both are called "Coalition Battle Management Language" and share members in order to enhance the cooperation [4]. The goal of BML is to unambiguously specify Command and Control information and Command and Control communication. To allow the specification of communication, it is defined as a language, and in order to cover the C2 domain it is based on doctrine and aligned to Coalition standards. In addition, specific BML extensions are under development. These extensions – e.g., geoBML for terrain reasoning [7] – specify communication about those domains that are highly connected to Command and Control, but focus on a different topic.

Originally BML was driven by the demand that an officer should be enabled to feed orders directly into a decision support system. Simulated forces could then execute these orders, and the system could report the results back to the officer. Testing different orders against simulation results would help to improve the officer's decisions. Furthermore, in HQ training a staff could send orders to simulated forces which would report back to the HQ. In these scenarios, BML is seen as a language which enables officers to communicate with simulated forces. In a broader perspective [4], BML is a language for C2 communication in general. C2 systems (as well as their operators) communicate to other C2 systems, to simulation systems, and to robotic forces via BML (cf. figure 1).

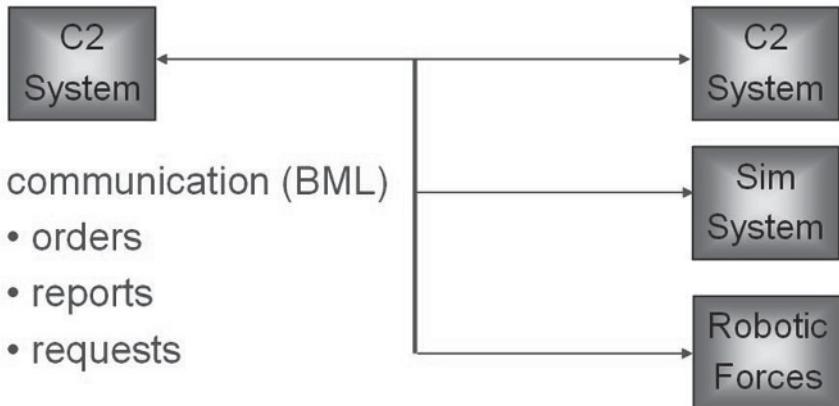


Fig. 1. BML is a language for communication among C2 systems, simulation systems, and future robotic forces

BML expressions have to be processed by systems. At the same time BML also has to be expressive enough to cover the C2 domain. In principle, this is a dilemma. Automatic processability reduces expressiveness and vice versa. Significantly, the dilemma can be reduced if the language in question is defined by a formal grammar. Such a grammar consists of a finite set of (well-defined) words and a finite set of production rules which determine how words can be concatenated to expressions. Schade and Hieb [14, 15] present such a grammar for BML orders and to some extent for BML reports as such. A more complete view on the grammar for BML reports is given by Schade and Hieb [16]. The grammars complement each other. They share lexical items as well as production rules. The lexical items and their meanings are taken from the standard data model “Command and Control Information Exchange Data Model” (C2IEDM) (cf. the MIP website) in order to ensure a common understanding of these C2 terms. The production rules of the grammars are built in accordance with Lexical Functional Grammar [9, 5]. Thus, a BML expression can be transformed first into a syntactic tree and then into a functional structure. In our system, the transformation is executed by a standard bottom-up shift-reduce parser. The resulting functional structures are so-called feature-value matrices which are standard within the field of computational linguistics. These matrices have many beneficial properties. First, under-specified information can be represented, which is important for fusion

processing. Second, the matrices obey XML schemata which allows further automatic processing and is therefore proposed as a BML requirement by Hieb, Tolk, Sudnikovich and Pullen [8]. Third, the matrices allow unification, a standard computational linguistic algorithm for merging information which we regard as also being beneficial in information fusion [13].

4 Unification

In this section, unification is introduced. Unification is a standard algorithm in computational linguistics. Mathematically, it is based on lattice theory. The first subsection therefore will give a definition of unification as it holds from the mathematical view. In the second subsection, we will show how unification is applied to information fusion.

4.1 The Mathematical Background of Unification

From the mathematical view, a feature-value matrix is a finite set of pairs. Each pair consists of a feature and a value. A feature is always a symbol, a value, however, can be a symbol or a feature-value matrix itself. In addition, the following uniqueness condition holds: Every feature has a unique value. In other words, in a matrix there cannot be two pairs which share the feature but not the value. However, different features may have the same value.

Unification is an operation on the set of feature-value matrices. In order to unify matrices A and B, one has to check all the feature-value pairs in A to see if any of its features is also the feature of a pair in B and vice versa. Say, we check the pair (feature_{aj}: value_{aj}) of A. If there is no pair (feature_{aj}: value_{bk}) in B, the pair (feature_{aj}: value_{aj}) will be an element of the unification result. The same holds for the pairs of B which have features which do not occur in pairs of A. If there is a pair (feature_{aj}: value_{bk}) in B, however, the pair (feature_{aj}, Ujk) is element of the unification result where Ujk is the result of unifying value_{aj} and value_{bk}. Under the mathematical view, a unification fails if one tries to unify a symbol with a matrix or with another non-identical symbol. If any subordinate unification fails, the main unification also fails.

4.2 Unification as Used for Information Fusion

In our system, reports written in BML are transformed into specific feature-value matrices represented in XML. The stream of reports is analyzed by trying to match the incoming reports to behavioral patterns. These patterns are stored in an ontology component of the system in order to model behavior. The ontology component also includes a taxonomy of C2 terms as well as other relations among these terms. Naturally, the terms are taken from C2IEDM's attributes and values and include the BML vocabulary.

A behavioral pattern consists of two parts, a sequence of actions and events and a result action which sums up the sequence. In the example given in the next section, the sequence in question consists of a disengage action and a move action which form a withdraw action as result. In other words, if a unit disengages and then moves away from the area of action it withdraws. Obviously, this statement includes conditions. For one, the unit has to move away from the opponent and not in the opponent's direction. Second, the move should start directly after disengaging. Third, it has to be the same unit which disengages and which moves (away from the fight). The last condition means that in order to fuse a reported disengage and a reported move to a withdraw, we have to check the conditions. In particular, we have to check whether the executers (agents in linguistic terminology) of the actions are identical. For the unification algorithm we want to apply, this means that we have to unify reported actions and events with actions and events of a behavioral pattern. The pattern then might demand a unification of parts of one action/event with parts of another action/event, e.g., the unification of some actions' executers.

To unify reported actions to pattern actions as well as the unification of different actions' parts during a fusion process cannot be as strict as the mathematical unification described above. In order to illustrate the problem, we take a look at the way BML denotes units. In BML, a unit as executer of an action is denoted by one of three kinds of expressions. First, there is denotation by name, e.g., “2./PzGrenBtl332” (second company of the infantry battalion 332). Denotation by name is clear without ambiguity. However, for some units reported, especially enemy units, the name might not be known. Second, there is denotation by unit type, e.g., “an infantry company”, and, third, there is denotation by material “four battle tanks (are approaching)”. If one and the same unit is denoted by different types of BML expressions (“2./PzGrenBtl332” vs. “an infantry company”) or by expressions differing in elaborateness (“four battle tanks” vs. “four Leopard2A5”), the resulting feature-value matrices also differ. Thus, pure mathematical unification of these matrices will fail in some cases. E.g., a

matrix representing “four battle tanks” will have the feature-value pair (**subtype**: `battle_tank`) whereas the matrix representing “four Leopard2A5” will have the pair (**subtype**: `leopard2A5`). Nevertheless, in our system, a unification is possible. It is done by ontological means. In the example, it can be derived from the ontology’s taxonomy that Leopard2A5 are battle tanks. E.g., with respect to feature subtype, two values of subtype can be unified if one value is a hyponym of the other. The resulting value is the hyponym. In the case of our example, Leopard2A5 is a battle tank. Thus, Leopard2A5 is a hyponym of `battle_tank`, and the unification is successful: (**subtype**: `leopard2A5`). Naturally, (**subtype**: `leopard2A5`) could not be unified to (**subtype**: `howitzer`).

5 Example

In this section, we will provide an example in order to illustrate how our system works. The example is based on a patrol scenario. Patrolling is a standard task for military units, especially for coalition forces in current coalition operations, and it is a task which becomes more and more important as war changes its face back to asymmetric warfare [10]. In our scenario, the patrol is run by a platoon in an urban environment. It consists of four armored wheeled vehicles. The patrol is ambushed by snipers that immobilized two of the vehicles and wound patrol members during the fire fight that follows. The battalion leader who commands and controls the patrol sends an armored company as relief and a UAV (Unmanned Aerial Vehicle) as reconnaissance to the back of the building where the sniper fire is coming from. The armored company joins the fire fight and tips the balance. The snipers cease fire. A few moments later the UAV reports that some individuals with weapons board a car and move away fast. At this point in time, information fusion should step in and draw the conclusion that the individuals with the weapons are most probably the snipers who withdraw. Based on this implication the battalion leader can order a light aircraft group to pursue the snipers and destroy them.

The information fusion process is triggered by incoming reports that match to parts of a schema. In the example, there are two critical reports. One is sent by the relief company; the other is sent by the UAV’s ground-station. The reports are given in the following in two variants, the (a)-variant displays the respective report in natural language and the (b)-variant in BML.

- (1) C2-Company to Battalion:
- (a) *Enemy ceases fire.*

- (b) disengage en C2 at XY at now fact;
- (2) UAV to Battalion:
- (a) *Four armed persons are leaving target building; they board a car and move away towards Z fast.*
- (b) move four suspect par (I1) from XY toward Z at now by car fast fact;

BML report (1b) is about the action disengage, which is C2IEDM's best approximation to express "ceasing fire". The name en for the executer of the disengage action is a label which had been introduced to the communication before and which refers to the enemy. The name C2 refers to the C2-Company, and the location statement XY denotes the building the sniper fire had come from. The report ends with a deictic temporal expression (at now) and a credibility statement (fact). BML report (2) is about the action move which sums up the leave and the move from the natural language report. The phrase four suspect par describes the executer of the move action. The term par is taken from C2IEDM's table "person-type-category-code". It denotes a "member of an irregular armed force" (C2IEDM, p. E-195). The whole phrase is automatically labelled by (11). The label can be used in subsequent reports. It expresses an anaphoral reference like "they" from the natural language version. In contrast to pronouns of natural languages, however, labels used in BML are unique in order to grant unambiguity. The BML reports are sent in their XML versions which represent feature-value matrices. The matrices are shown in figure 2.

<table border="1"> <thead> <tr> <th colspan="2">type: disengage</th></tr> </thead> <tbody> <tr> <td>executer:</td><td>name: en type: unit subtype: special arm-cat: sniper size: squad hostility: hostile</td></tr> <tr> <td>affected:</td><td>name: C2 type: unit subtype: combat arm_cat: inf size: coy hostility: friend</td></tr> <tr> <td>loc:</td><td>XY</td></tr> <tr> <td>start:</td><td>...</td></tr> <tr> <td>cred:</td><td>fact</td></tr> </tbody> </table>	type: disengage		executer:	name: en type: unit subtype: special arm-cat: sniper size: squad hostility: hostile	affected:	name: C2 type: unit subtype: combat arm_cat: inf size: coy hostility: friend	loc:	XY	start:	...	cred:	fact	<table border="1"> <thead> <tr> <th colspan="2">type: move</th></tr> </thead> <tbody> <tr> <td>executer:</td><td>label: I1 type: par count: 4 hostility: suspect</td></tr> <tr> <td>source:</td><td>XY</td></tr> <tr> <td>direction:</td><td>Z</td></tr> <tr> <td>start:</td><td>...</td></tr> <tr> <td>modifier:</td><td>instrument: car speed: fast</td></tr> <tr> <td>cred:</td><td>fact</td></tr> </tbody> </table>	type: move		executer:	label: I1 type: par count: 4 hostility: suspect	source:	XY	direction:	Z	start:	...	modifier:	instrument: car speed: fast	cred:	fact
type: disengage																											
executer:	name: en type: unit subtype: special arm-cat: sniper size: squad hostility: hostile																										
affected:	name: C2 type: unit subtype: combat arm_cat: inf size: coy hostility: friend																										
loc:	XY																										
start:	...																										
cred:	fact																										
type: move																											
executer:	label: I1 type: par count: 4 hostility: suspect																										
source:	XY																										
direction:	Z																										
start:	...																										
modifier:	instrument: car speed: fast																										
cred:	fact																										

Fig. 2. Feature-value matrices for the disengage report (left) and the move report (right)

<table border="1"> <tr><td>action:</td><td>type: disengage</td></tr> <tr><td> executer:</td><td>_1</td></tr> <tr><td> affected:</td><td>_2</td></tr> <tr><td> loc:</td><td>_3</td></tr> <tr><td> start:</td><td>_4</td></tr> </table> <table border="1"> <tr><td>action:</td><td>type: move</td></tr> <tr><td> executer:</td><td>_1</td></tr> <tr><td> source:</td><td>_3</td></tr> <tr><td> destination:</td><td>_6</td></tr> <tr><td> direction:</td><td>_7</td></tr> <tr><td> start:</td><td>_5</td></tr> <tr><td> mod:</td><td>_8</td></tr> </table>	action:	type: disengage	executer:	_1	affected:	_2	loc:	_3	start:	_4	action:	type: move	executer:	_1	source:	_3	destination:	_6	direction:	_7	start:	_5	mod:	_8	<table border="1"> <tr><td>action:</td><td>type: withdraw</td></tr> <tr><td> executer:</td><td>_1</td></tr> <tr><td> source:</td><td>_3</td></tr> <tr><td> destination:</td><td>_6</td></tr> <tr><td> direction:</td><td>_7</td></tr> <tr><td> start:</td><td>_4</td></tr> <tr><td> mod:</td><td>_8</td></tr> </table>	action:	type: withdraw	executer:	_1	source:	_3	destination:	_6	direction:	_7	start:	_4	mod:	_8
action:	type: disengage																																						
executer:	_1																																						
affected:	_2																																						
loc:	_3																																						
start:	_4																																						
action:	type: move																																						
executer:	_1																																						
source:	_3																																						
destination:	_6																																						
direction:	_7																																						
start:	_5																																						
mod:	_8																																						
action:	type: withdraw																																						
executer:	_1																																						
source:	_3																																						
destination:	_6																																						
direction:	_7																																						
start:	_4																																						
mod:	_8																																						

Fig. 3. The operational withdraw schema. On the left side there is a list of actions which members have to be unified with reported actions. On the right side is the resulting matrix. _1 to _8 are variables which are instantiated during the unification

Both matrices can be merged (unified) with the withdraw schema (figure 3). Even more, they can be merged into the same withdraw schema, because the following conditions hold: a) The timestamp of the move-report is close to the timestamp of the disengage-report; b) “four suspect par”, the actor of the move report, perfectly matches the actor of the disengage report labelled “en”. In particular, this match holds because of three subordinated matches, the matches about size/count, hostility, and type. The label “en” refers to the hostile snipers that have been attributed “size: squad; hostility: hostile; subtype: special”. Obviously, hostility as well as size of the two actors match. Hostility matches because suspect individuals can be hostile individuals. The pair “size: squad” matches the pair “count: four” due to a quite trivial ontological process. In addition, in the given scenario, enemy troops are supposed to be not only regular units but also paramilitary (or even terrorist) groups. Thus, “paramilitary”, the value of “type” analyzed from “four suspect par”, does not prevent the match.

The result of the match is inferred information about a withdraw action (cf. figure 4). The operator of the system now is asked to accept or to decline the result. If it is accepted, a credibility statement has to be added by the operator. In this case it might be “plausible”.

	type: withdraw	
executer:	name: en type: unit subtype: special arm-cat: sniper size: squad hostility: hostile	
source:	XY	
direction:	Z	
start:	...	
modifier:	instrument: car speed: fast	
cred:	plausible	

Fig. 4. The matrix of the inferred withdraw

After acceptance, the inferred information is stored in the data base. In addition, it can be distributed to other units. In our case, the information is made available to an air-force strike unit together with the commanding unit's order to pursue and to destroy the enemy:

- (3) commanding unit (TF) to air force strike unit (Cav):
- (a) *Enemy withdraws from XY towards Z by car fast;
pursue enemy in order to destroy.*
- (b) move en from XY toward Z at now by car fast fact;
pursue TF Cav en from XY toward Z begin at now in order to de-
stroy.

The last line of (3b) shows the pursue order. It starts with the tasker (**pursue**), the tasker (**TF**) – the one who orders the task –, the taskee (**Cav**) – the one ordered to execute the task – and the affected (**en**).

6 Conclusion

The fusion algorithm presented above is based on unification, a well-defined algorithm used in the field of computational linguistics. In order to fuse information automatically, the reports which bear the individual in-

formation pieces have to be represented in a structured way, in our case in feature-value matrices. This is the case if the reports are originally expressed in BML and parsed during the collation step of information fusion. The fusion itself is done by unifying sequences of reported actions and events with patterns that model behavior and activity. The patterns are part of the ontology component. They represent the knowledge that drives the fusion. In our system this knowledge is independent from the fusion algorithm. On the one hand, the algorithm is based on unification and the differences between our algorithm and the pure mathematical unification are defined in *general* terms. E.g., an atomic value of the features type or subtype can be unified to a different atomic value of the same type if and only if one of these two values is a hyponym of the other according to the ontology's taxonomy. On the other hand, explicit and specific military knowledge is represented in the behavior patterns and their schema representations. This division of work is intentional. It ensures that to adapt the information fusion process to new behavior of an enemy – which has to be done constantly in times when the face of war changes back to asymmetric warfare – it is sufficient to update the behavior patterns whereas the fusion algorithm remains untouched.

References

1. Alberts, D.S. & Hayes, R.E. (2003). Power to the Edge. Washington: CCRP.
2. Alberts, D.S. & Hayes, R.E. (2006). Understanding Command and Control. Washington: CCRP.
3. Biermann, J. (2006). Understanding Military Information Processing – An Approach To Supporting the Production of Intelligence in Defence and Security. In: Shahbazian E. & Rogova, G. (Eds.), NATO Science Series: Computer & Systems Sciences: Data Fusion Technologies on Harbour Protection. Amsterdam, NL: IOS Press.
4. Blais, C., Hieb, M.R. & Galvin, K. (2005). Coalition Battle Management Language (C-BML) Study Group Report. Paper 05F-SIW-041, Fall Simulation Interoperability Workshop, Orlando, FL, September 2005.
5. Bresnan, J. (2001). Lexical-Functional Syntax. Malden, MA: Blackwell.
6. Clausewitz, C. von (1832). Vom Kriege. Berlin: Dümmlers Verlag. Cited after Clausewitz, C. von, On war, edited by Rapoport, A. Baltimore, MD: Penguin Books, 1968.
7. Hieb, M.R., Powers, M.W., Pullen, J.M & Kleiner, M. (2006). A Geospatial Battle Management Language (geoBML) for Terrain Reasoning. Proceedings of the 11th International Command and Control Research and Technology Symposium, September 2006. Cambridge, UK.

8. Hieb, M.R., Tolk, A., Sudnikovich, W.P. & J.M. Pullen (2004). Developing Extensible Battle Management Language to Enable Coalition Interoperability. Paper 04E-SIW-064, Euro Simulation Interoperability Workshop, Edinburgh, UK, 2004.
9. Kaplan, R. & Bresnan, J. (1982). Lexical-Functional Grammar: A formal system for grammatical representation. In: Bresnan, J. (Ed.), *The Mental Representation of Grammatical Relations*. Cambridge, MA: MIT Press.
10. Münkler, H. (2006). *Der Wandel des Krieges*. Weilerswist: Velbrück.
11. MIP website: <http://www.mip-site.org>.
12. NATO Standardization Agency (NSA) (2003). AAP-6(2002) NATO Glossary of Terms and Definitions, <http://www.nato.int/docu/stanag/aap006/aap6.htm>.
13. Schade, U., Biermann, J. & Frey, M. (2005). Towards Automatic Threat Recognition. In: Proceedings of the NATO-RTO Specialists' Meeting on Information Fusion for Command Support (IST-055/RSM-001), Den Haag.
14. Schade, U. & Hieb, M.R. (2006a). Formalizing Battle Management Language: A Grammar for Specifying Orders. Paper 06S-SIW-068, Spring Simulation Interoperability Workshop, Huntsville, AL, April 2006.
15. Schade, U. & Hieb, M.R. (2006b). Development of Formal Grammars to Support Coalition Command and Control: A Battle Management Language for Orders, Requests, and Reports. Proceedings of the 11th International Command and Control Research and Technology Symposium, September 2006. Cambridge, UK.
16. Schade, U. & Hieb, M. (2007). Battle Management Language: A Grammar for Specifying Reports. Paper 07S-SIW-036, Spring Simulation Interoperability Workshop, Norfolk, VA, March 2007.
17. Schwarzkopf, H.N. & Petre, P. (1992). *It Doesn't Take a Hero*. New York: Bantam.
18. Shieber, S.M. (1986). An Introduction to Unification-Based Approaches to Grammar (= Volume 4 of CSLI Lecture Notes Series). Stanford, CA: Center for the Study of Language and Information.

Centrope MAP: Combining X-Border Information from Multiple Sources for Planning Purposes

Manfred Schrenk¹, Walter Pozarek²

¹ MULTIMEDIAPLAN.AT and CEIT ALANOVA – Central European Institute of Technology, Dept. for Urbanism, Transport, Environment and Information Society

² PGO – Planungsgemeinschaft Ost,
www.centropemap.org

Keywords: SDI, Spatial Data Infrastructure, Cross Border co-operation, Cross Border Information Systems, Spatial Planning, Monitoring, Centrope, INSPIRE, CENTROPE, Web Map Services, WMS, OGC, UMN MapServer, Mapbender

Summary

“CENTROPE” is the new common name for the cross-border-region in Central Europe consisting of parts of Czech Republic, Slovakia, Hungary and Austria and including the cities of Vienna, Bratislava, Trnava, Brno, Győr and Sopron. The CENTROPE region, strategically located in the heart of the “New Europe”, is one of the continent’s most dynamically developing regions.

Harmonised cross-border geo-information is essential to support political and economic decision-making. “Centrope Map” represents a cross-border Spatial Data Infrastructure for the region. It provides a framework for a common map representation of cross-border spatial data via OGC-compliant Web Map Services (WMS) originating from different organizations and institutions throughout the region.

Centrope Map as a Cross-Border Spatial Data Infrastructure (SDI) is seen as a precondition for a successful and common development of the CENTROPE-region as a whole and all it’s parts.

1 Introduction: The CENTROPE-Region

1.1 Geographical Situation

“CENTROPE” is the new common name for the cross-border-region in Central Europe consisting of parts of Czech Republic, Slovakia, Hungary and Austria. Official Centrope partners are the region South Moravia and the City of Brno (Czech Republic), the Regions and Cities of Bratislava and Trnava (Slovakia), Győr-Moson-Sopron County and the Cities of Győr and Sopron (Hungary), the federal states of Burgenland, Lower Austria, and Vienna as well as the Cities of Eisenstadt, St. Pölten, Vienna (Austria) (see Fig. 1).

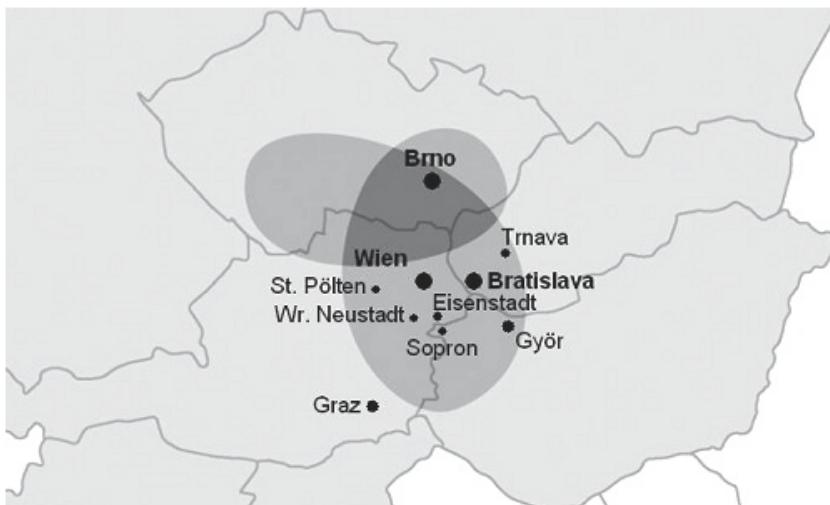


Fig. 1. Overview of Centrope region and partners

The geopolitical position of CENTROPE region is unique from many perspectives, some of the most important being the location in the geographical and in some respects also political center of Europe, positioned on the former dividing line of the “iron curtain” as well as the “Twin-City-situation” of Vienna and Bratislava that are both the capital cities of their countries and only 60 km apart from each other.

1.2 Political Framework

The official birth of CENTROPE was the joint signature of the “Kittsee Declaration” by the political leaders of the involved regions and cities in September 2003. The overall goal of this agreement was to create an internationally attractive and competitive common cross-border region that in economic, ecological and social terms ensures a high quality of life.

The Kittsee Declaration was followed by the “St.Pölten Memorandum” in April 2005 under the motto “We grow together. Together we grow”. It contains agreements on specific thematic areas such as the mutual support of consultations at the EU level, the development of joint proposals for sustainable, socially viable functioning in the cross-border labor market, the intensification of cooperation between existing institutions in order to integrate the cooperative activities in the region and more.

Meanwhile, a lot of thematic projects have been launched to support these political goals and, furthermore, the economic integration as well as the flourishing initiatives of NGOs.

1.3 The Change of Paradigms in Public Administration: “from Chess to Soccer”

To illustrate the fundamental paradigmatic changes that are necessary especially for public administrations to form a successful cross-border region let us illustrate it with the example of switching “from chess to soccer”:

Not so long ago, the tasks for public administrations were clearly defined both in form and content and in administrative boundaries and future development seemed predictable as a prolongation of the presence. It was like in chess: the fields were strictly divided, the rules and possible moves clearly defined, success was largely dependent on the position of the players. Co-operation across institutions, especially cross-border-co-operation, was very limited - it was not really necessary and almost impossible.

With a rapidly changing world – the fall of the iron curtain, the ongoing expansion of the EU, technological changes especially in information- and communication technologies – those structures became obsolete, so new tasks and rules had to (and still have to) be adopted: modern politics and public administration was becoming more like a soccer field. Players built teams, could run around freely, think outside their positions, and everyone was welcome to score. This new game definitely required a new approach including a new set of tools.

2 CENTROPE MAP – the Spatial Data Infrastructure for Centrope

The project “Centrope MAP” was launched by the “Austrian Planning Association East” (PGO), an association with extensive experience in co-operation across administrative boundaries, to support the Centrope-activities and to prepare for cross-border spatial planning and decision making.

It evolved out of the realization that harmonized cross-border geographic data sets are essential for a common development of the Centrope region. During the early project phase, heterogeneous geodata and related attributes of the involved regions were shared by exchanging data via email and CDs/DVDs, which of course proved inappropriate for teamwork on a regular basis. Accordingly, the project partners agreed on the need for a *“relevant base collection of technologies, policies, and institutional arrangements that facilitate the availability of and access to spatial data”* which also represents the definition of a Spatial Data Infrastructure (SDI) (Nebert / SDI Cookbook, 2004, p.8). An SDI consists of spatial data and related attributes, metadata which document specific data about spatial data, applications like catalogue and web mapping services to discover, visualize and evaluate the data as well as methods to offer access to geographic data. As a result, the expected outcome of the project “Centrope MAP” was an advanced Spatial Data Infrastructure that consisted of various services originating in the participating regions of Austria, Czech Republic, Hungary, and Slovakia.

A framework for a geographic information strategy at the European level was already provided by EUROGI, the European Umbrella Organization for Geographic Information, but there was still a lack of clear rules and policies concerning the interoperability and harmonization for geographic information (GI) and Geographic Information Systems (GIS). The debatable INSPIRE (INfrastructure for SPatial InfoRmation in Europe) directive aspired to provide a framework for coordinating, harmonizing, sharing and making accessible spatial information to support policy actions that directly or indirectly affect the environment. This framework should enforce the common chance of the EU member states to create a European Spatial Data Infrastructure (ESDI). *“The goal is an open, cooperative infrastructure for accessing and distributing information products and services online”* (Smits et al. 2002, p. 6). Such a European Spatial Data Infrastructure (ESDI) can only be developed through the synergies of smaller initiatives at regional and national levels. As Hecht (2002) argues, the aims of INSPIRE can only become a reality through the creation of a net-

work consisting of local and regional SDI initiatives. It's obvious that there is an urgent need to set up national as well as transnational SDI initiatives according to the INSPIRE principles.

“Centrope Map” represents such an initiative at regional level and becomes an instrument for the “new game of co-operation”: an Internet portal providing various spatial, economical, social and environmental data about the Centrope region visualized in an easy-to-understand web-mapping application. Its main goal is to become a decision-support tool for ongoing and future supra- and interregional projects.

Of course, Centrope Map did not start from the scratch, but evolved from existing well-established national and regional Geo-information-systems -systems that were typically limited in scope or too specific to their areas of research. For cross-border planning, projects and strategies input from a dozen of such sources is/was often needed, which can make the planning process complicated and time-consuming.

“Centrope Map” helps to overcome these issues by interlinking all these scattered national and regional systems into a unified platform that acts as a single source of information for the entire Centrope region.

“Centrope Map” does not produce geodata itself - instead, it collects, bundles and structures geodata from existing sources. Therefore, the quality of the “Centrope Map” data is highly dependent on the quality and accuracy of the source data. It is an open platform where partner systems provide their data, and might as well obtain data from other systems. All this data is then made freely available to the public as standard OGC-conform web maps services. Because of its evolving nature, it should not be considered as a single project, but as a cooperation process, that leads to a win-win situation for all partners involved and the Centrope region as a whole.

3 Centrope_Map Approach and Challenges

3.1 (dis)organisation

As mentioned previously, the regions involved in the project “Centrope MAP” are situated in four different countries, with various languages and diverse cultural backgrounds. For the fusion of geographic information across these regions there is a demand for clear rules and policies concerning the interoperability and harmonization for geographic information (GI) and Geographic Information Systems (GIS).

Therefore it has to be considered that an SDI becomes only useful when the organizational arrangements for the coordination and administration at local, regional, national, and transnational level are given (Nebert 2004). For sustainable decision making each of these levels requires the access, integration, and usage of geodata from various sources. A framework of policies, institutional agreements, technologies, data and people is necessary to enable an efficient exchange and use of geographic information (Craglia et al. 2003b). To support the interoperability of software, such a framework can only be established by combining international standards. According to the SDI Cookbook (Nebert 2004), the standards of most relevance to access components of Spatial Data Infrastructures include those from ISO/TC211, Open Geospatial Consortium® and Internet-related bodies including the World Wide Web consortium (W3C) and the Internet Engineering Task Force (IETF).

3.2 Multiple Languages

The regions that belong to the Centrope region are part of different countries with different official languages. These are Czech, German, Hungarian, and Slovakian.

The European Environment Information and Observation Network (EIONET) has also focused on a problem with specific terms of various European languages. For that reason, the EIONET has established a general multilingual environmental thesaurus (GEMET thesaurus) as a reference indexing and retrieval tool for the European Environment Agency (EEA) catalogue of data sources (GEMET thesaurus 2006). The GEMET thesaurus provides specific vocabulary in the fields of environment and ecology.

For the “Centrope Map” application, there is also the need for such a multilingual dictionary. Moreover, using digital spatial data originating from different countries in common computer systems often causes character encoding problems. All the various languages spoken in the Centrope region tend to use a different encoding standard for the representation of their national (mostly accented) characters (e.g. ISO-8859-2 for Czech or Hungarian datasets, ISO-8859-1 for German datasets). This can lead to unexpected results for the end user of the information system (e.g. incorrect search results). However, advanced conversion algorithms and unified character sets (UTF-8) will eventually help to resolve these issues.

3.3 Interoperability

The geodata for a common map representation of the Centrope region are derived from different data sources and servers hosted by different institutions using various languages and standards. As already mentioned before, this implies a framework of standards and tools. General regulations increase the advantage of the use of already existing data coming from different sources for the future development of SDIs, e.g. by a stronger cooperation concerning research and development. Fig. 2 shows the complexity of such a framework (for a more precise explanation see Nebert 2004).

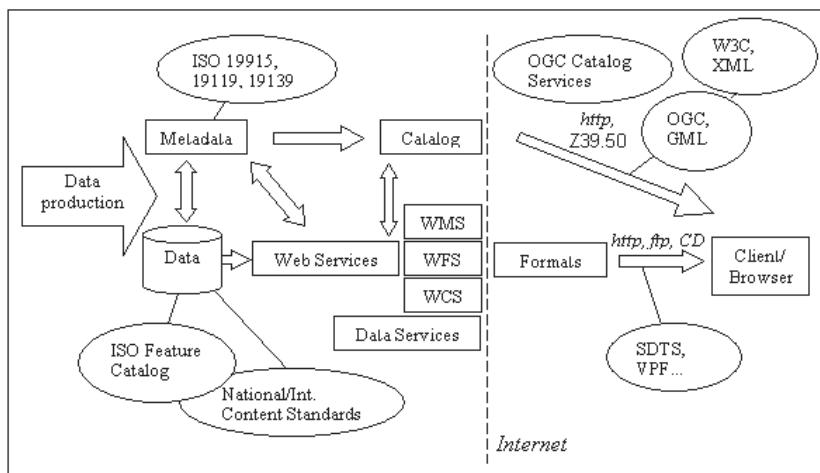


Fig. 2. Activities and standards in SDI (after Nebert 2004)

Especially when different systems and organizations have to communicate with each other there is an urgent need of standards. Without the use of standards, a systematic exchange of information can hardly take place. In addition, standards require a definite data structure that also arranges the available services. The organization offering the most important standards for building Centrope Map is Open Geospatial Consortium® (OGC 2006) which is a non-profit, international, voluntary consensus-based standards organization that is leading the development of standards for geospatial and location based services.

3.4 Various Open Tasks

Although succeeding in achieving very good results in the representation of cross-border-datasets coming from different sources, there are still some open problems that must be resolved step by step, e.g.

- Overlaps between datasets (problems with transparency),
- the cartographic styles of the regions still vary,
- some datasets are displayed, while others are not, although the settings appear identical.

No common basic map representation is guaranteed to present all the geodata of the whole region at the same time.

4 Centrope_Map: Realization and Result

4.1 Four and more WMS combined in 1 SDI

The particular challenge during the development of an SDI for Centrope region was the fusion of geographic information. As a first step, existing OGC-conform Web-Map-Services (WMS) were bundled and made accessible via a Web portal.

The current map visualizes the geodata of the Centrope region using the projection Austria Lambert 47°30'. This projection gives the best representation of this specific region.

The integration was not only a technical problem, rather more an organizational problem to solve, so in parallel to collecting and combining what is already available in the different regions there were workshops, meetings and negotiations on how to standardize and harmonize the different services.

In technical terms, despite the wide availability of spatial data via Web Map Services (WMS) there are still problems with interoperability, the most common of which are

- not every WMS is compliant with the Open Geospatial Consortium (OGC) WMS Implementation Specification® (OGC 2006).
- in some cases, an appropriate spatial reference system is not supported.

As the WMS for “Centrope MAP” are originate in institutions of different administrative levels, there are still a number of inconsistencies in content and accuracy (i.e. regional data coming from South Bohemia, Vy-

socina, South Moravia, as well as data covering the entire Czech Republic (offered by the Czech Environmental Information Agency, CENIA 2006) whereas geodata coming from Slovakia (offered by Slovak Environment Agency, SZAP 2006) offers data covering the entire country, the Czech geodata for South Bohemia and South Moravia are only available at the regional level). Another problem is the different map representation of Web Map Services belonging to the different regions as shown in Fig. 3.

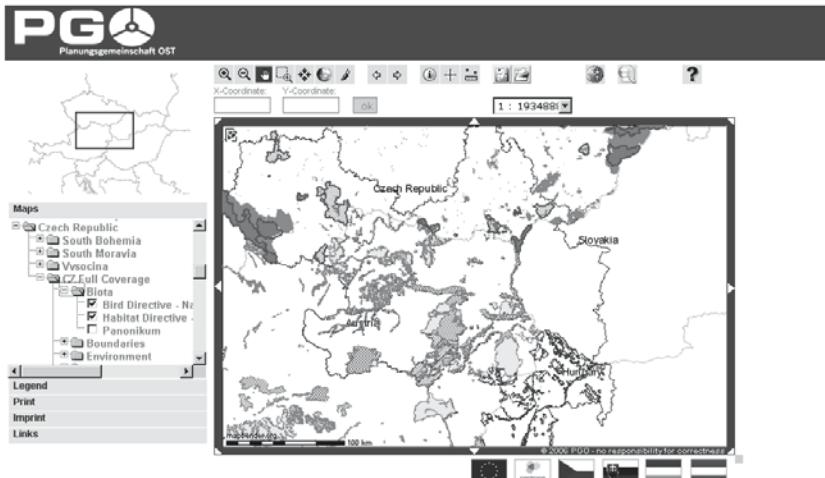


Fig. 3. Different styles for the same content

Moreover, the heterogeneity of languages delivers the layer titles coming via Web Map Service in the language of the requested country. Therefore the implementation is made with the help of a Geodata Explorer. A tree structure is built up to integrate the diverse layer considering a certain order and standard. As shown in Fig. 3, on the left side, first the layers in the Geodata Explorer are categorized by country and region. Second, each layer is ordered by the categories of the ISO 19115 code list “Core metadata for geographic datasets” (e.g. Biota, Boundaries, Environment). Moreover, all layers are translated from the official language of each country into the assigned working language - English.

4.2 Four Languages and 1 Multilingual Dictionary

Because of the multilingual nature of the Centrope Region, English was assigned as the working language for “Centrope MAP”. An online dictionary was added to dissolve the heterogeneity of terms. The Centrope

online dictionary (see Fig 4) can be accessed via a link in the menu window of the Centrope web map application: <http://map.centropemap.org/>.

German	English	Czech	Slovak	Hungarian
Bevölkerung	population	obyvatelstvo	obyvatelstvo	lakosság
Gemeinde	community, municipality	obce	obce	település
Haushalt	household	májnost, dům, rodina, služebnictvo	domácnosť	háztartás
männlich/weiblich	male/female	samec/ samice	samec/samica	férfi / nő
Altersklasse	age group	věková skupina	vek skupina	korosztály
Geschlecht	sex, gender	pohlaví	pohlavie	nem
Geburtenbilanz	balance of births	balancovat of narození	rovnováha zo narození	születések aránya
Arbeitnehmer	employee	zaměstnanec	zaměstnanec	alkalmazott
Wasserlauf	watercourse	Vodní tok		vizolyam
Kanal	channel	kanál	Kanalizácia	Csatloma
Bach	stream	tok	Prúd	ér
Quelle, Brunnen	spring well	skoda dobré	zdroj prameň	Forrás, kút
Feld	field	prostor,pole	Pole	Mező
Straße	way, road	cesta	Cesta	Ut
Komitat, Bezirk	county	hrabství	Zupa	Megye
Verwaltungsgrenze	administrative boundary	Správni hranice	Správne hranice	Kozigazgatási határ
Natura2000 Habitat-Gebiete	Natura2000 habitat areas	Natura2000 Lokality plochy	Natura 2000 lokality	Natura2000 pSC terület
Natura2000 Vogelschutz-Gebiete	Natura2000 bird areas	Natura2000 letoadlo	Natura 2000 vtácie územia	Natura2000 SPA terület
Nationale Naturschutz-Gebiete	National protected areas	Národní chránit plochy	Štátne chránené územia	természetvédelmi terület

Fig. 4. “Centrope MAP” Online Dictionary

Each partner can make new entries into the Online Dictionary. The color of the entry (see) signifies the status of rectification: green means that the vocabulary has been checked and seems to be correct, red signs an incorrect entry and blue that the word has not been verified yet.

4.3 Technical Details on Software and Services

For the technical implementation of “Centrope MAP”, mainly Free and Open Source software for Geoinformatics was used. The core of the Centrope SDI consists of UMN MapServer and Mapbender, a WebGIS client suite.

UMN MapServer is an Open Source development environment that allows the creation of image maps out of geodata (e.g. ESRI shapefiles, ESRI ArcSDE, PostGIS, Oracle Spatial, MySQL, TIFF/GeoTiff, EPPL7) and visualization of the results via the Web (MapServer 2006). The UMN

MapServer was initially developed by the University of Minnesota (UMN).

Mapbender is an Open Source Geospatial Foundation (OSGeo) project founded by the Consulting Center for Geographic Information Systems (CCGIS 2006), a company in Germany. It is implemented in PHP, JavaScript and XML and acts as a geoportal that offers the management of OGC compliant WMS. The Mapbender framework enables an authentication and authorization process for WMS, the management of graphical user interfaces for different user groups, and the administration of services.

Mapbender facilitates the agglomeration of OGC compliant WMS from distributed servers.

At the current state in Centrope Map data and services from the following institutions are integrated:

- Federal state of Burgenland (Austria)
- Geoland.at (Austria)
- CENIA (Czech Republic)
- SAZP (Slovakia)
- South Moravia (Kraj Jihomoravsky)
- Municipality of Vienna (Austria)
- Federal State of Lower Austria (Austria)
- South Bohemia (Czech Republic)
- Hungary

There are ongoing negotiations with more organizations and institutions, whose WMS can be integrated soon.

4.4 Topographic Basemap for Centrope-Region

As the national mapping agencies in the involved countries use different representations for their topographic maps, the project consortium agreed that the product “GeoAtlas” would be used as a common topographic basemap for the whole region (see Fig. 5). These datasets are extracted from datasets by Teleatlas® data. So Centrope Map not only is an example of the combination of data from different administrative bodies, but also from the private sector – an approach that also shall be extended.

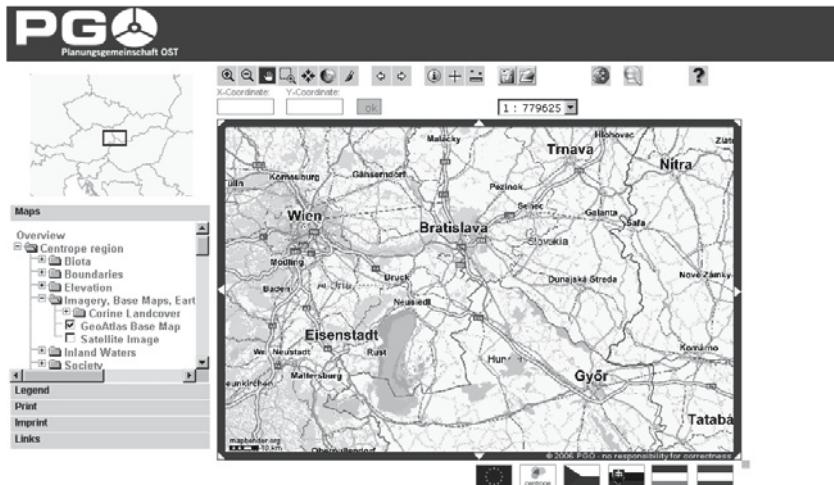


Fig. 5. GeoAtlas data, derived from Teleatlas-datasets, provides a homogenous topographic map

5 Conclusions and Work to Do

Albeit not an easy task, CENTROPE MAP shows that cross-border geospatial infrastructures can be successfully implemented and data and services from various sources can be integrated and used together.

In the case of CENTROPE MAP, it was essential not to see the project as an end in itself, but to have in mind the “CENTROPE-IDEA” and to deliver useful and visible results for the project partners as well as for the project sponsors in an appropriate timeframe – therefore sometimes it has to be accepted that not each single technical problem can be finally resolved in such a project but pragmatic, down-to-earth solutions and work-arounds have to be found.

There is an obvious demand for integrating spatial data from different sources, belonging to heterogeneous “information communities”. Usually this integration is non-trivial because of the various semantics used for spatial data in singular information communities. International standards such as those offered by the OGC and ISO support such efforts. In “Centrope Map” the ongoing activities to combine spatial data from new partners demonstrate that the overall benefit of the project is growing hand-in-hand with the number of partners contributing their knowledge and data.

On the one hand, there are still many non-standard-compliant and therefore non-compatible geodata sets (e.g. data format, granularity, attributes, accuracy) that avoid or limit the exchange of geodata – and there are even newly generated information sources that do not follow well- accepted standards.

On the other hand, the good news is that skilled and qualified experts can be found in every region, so from a technical point of view we can be optimistic about creating and improving ever more cross-border infrastructures.

An SDI can only be as strong as the cooperation of its partners. There is a pressing need for improved transnational coordination as well as the use of international standards to foster the exchange of spatial data and services.

5.2 Future Work

As previous project phases show, many tasks can only be solved jointly. For the ongoing project, the homogenization of the contents (mainly represented by spatial data) signifies an essential task. As an example, the integrated Natura 2000 data of Austria, Czech Republic, Hungary, and Slovakia have already nearly acquired a common style (see Fig. 6).

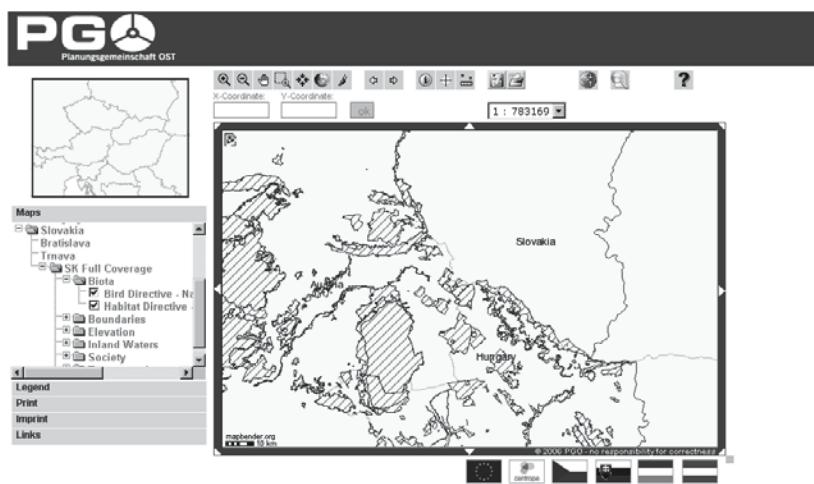


Fig. 6. Natura 2000 data

Therefore, one of the next technical steps for “Centrope Map” will be to improve the cartography. A common style guide and the use of Styled

Layer Descriptor (SLD) should combine the layout of the various Web Map Services displayed in one map. As a trained eye can see, in

Fig. 6 shown above, the data of Slovakia are missing. Obviously, the server is shut down or has some other problem. An inexperienced user will not understand why the requested data does not show up.

Furthermore, in the upcoming project phase, the function of online-digitizing new layers will be implemented using transactional Web Feature Services (WFS-T). Thereby different users will be able to add new features for a common use.

Finally, the content of the web mapping application will be widened by adding new datasets, especially realtime-information for monitoring purposes. Therefore, a close cooperation with neighboring partners in the whole region is necessary.

References

1. Craglia, M., Annoni, A., Smits, P. and Smith, R., 2003, SDI Developments in Western Europe. In GI in the wider Europe - GINIE: Geographic Information Network in Europe, edited by University of Sheffield (USFD), Open GIS Consortium Europe (OGCE), European Umbrella Organisation for Geographic Information (EUROGI) and Joint Research Centre of the European Commission (JRC) pp. 19-70.
2. Burrough, P. and Masser, I., (eds.), 1998, European Geographic Information Infrastructures opportunities and pitfalls. (London: Taylor & Francis Ltd).
3. European Umbrella Organisation for Geographic Information - EUROGI, 2005, Official Website of the European Umbrella Organization for Geographic Information - EUROGI. http://www.eurogi.org/index_1024.html accessed on June 21, 2005.
4. Hecht, L., 2002, How to build an INSPIRE Node. Geoinformatics - Magazine for Geo-IT Professionals, Volume 6.
5. Nebert, D., 2004, Developing Spatial Data Infrastructures: The SDI Cookbook. <http://www.gsdi.org/> accessed on June 21, 2005, Version 2.0.
6. Open Geospatial Consortium - OGC (2006). OpenGIS® Implementation Specification. OpenGIS® Web Map Server Implementation Specification, Open Geospatial Consortium Inc
7. Smits, P. C., Düren, U., Østensen, O., Murre, L., Gould, M., Sandgren, U., Marinelli, M., Murray, K., Pross, E., Wirthmann, A., Salgé, F. and Konecny, M., 2002, INSPIRE Architecture and Standards, European Commission, Joint Research Centre, Ispra, Position Paper No. EUR 20518 EN.
8. Central European Region - Centrope, <http://www.centrope.info> accessed on November 30, 2006
9. "Centrope Web MAP", <http://www.centropemap.org> accessed on November 30, 2006
10. "Centrope Web MAP Server", <http://map.centropemap.org> accessed on November 30, 2006
11. Consulting Center for Geographic Information Systems - CCGIS, <http://www.ccgis.de/> accessed on November 29, 2006
12. County of South Moravia –Jihomoravský kraj, <http://www.kr-jihomoravsky.cz/> accessed on November 30, 2006

13. Czech Environmental Information Agency - CENIA, <http://www.cenia.cz/> accessed on November 30, 2006
14. General Multilingual Environmental Thesaurus - GEMET thesaurus, <http://www.eionet.europa.eu/gemet> accessed on November 29, 2006
15. Geodata portal of the Austrian provinces - Geoland.at, <http://www.geoland.at> accessed on November 30, 2006
16. Mapbender, <http://www.mapbender.org> accessed on November 30, 2006
17. Open Geospatial Consortium - OGC, <http://www.opengeospatial.org/> accessed on November 29, 2006
18. Planungsgemeinschaft Ost - PGO, <http://www.planungsgemeinschaft-ost.org> accessed on November 30, 2006
19. Slovak Environment Agency - SAZP, <http://www.sazp.sk/> accessed on November 30, 2006
20. UMN MapServer, <http://mapserver.gis.umn.edu/> accessed on November 29, 2006

Software Environment for Simulation and Evaluation of a Security Operation Center

Julien Bourgeois¹, Abdoul Karim Ganame¹, Igor Kotenko² and Alexander Ulanov²

¹LIFC, Universite de Franche Comte, 4, place Tharradin, 25211 Montbeliard, France
{ganame, bourgeois}@lifc.univ-fcomte.fr

²St. Petersburg Institute for Informatics and Automation (SPIIRAS), 39, 14 Linia, St.Petersburg, 199178, Russia {ivkote, ulanov}@comsec.spb.ru

Abstract. It is somewhat problematic to evaluate the performance of security systems in the Internet due to complexity of these systems and the Internet itself. Therefore, modeling and simulation are becoming more and more important in optimizing the behavior of security systems, including security components intended for protecting various distributed geographic information systems (GIS). This paper presents an approach and software simulation environment for comprehensive investigation of the Security Operation Center (SOCBox) system which is in essence an intrusion detection “metasystem”. SOCBox collects data from a wide range of sources (intrusion detection systems (IDS), firewalls, routers, workstations, etc.) and therefore has a global view on the network. The simulation environment has been developed formerly for Distributed Denial of Service (DDoS) attacks and defense simulation. This tool is characterized by agent-oriented approach, the packet-based imitation of network security processes and the open library of different attacks and defense mechanisms. We consider the SOCBox structure, the simulation environment architecture, the SOCBox models in the simulation environment and peculiarities of SOCBox simulation.

Keywords: Security modeling and simulation, infrastructure security, intrusion detection, DDoS.

1 Introduction

The design of reliable defense system for complex and distributed computer systems including geographic information systems (GIS) is a complicated problem. Such a system has to include the mechanisms of prevention, detection, unauthorized access (or attack) source tracing and

malicious actions counteraction. It is obvious that the more useful data such system has the more effective it is. This especially concerns the detection mechanisms. The large-scaled system of distributed sensors gives the following opportunity: one can represent the more complete picture of current state and detect an attack in a long distance from the attack goal and then take some countermeasures.

Such approach is suggested in the Security Operation Center (SOCBox) [1-3], that could be called intrusion detection “metasystem”. SOCBox collects data from a wide range of sources (intrusion detection systems (IDS), firewalls, routers, workstations, etc.) and therefore has a global view on the network. The SOCBox analysis engine can correlate all messages generated by all network components and find patterns of intrusion. However, the implementation of SOCBox is a complicated problem.

An advanced hardware test bed is needed to debug, test and evaluate the effectiveness of SOCBox. It can be a laborious independent network or the Internet fragment. A.K.Ganame et al. [3] describe different experiments conducted with SOCBox. They presented the intrusion detection capabilities and performance of SOCBox in comparison with Snort IDS. Experiments were carried out in a real Internet service provider (ISP) network for over a week. This ISP manages more than 50000 subscribers. But, it is hard (or impossible) to implement a set of periodical global experiments in the networks of real ISP (e.g. when applying a new detection technique). Furthermore, the important conditions for a scientific experiment are repetition and controllability. These conditions are hard to fulfill on a hardware test bed, but they can be provided in a simulation environment.

The choice of simulation type depends on scalability and fidelity requirements. The variety of simulation tools spreads from hardware test beds to analytical models. Hardware test beds, e.g., EmuLab [4, 5], offer the real network incorporating hundreds or even thousands of hosts. One can simulate up to dozens of thousands of hosts using the network emulation, e.g., NetLab [4, 5], ModelNet [4, 6]. Both hardware test beds as well as emulation systems are by definition executed in real-time. Next alternative is a packet-level simulation: OMNeT++ INET Framework [7, 8, 9], NS-2, SSFNet, J-Sim INET Framework [7, 8]. Packet-level simulation exhibits one of the best tradeoffs between scalability and fidelity. Mixed simulation is a combination of packet-level simulation and analytical models [4, 10]. The latter is the most scalable but the most simplified simulation approach [4, 11].

Test beds and SOCBox itself can be simulated with the given scalability and fidelity (accounting for certain assumptions). Multi-agent simulation

environment of Distributed Denial of Service (DDoS) attack and defense mechanisms [7, 8, 12] is an example of such approach implementation. It has been developed for a comprehensive investigation of the Internet DDoS attacks and defense mechanisms. This tool can be characterized by three main features: agent-oriented approach to simulation, packet-based imitation of network security processes, and open library of different DDoS attacks and defense mechanisms.

This paper presents the methodological and practical foundations for the development of the SOCBox simulation models and their implementation based on the developed multi-agent environment.

The rest of the paper is structured as follows. Section 2 considers the SOCBox structure and main mechanisms realized. We discuss the functional architecture needed to simulate the SOCBox modules. Section 3 suggests a common approach and simulation environment developed for the investigation and elaboration of the adequate defense against attacks methods. This environment can produce well-grounded recommendations on the choice of defense mechanisms that are the most efficient under particular conditions. Section 4 and 5 describe the basis for SOCBox simulation and the architecture of simulation tool needed. Section 6 represents the implementation aspects of the environment for the SOCBox simulation. The conclusion presents the main results and outlines future research.

2 Security Operation Center

The Security Operation Center (SOCBox) is built up of *six distinct classes of modules* (Fig.1): event generators (E-Box), event collectors (C-Box), message database (D-Box), analysis engine (A-Box), knowledge base (K-Box) and reaction management module (R-Box). The main problem encountered when building SOCBox is the integration of all these modules, usually built as autonomous parts, while matching availability, integrity and security of data and their transmission channels.

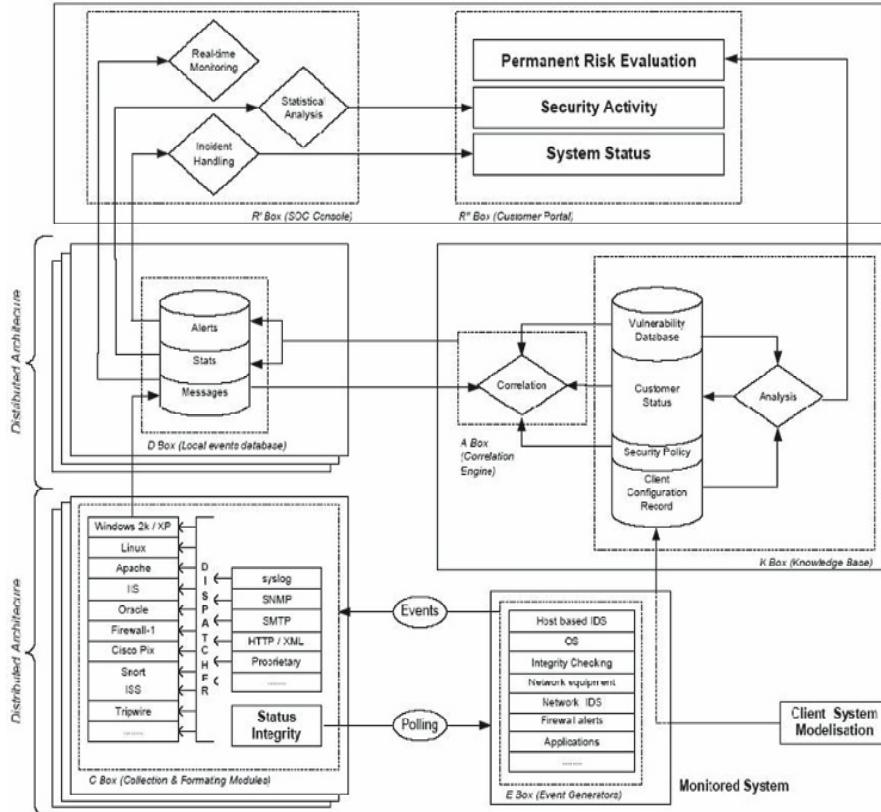


Fig.1. The SOCBox architecture

E-Boxes are responsible for event generation. We can distinguish two main families of such Boxes: *event based data generators* generating events according to a specific operation performed on the operating system (OS), applications or over the network, and *status based data generators*, generating an event according to the reaction to an external stimulus such as ping, data integrity checking or daemon status check.

C-Box purpose is to collect information from different sensors and to translate it into a standard format, in order to have a homogeneous base of messages. Availability and scalability of these boxes appears to be a major concern.

D-Box is a database storing events captured by C-Box as well as the alerts generated by the Analysis engine (A-Box). Beside the classical concerns regarding database availability, integrity and confidentiality, D-Box often faces the performance problems because sensors may generate dozens of messages every second. Those messages will have to be stored,

processed and analyzed as quick as possible, in order to allow a timely reaction to intrusions.

A-Box is responsible for the analysis of events stored in D-Box. It performs various operations in order to provide for qualified alert messages. An approach dealing with the structural analysis of intrusion attempts is used, as well as behavior analysis. It is evident that the analysis process needs inputs from a database in which intrusion path characteristics, the protected system model and the security policy are stored. This is the very purpose of the *K-Box*.

R-Box is a response and reporting tool used to react against the offending events taking place on or within the supervised systems.

2.1 Data Acquisition

Before setting up sensors and designing any correlation or analysis rule, it is necessary to evaluate the overall security level of the information technology infrastructure to be supervised.

This will make it possible to determine if an intrusion path may actually lead to an intrusion on the target system and also a criticality associated to such an intrusion attempt. Another point to be defined is the security policy, mostly in the terms of access rights, permitted operations, etc.

2.1.1 Vulnerability Database

The vulnerability database holds information about security breaches and insecure behavior that would either impact the overall security level or could be exploited by an attacker in order to perform an intrusion.

The database format must allow including the following *types of vulnerabilities*:

- *Structural vulnerabilities*, i.e. vulnerabilities internal to specific software such as buffer overflow, format string, race conditions, etc.
- *Functional vulnerabilities*, depending on configuration, operational behavior, users, etc. These vulnerabilities differ from the previous ones as they strongly depend on the environment where they exist.
- *Topology-based vulnerabilities*, including networking impact on intrusions and their consequences.

2.1.2 Security Policy

The next step of the supervised system inventory is an organizational one and, more specifically, a review of security policy aspects that would affect either event generation and/or the reaction-reporting processes.

It is clear that two major aspects of security policy required to be reviewed are authorization and testing/audit procedures. These two aspects will provide information concerning behavior that sensors will detect.

Events generated (administrator login, port scans, etc.) will then be marked as matching security policy criteria. Other will be analyzed as a possible part of an intrusion attempt. This information is stored in the Knowledge Base (K-Box).

2.2 Data Collection

Collecting data from heterogeneous sources implies the setup of two kinds of agents: protocol and application. The former collects information from E-Boxes, the latter parses information for storage in a “pseudo-standard” format.

These two modules are connected by a dispatcher. Such an architecture allows the high-availability and load-balancing systems to be set at any level of the SOCBox architecture.

2.2.1 Protocol agents

Protocol agents are designed to receive information from specific transport protocols, such as syslog, SNMP, SMTP, HTML, etc.

They act like server side applications and their only purpose is to listen to incoming connections from E-Boxes and make collected data available to the dispatcher. The simplicity of such agents makes them easy to implement and maintain.

The raw format storage is usually a simple file, though direct transfer to the dispatcher through named pipes, sockets or shared memory ensures a better performance. From a security point of view, the most important point is to ensure the integrity of data collected by agents. Therefore, data are encapsulated into a secure tunnel.

2.2.2 Dispatcher and application agents

The dispatcher's purpose is to determine the source-type of an incoming event and then forward the original message to the appropriate application agent.

Again, implementation is relatively trivial, once a specific pattern has been found for each source-type where the data may be received from.

Autonomous operations performed by the dispatcher are as follows:

- Listening to an incoming channel from protocol agents, such as socket, named pipe, system V message queue, etc.
- Checking pattern matching against a patterns database that should be pre-loaded in memory for performances considerations.
- Sending the original message to an E-Box specific application agent through any suitable outgoing channel.

Application agents perform formatting of messages so that they match with the generic model of the message database. Autonomous operations performed by application agents are as follows:

- Listening to an incoming channel from dispatchers, such as a socket, named pipe, system V message queue, etc.
- Parsing the original message into standard fields.
- Transmitting the formatted message to the D-Box.

2.3 Data Analysis

The main operations performed to generate alerts are: correlation, structural analysis, intrusion path analysis and behavior analysis. Correlation is a stand-alone operation leading to the creation of contexts in which further analysis will be made, in order to check if they match the characteristics of an intrusion attempt. Structural analysis may be compared with an advanced pattern matching process, used to determine whether events stored within a certain context lead to a known intrusion path or to an attack tree [13]. Intrusion path analysis is the next step whose output provides the intrusion attempt detected with information about the exposure of the target system.

Then, the behavior analysis integrates elements from the security policy in order to determine if the intrusion attempt allowed or not. The purpose of such operations is to generate alerts that do not only match the structural path of intrusion (i.e. scan, fingerprinting, exploiting, backdooring and cleaning), but also take care of the security policy defined, as well as the criticality of target systems.

2.4 Correlation Overview

The correlation engine purpose is to analyze complex information sequences and produce simple, synthesized and accurate events. In order to generate such qualified events, *five operations* are to be performed:

- The first, obvious, operation is to identify duplicates and set a specific flag in order to keep the information and continue without the multiple identical messages.
- Sequence patterns matching is the most common operation performed by the correlation engine. Its purpose is to identify a sequence of messages which would be a characteristic of an intrusion attempt. This allows identifying on-going intrusion processes, as well as complex intrusion scenarios.
- Time pattern matching is designed to include another important dimension in intrusion analysis: time. This is mainly used for context management, as well as slow and distributed intrusion processes.
- System exposure and criticality analysis provide information about the target system's vulnerability to detected intrusion attempts. Indeed, it seems inappropriate to have SOCBox generating alarms concerning an intrusion scenario based on a vulnerability that the target system is not exposed to. Another piece of information is the criticality of the intrusion, i.e. its overall impact on the supervised system. This helps to manage the priorities in terms of reaction to multiple incidents.
- Security policy matching is a behavior-based filter that eliminates specific events if they match security policy criteria such as administrator login, identification processes and authorizations / restrictions.

3 Simulation Environment

We have developed the common approach and software environment for simulation and investigation of DDoS attacks and defense systems [7, 8]. The main attention in this work was drawn to the integrated agent-oriented and packet-level approach to the simulation of security processes in the Internet which can provide the acceptable fidelity and scalability for the implementation of computer attacks and defenses.

According to this approach, the cybernetic counteraction is represented as the interaction of various software agent teams. At least two agent teams are selected that affect computer network and each other: agent-malefactor team and defense agent team.

Attack agents are divided at least into two classes: “daemons” and “master”. “Daemons” are attack executors; “master” coordinates them. At the preliminary stage “daemons” and “master” are deployed on available (already compromised) hosts. The attack class is defined by the set of different parameters.

According to the general DDoS defense approach, the defense agents are divided into the following classes: information processing (“sampler”); attack detection (“detector”); filtering and balancing (“filter”); investigation (“investigator”).

Samplers collect and process network data for anomaly and misuse detection.

Detector coordinates the team and correlates data from samplers.

Filters are responsible for traffic filtering using the rules provided by detector.

Investigator tries to defeat attack agents.

Defense team implements certain investigated defense mechanism.

The proposed simulation approach assume the following components of the simulation environment developed (Fig.2):

- Simulation Framework;
- Internet Simulation Framework;
- DDoS Framework (Library of attacks and defenses);
- Multi-Agent Simulation Framework.

The discrete event simulator OMNeT++ [9] and the set of modules OMNeT++ INET Framework (that allows simulating the Internet hosts and protocols) were used as the kernel for the simulation environment.

On the basis of these tools, DDoS Framework and Multi-Agent Simulation Framework were developed. DDoS Framework includes DDoS attack and defense modules and the modules that expand the hosts of INET Framework: filter table and packet analyzer.

The following parameters are used in the environment to define the attack: victim type; type of attack; attack rate dynamics; impact on a victim; persistence of agent set; possibility of exposure; source address validity; degree of automation.

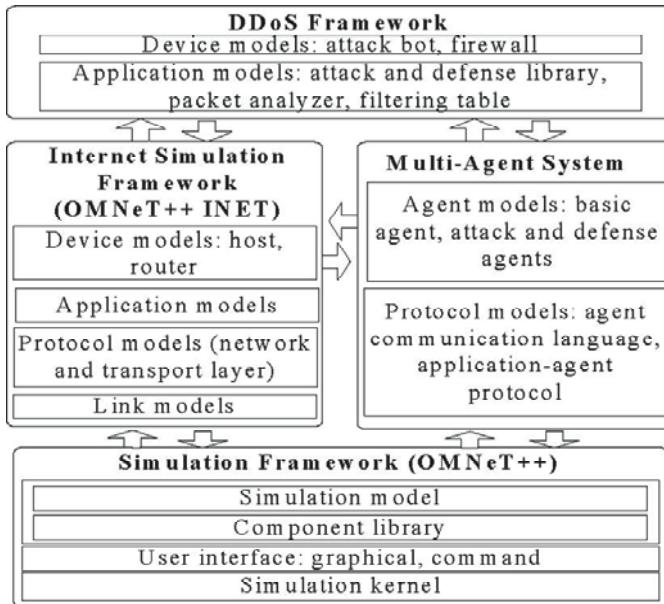


Fig.2. Environment architecture

Defense mechanisms are determined in the environment by the following parameters: deployment location; mechanism of component cooperation; covered defense stages; attack detection technique; attack source detection technique; attack prevention and counteraction technique; model data collecting technique; technique determining the deviation from model data.

Multi-agent Framework consists of modules representing intelligent agents implemented as applications.

The following elements of abstract FIPA architecture [14] were used:

- agent communication language,
- agent message transport protocol,
- agent directory.

The idea of such a representation is to provide agent interactions and reusability.

Fig.3 shows the example of the simulation environment multi-window interface. One can see the simulation management window (Fig.3, at top left), state visualization windows (Fig.3, at right), and the fragment of simulated network window (Fig.3, at bottom left).

The main point of defense mechanisms evaluation is to carry out the series of experiments in the developed simulation environment. The experi-

ments are fulfilled for various values of input parameters to measure the defense mechanisms effectiveness and efficacy and to analyze them.

The made experiments showed the applicability of the proposed approach to simulating prospective defense mechanisms and to analyzing existing and designed networks.

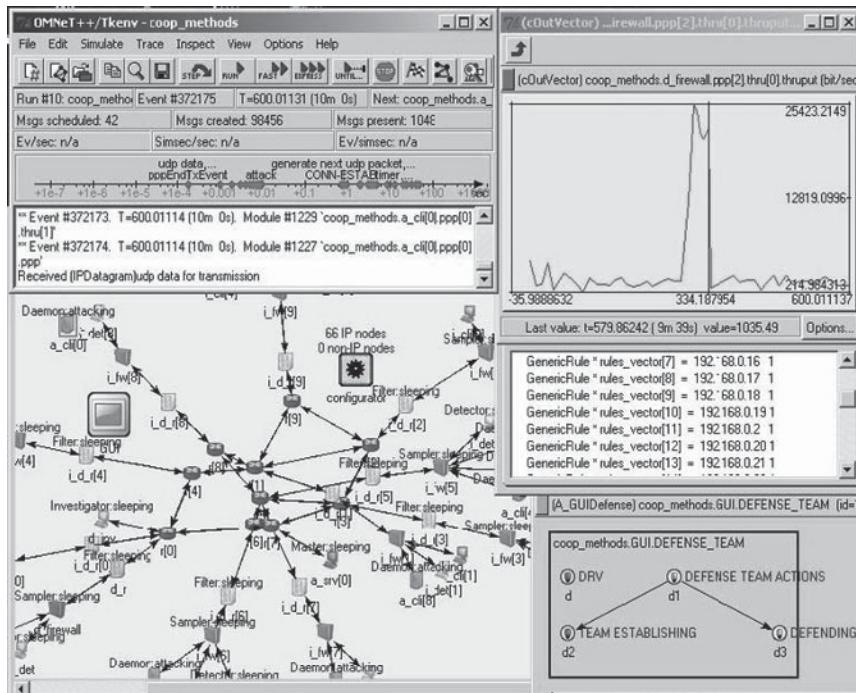


Fig. 3. User interface of simulation environment

4 SOCBox Simulation Approach

Here it is proposed to use *the agent-oriented approach based on the packet-based imitation of network security processes* [7, 8] for the SOCBox simulation. It is briefly described in Section 3. There are two agent teams: attack team and defense team that implement SOCBox and its environment.

Attack team has the same structure as in the considered approach [7, 8]: agents “daemons” and “master”. “Daemons” are able to execute various attack classes. This gives a possibility to comprehensively evaluate the SOCBox capability of attacks detecting.

The attack classes are represented in the Fig.4.

There are the following *defense agent classes* according to the SOCBox architecture: E-Box, C-Box, D-Box, A-Box, K-Box, and R-Box.

E-Box agents are the event sources. The event is based on the message from the external source (sensor) or the generated status message (poller).

C-Box agents are the collectors of events coming from E-Box. They confront the event with the certain type and translate it into the common format to transmit it to D-Box agent.

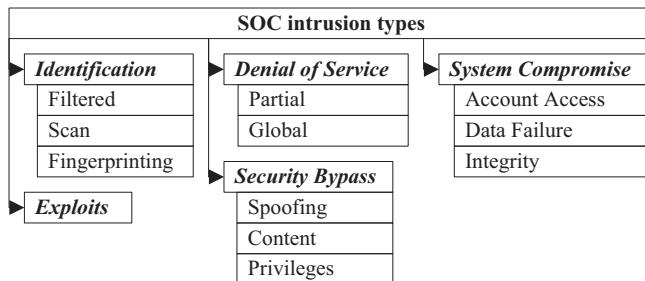


Fig.4. Classes of attack parameters

D-Box agents are the databases of events received from C-Box. They do the basic functions on deleting events-duplicates and on event filtering to reduce the load.

A-Box agents analyze the events stored in D-Box and make the decision about the attack. They send the messages about these events to the R-Box agents.

K-Box agents contain the vulnerabilities database, system status and security policy. A-Box uses the data during event analysis.

R-Box agents receive the messages about the attack from A-Box, propose the ways of the problem solution and form the reports.

It is needed to simulate the external event sources to cause the E-Box generate the events. These sources can be the following: firewalls, routers, IDS, sniffers, web-servers, etc.

The following components are used as the external sources in the proposed simulation (they are based on the implemented basic components of simulation environment, see Section 3):

- Defense team as the IDS: E-Box can receive the messages about attack detection. The defense team has to be in one of cooperative modes [12].
- Sampler from the defense team: it generates the statistics on network traffic.

- Filter from the defense team: it generates the information about dropped or passed packets.
- Sniffers: they sniffer the traffic from the whole subnet or from the supervised host.
- HTTP-server model: it receives the requests from the clients and produces the certain amount of replies of certain sizes via HTTP.

Generation of status messages is done due to calculation of HTTP-server response time to the periodical request of E-Box. Response time is sent to the C-Box agent in the form of event. E-Box is deployed in the various networks or hosts depending on the external source type.

The use of real intrusion detection systems (e.g., Snort, Bro) is the possible investigation task aimed at simulation reinforcement. It is needed to plug these systems to the network traffic in the simulation environment. One can also simulate these systems.

There exists a different number of C-Box agents depending on the number of events coming from E-Box agents. The former can receive events from various types of E-Box working on different external sources.

D-Box agent has to be deployed so that all C-Box agents will be able to access it since it stores all events from C-Box.

A-Box agent analyzes the events from D-Box database due to data stored in K-Box. The latter stores the vulnerabilities database, model of defended system (its status) and security policy. Various attack detection mechanisms are proposed to be used in this model. They and their parameters are defined before simulation starts. Certain mechanism can be changed to the other one during simulation. Mechanism implementation is stored in A-Box. System status may have various sets of attributes depending on the type of mechanism. In particular, there can be the model data about normal (or abnormal) traffic for the defended network obtained at the learning stage. A-Box and K-Box agents have the functionality which is similar to the “detector” agents (see Section 3).

R-Box agent consists of monitoring and management components. Thus, a user can track the current system security state and compose the reports. A-Box and K-Box mode is set due to management component. The mode can be, for example, “learning”, “normal”, “simulation repetition”, etc. Experiment results recorded by R-Box can be used to repeat the situation and its further analysis.

5 SOCBox Simulation: Architecture and Implementation

The SOCBox simulation environment is developed based on multi-agent simulation environment of DDoS attack and defense mechanisms.

Attack and defense library is extended to implement new environment. The modules of the following attacks realization: Identification, Security Bypass, System Compromise - are included. Multi-agent simulation framework is expanded with the modules of “E-Box”, “C-Box”, “D-Box”, “A-Box”, “K-Box” and “R-Box” agents. “E-Box” agents are built upon the “sampler” agent class while “A-Box” and “K-Box” – upon “detector” agent class since they have similar functionality. It is needed to implement the agent communication language on the basis of “Message” and “Context” structures used in SOCBox.

The simulation environment proposed for SOCBox allows running various experiments to investigate the strategies of attacks and prospective attack detection mechanisms. One can vary during experiments: network topology and configuration, structure and configuration of attack and the SOCBox teams, the SOCBox team deployment location, attack mechanisms etc. Various SOCBox effectiveness parameters are measured through the experiment results. The analysis of conditions and possibilities of their application is performed. There exist different effectiveness parameters, e.g., false negative rate, false positive rate, performance, attack reaction time.

Networks used for simulation consist of various subnets that are (for example) the responsibility areas of different ISPs. At that the following subnets are distinguished: the defense subnet where attack target resource is, the intermediate subnet where the hosts generating typical traffic are, and the attack subnet where the attack agents are. Simulated networks are built using algorithms that allow creating subnets close to real Internet configurations. Networks consist of routers, client and agent hosts.

The experiment aimed at evaluation of SOCBox capacity and effectiveness is currently implemented. It is similar to the experiment described in [3] where tests were carried out in the real ISP network having over 50000 subscribers. First, the SOCBox model is being tested for a capability to manage the heterogeneous events coming from different sources. After that, some exploits are executed against the network to check for the SOCBox capacity to detect various classes of intrusions. Then, the SOCBox capability to detect distributed intrusions is evaluated. After that, the clarity and the relevance of the SOCBox reports are studied. Finally, performance evaluation is made.

6 Conclusion

The paper proposes the approach to investigate the capabilities of an intrusion detection “metasystem” called SOCBox which can be used for protecting various distributed systems, including networked GIS. The approach is based on the use of developed simulation environment [7, 8, 12].

The main feature of the SOCBox is to collect and process data from different sources for the defended system global monitoring. It assumes the presence of particularized agents that collect data about system functioning. The analysis engine of the SOCBox processes the data and finds intrusion patterns in them. Afterwards, it correlates the generated alerts and makes the decision about an attack. The SOCBox includes also a module responsible for attack reaction and reports generation.

Here presented modeling environment allows simulating the attacks aimed at denial of service of information resources and appropriate defense mechanisms. It can be used to analyze the effectiveness of defense mechanisms implemented due to various experiments. They are in the investigation of various mechanisms effectiveness parameters dependence (false negative, false positive rates etc.) on various input parameters (network topology, attack type etc.).

New classes of defense agents are created. They are needed for the SOCBox simulation based upon the developed simulation environment and form the defense team with already existing agents. New agent communication protocol is implemented in this team. The attack and defense library were also extended to let the attack team realize big variety of attack classes. Simulation allows investigating the effectiveness of SOCBox due to testing it in various conditions and estimating the optimal parameters of different modules. These results can be used at implementation and addition of new facilities to SOCBox.

Future work is related to the further development of the theoretical approach suggested and the simulation environment developed. The families of various models (analytical, mixed abstract, packet-level simulation, emulation, hardware) are intended to be used for research simulation. The choice of models is determined first of all by the required fidelity and scalability of simulation. In the future research it is planned to expand the attacks and defenses library, elaborate particular components functionalities. The important constituents of future research are numerous experiments to investigate various attacks, defense mechanisms and optimal defense combinations.

Acknowledgments

The research is supported by grants of Russian Foundation of Basic Research, Department for Informational Technologies and Computation Systems of the Russian Academy of Sciences (contract No 3.2/03), Russian Science Support Foundation and by EC as part of the POSITIF project (contract IST-2002-002314), RE-TRUST (contract No. 021186-2), CAPM, CRFC, French Government and European Union programme under the STIC pole.

References

1. Bidou, R., Bourgeois, J., Spies, F.: Towards a global security architecture for intrusion detection and reaction management. Proceedings of the 4th International Workshop on Information Security Applications (WISA 2003). Lecture Notes in Computer Science, Vol.2908 (2003)
2. Ganame, A.K., Bourgeois, J., Bidou, R., Spies, F.: A High Performance System for Intrusion Detection and Reaction Management. Journal of Information Assurance and Security, Vol.3 (2006)
3. Ganame, A.K., Bourgeois, J., Bidou, R., Spies, F.: Evaluation of the Intrusion Detection Capabilities and Performance of a Security Operation Center. International Conference on Security and Cryptography (SECRYPT 2006) Proceedings. Portugal (2006)
4. Perumalla, K., Sundaragopalan, S.: High-Fidelity Modeling of Computer Network Worms. Annual Computer Security Applications Conference (ACSAC), Tucson, AZ (2004)
5. White, B., Lepreau, J., Stoller, L., Ricci, R., Guruprasad, S., Newbold, M., Hibler, M., Barb, C., Joglekar, A.: An Integrated Experimental Environment for Distributed Systems and Networks. Proceedings of the 5th Symposium on Operating Systems Design and Implementation, Boston, MA (2002)
6. Durst, R., Champion, T., Witten, B., Miller, E., Spagnuolo, L.: Testing and Evaluating Computer Intrusion Detection Systems. Communications of the ACM, Vol.42, No.7 (1999)
7. Kotenko, I., Ulanov, A.: Simulation of Internet DDoS Attacks and Defense. 9th Information Security Conference (ISC 2006). Samos, Greece. August 30 - September 2, 2006. Lecture Notes in Computer Science, Vol.4176 (2006).
8. Kotenko, I., Ulanov, A.: Agent-based Simulation of Distributed Defense against Computer Network Attacks. Proceedings of 20th European Conference on Modeling and Simulation (ECMS 2006). Bonn. Germany (2006)
9. Omnet++. <http://www.omnetpp.org>
10. Kiddie, C., Simmonds, R., Williamson, C., Unger, B.: Hybrid Packet/Fluid Flow Network Simulation. IEEE/ACM Workshop on Parallel and Distributed Simulation (PADS) (2003)

11. Nicol, D., Liljenstam, M., Liu, J.: Multiscale Modeling and Simulation of Worm Effects on the Internet Routing Infrastructure. International Conference on Modeling Techniques and Tools for Computer Performance Evaluation (Performance TOOLS) (2003)
12. Kotenko, I., Ulanov, A.: Simulation Environment for Investigation of Cooperative Distributed Attacks and Defense. 9th International Symposium on Recent Advances in Intrusion Detection (RAID 2006). Abstract and Poster sessions. Hamburg, Germany September 20-22 (2006)
13. Schneier, B.: Attacks trees. Dr. Dobb (1999)
14. The Foundation for Intelligent Physical Agents (FIPA). <http://www.fipa.org>.

Security Policy Verification Tool for Geographical Information Systems

Igor Kotenko¹, Artem Tishkov², Olga Chervatuk¹ and Ekaterina Sidelnikova¹

St. Petersburg Institute for Informatics and Automation (SPIIRAS)

39, 14 Linia, St. Petersburg, Russia

¹{ivkote, ovch, sidelnikova}@comsec.spb.ru

²avt@iias.spb.su

Abstract. It is universally recognized, that one of the most effective approaches to security management consists in the use of policy-based security systems. This approach assumes that all actions of the system under defense are performed according to a policy incorporating a multitude of if-else rules describing the system behavior. It is hard for a system administrator while constructing a security policy to detect and resolve without an appropriate software tool all possible inconsistencies even inside one category of security rules (authentication, authorization, filtering, channel protection, operational, etc.), not to mention inter-category inconsistencies. The paper describes a general approach to the security policy verification and presents a software tool “Security Checker” that can serve as a security policy debugger for various policy categories. Security Checker can also be used as Security policy verification tool for complex distributed Geographical Information Systems (GIS).

Keywords: Security, security policy, verification.

1 Introduction

Security systems designed to protect various information systems (including geographical information systems (GIS)), based on policies, form one of the main streams in the development of protected systems due to the flexibility of management and the convenience of administration. Security policy describes a set of system behavior rules, including rules to maintain confidentiality, integrity and availability of system resources.

In large-scaled networks where a security policy includes hundreds or may be thousands of rules, an administrator is unable to track all inconsistencies emerging during the policy creation without corresponding automatic verification tool. This tool should (1) provide a way to search for and resolve contradictions between different policy rules, (2) allow to

check the possibility of policy application to system and also (3) give the estimation of network security level with supplied security policy.

To build a powerful and flexible security policy verification tool, it is very important to use an approach which allows covering all possible inconsistencies and has open (extendable) architecture and efficient conflict verification implementation. Such approach can be based on using a family of different verification modules each on working better and with acceptable computational complexity for the particular types of conflicts.

Security policy conflicts and anomalies have been carefully studied by many authors. The approach to conflict management and verification suggested by Morris Sloman's et al. [5, 26], Flexible Authorization Framework destined for authorization conflicts resolution developed by Sushil Jajodia et al. [22], filtering and IPsec anomalies detection techniques analyzed by Ehab Al-Shaer et al. [1, 15], etc. should be outlined.

Different common approaches are available for analysis and checking of formal specifications (see, for example, [5, 30, 39]), however, their applicability to the policy is limited. In fact, not all the security and dependability requirements are well modeled for complex systems [4, 16, 27, 33, 34, 35]. Additionally, the policy specifications as a rule are greater than usual calculus expressions.

General purpose methods for complex systems verification can be grouped into three categories [9, 19, 20, 21]: theorem provers, model checkers, and hybrid approaches.

Model checkers try to verify that a property is satisfied by exploring all the possible execution paths. Although model checkers can be configured and used relatively easy, they can not be reasonably applied to complex problems (like the consistency checking of a large-scaled system policy) because a state space dimension diverges rapidly. Examples of the most important model checkers are SPIN [18, 32], SMV and NuSMV [8, 28], VIS [14], MOCHA [2, 3] and PV [31]. All these tools are provided with an internal specification language. A standardization effort is done in this field for a verification language, the candidate is the Process Meta Language (PROMELA) [17].

Automatic reasoning systems and theorem provers, starting from a set of specifications (axioms), try to prove that a property (theorem) is satisfied. The major advantage is that these methods and corresponding tools can manage problems of large dimension at an expense of a very complex set up. Examples of theorem provers are given by HOL [13], ACL2 [24], Isabelle [7], etc.

Model checking and theorem proving are methods of the major paper's interest. However, it should be noted that along with general purpose

approaches, there exist the *specialized methods and techniques* aimed at verification of particular types of security policy rules.

The paper presents the concept and prototype of novel security policy verification tool “Security Checker” (SEC) that allows effectively detecting and resolving inconsistencies in computer network security policy. *The peculiarity of the approach is a hybrid multi-module architecture of SEC. Utilizing several verification modules, the Security Checker combines general verification methods with specialized algorithms which handle particular types of inconsistencies. The architecture is open for addition of new modules and therefore offers a flexible and scalable solution for security policy verification.* The input languages for SEC are XML-based System Description Language (SDL) and Security Policy Language (SPL). SDL allows specifying network topology and nodes’ functionality including security capabilities and services. SPL is a language for the specification of policy rules.

The rest of the paper is structured as follows. *Section 2* considers the classification of possible inconsistencies in security policy. *Section 3* describes Security Checker architecture. *Section 4 and 5* outline two general purpose modules, one based on combining Event Calculus with abductive reasoning and another – on model checking. *Section 6* gives a brief overview of specialized modules. *Section 7* represents the implementation aspects of Security Checker and its graphical user interface. The *conclusion* contains the results of work and further research.

2 Classification of Inconsistencies

In a policy verification problem, it is important to determine what inconsistencies of a security policy will lead security system to a deadlock, and what ones do not affect the stability of its work though deteriorate its effectiveness.

Therefore, all security policy inconsistencies are classified by the paper into two types:

- *Conflict* is a contradiction either between policy rules or between policy and system description which leads to a possibility of nondeterministic behavior of the security system. There are two situations that cause the uncertainty: (1) the security system must activate two or more rules at the same moment, but the actions of these rules are contradictory [37]; (2) the security system must activate a rule, but the rule action cannot be performed;

- *Anomaly* is a contradiction either between policy rules or between the policy and system description when one or more policy rules will never be considered for activation by the security system.

As an example of conflict one can imagine a situation when a user is assigned to two roles, one allowing some action and another denying it. Anomaly, for instance, appears in an ordered rule set, when a rule with higher priority “shadows” a rule with lower priority, thus, the action of the second rule will never be performed.

Policy conflicts lead security system to an ambiguous state of choice between contradictory actions, while in case of policy anomaly the system will make unambiguous solution. However, from the administrator’s (user’s) point of view, anomalies are as important as conflicts. Anomalies force the system to ignore rules, possibly including very important ones for correct system behavior. Thus, analyzing policy, it is necessary to define its critical parts and carefully check for all anomalies.

The classification of inconsistencies can be constructed using several criteria:

- (1) Detection time – either during policy formulation or runtime;
- (2) The degree of dependence on network topology and network nodes functionality – either not dependent (an inconsistency appears only for policy rules) or dependent (the inconsistency of policy application to a network configuration);
- (3) Policy rule category, etc.

The classification of inconsistencies by a security rule category is a basis for multi-module SEC architecture. We define five policy rule categories: authentication, authorizations, filtering, channel protection, and operational rules.

Inconsistencies between policy rules can be classified into two types: inside one category and between different categories. Let us describe at first *inconsistencies within one policy category*.

The inconsistencies for the authentication policy appear when for the same subject-object pair irrelevant authentication types are provided. There is the admission that allowable authentication rules do not contradict each other: two authentication rules for the same subject-object pair provide alternative authentication ways.

The authorization policy regulates access management. An authorization rule is specified as a group of four elements – “subject-action-object-access right (allow/deny)”. Subject can be represented either by a role or a particular user. Object can be compound and contain the collection of sub-objects. The inconsistency appears if allowing and denying privileges are given or derived for some “subject-action-object” triple. There is a hierarchical structure with partial order relation on the set

of roles and users. The relation determines the priority between two roles or the membership of the user to the role. For each object and each action we consider authorization rules in which the object and the action participate. Each rule adds allowing or denying to one of the nodes of the roles hierarchy graph. For the nodes, having no corresponding rule, the access right is derived from the values of higher nodes.

The *filtering policy rule* allows or denies network traffic for the given source and destination addresses, ports, and a protocol type. As a refinement procedure result, the filtering policy looks like a set of access control lists (ACLs). Each ACL is designed for a certain node (firewall) that can filter network packets. Since the rules in ACL are ordered, there are no inconsistencies, which lead system to ambiguous state within particular ACL. The possible contradictions between rules from different ACLs can result in unexpected traffic filtering, but again do not lead the security system to ambiguous state. So, the filtering policy can contain only anomalies when there exist rules which filtering conditions intersect. If these rules are in the same ACL, then the rule with the greater ordinal number is not activated. If the rules are in different ACLs and a packet goes through both firewalls, then the possible anomalies of superfluous or contradictory packet filtering can exist.

The anomalies of *channel protection policy* are concerned with incorrect assignment of security protocols (IPSec, SSL/TLS, etc). For instance, SEC checks that IPSec tunnels are either alternate or nested. Furthermore, since the channel protection policy contains filtering rules, filtering policy anomalies are to be searched.

The *operational rules* are specified in SPL as system method calls, when certain conditions are held (for instance, at some time point). Usually, operational rules start or stop services. Inconsistency appears when the conditions of operational rules are held simultaneously but actions are contradictory or to be done in certain sequence.

Let us now briefly describe *inconsistencies between policy categories*.

Almost all inconsistencies between categories are anomalies: the rules from one category can never be activated because of the rules from another one. The exception is the case when some rule prevents an operational rule action performing. In the last case the inconsistency is a conflict.

The responsibilities of some security categories are independent, and their rules can not contradict. Since authorization is defined only for authenticated users, authorization rules can not prevent authentication. In turn, authentication rules cannot contradict authorization rules, because access privilege does not depend on authentication type. Channel protection rules specify a traffic encryption protocol. They can not affect authentication and authorization procedures and operational rules.

The following inconsistencies can appear between policy categories:

- *Authentication – filtering.* The contradiction appears, for instance, if user can not send an authentication request to a service because of a denying filtering policy. Authentication rules cannot prevent a filtering policy activity, because user's authentication does not affect traffic filtering.
- *Authentication – operational rules.* Inconsistency appears, for instance, if operational rule's action requires authentication, but authentication is not allowed. In turn, operational rule can make authentication senseless if it stops the service.
- *Authorization – filtering.* Anomaly is a situation when authorization policy allows access to a service, but filtering policy denies access to the corresponding port of the service.
- *Authorization – operational rules.* Conflict appears if there are not enough privileges to perform an operational rule action. Similarly to authentication situation, access allowance to a service is senseless if an operational rule stops it.
- *Filtering – channel protection.* Since the channel protection policy contains filtering, the search of the anomalies for the given policies combination is similar to the searching of anomalies in the filtering policy.
- *Filtering – operational rules.* A conflict appears when an action of operational rule is defined on a service while access to the service is denied by the filtering policy. In other side, if an operational rule stops firewall then the filtering will be stopped on the given node. This situation can be considered as anomaly and the list of the filtering rules, which will not be activated, are to be sent to the administrator.

3 Security Checker Architecture

The methods and techniques for security policy verification as well as the types of conflicts and anomalies that are possible to detect and resolve are identified through SEC developing. These studies as well as practical and complexity considerations stipulated the methods to implement in SEC verification modules (Fig.1).

The input data for verification are the data from XML repository – the specifications of the secured system (in SDL) and the security policy (in SPL). Those specifications are transformed to the internal representation by the Parser component. The verification manager controls verification process. Verification modules perform major work on the detection and

resolution of inconsistencies. As the result of their work the following output is generated: verification result (whether inconsistencies are detected), information about inconsistencies identification and a set of commands for the policy modification.

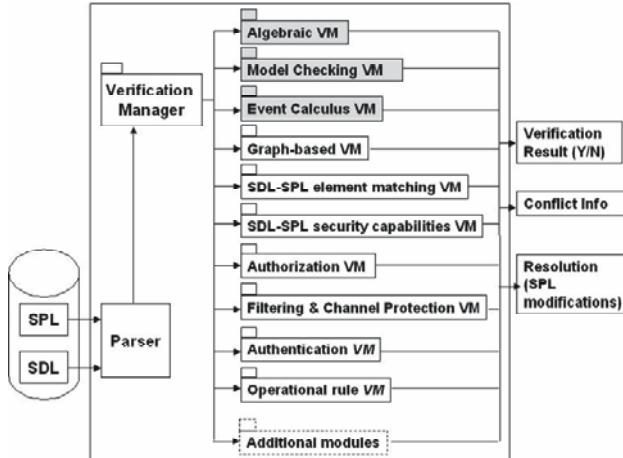


Fig. 1. Security checker verification modules

Modules are classified into two types by the level of generalization (cardinality of handled inconsistencies set): general purpose modules and specialized ones.

Three general purpose modules (grayed in Fig.1) implement a common approach for detection and resolution of inconsistencies, including ones between the rules of different categories and the inconsistencies of the policy application to the system description. One of the modules is based on algebraic approach [6], another one uses model checking, third one applies an abductive procedure for Event Calculus [25]. These modules work less effective than the specialized ones since they use more general approach to formalization of the security policy, system description and inconsistencies specification. However, these modules manage any security policy inconsistencies identified.

Specialized modules implement algorithms that are appropriate to particular security categories: authorization, filtering, channel protection and operational.

SEC architecture allows uploading new verification modules.

4 Event Calculus Verification Module

One of theorem proving methods, prospective for security policy verification, represents a combination of *Event Calculus and abductive search*. The Event Calculus (EC) [25] formalizes the common sense principle of inertia: “normally, nothing changes”. It states that *fluent* (time-varying property of the world) holds at particular *timepoints* (real or integer numbers) if it was initiated by an *event* occurrence at some earlier timepoint and was not terminated by another event occurrence in meantime. Similarly, a fluent does not hold at a particular time-point if it was previously terminated and was not initiated in the meantime. Fluents, events and timepoints are sorts of first-order language used for event calculus formal representation.

To describe a basic calculus, the following seven predicates are introduced:

- *InitiallyTrue(f)* – fluent f is true at initial time;
- *InitiallyFalse(f)* – fluent f is false at initial time;
- *Happens(e, t)* – action e occurs at time t ;
- *Initiates(e, f, t)* – if e occurs at t it will initiate fluent f ;
- *Terminates(e, f, t)* – if e occurs at t it will terminate fluent f ;
- *HoldsAt(f, t)* – fluent f is true at time t ;
- *Clipped(t1, f, t2)* – the fluent f is terminated between times $t1$ and $t2$, where f is a variable over fluents, e is a variable over events, and $t, t1, t2$ are variables over timepoints.

The following variant of domain-independent EC axiomatics uses circumscription for the most convenient use in the abductive proof procedure [23, 12]:

- $\text{HoldsAt}(f, t) \equiv [\text{Happens}(e, t1) \wedge \text{Initiates}(e, f, t1) \wedge t1 < t \wedge \text{not}(\text{Clipped}(t1, f, t2))] \vee [\text{InitiallyTrue}(f) \wedge \text{not}(\text{Clipped}(0, f, t))] \vee [t=0 \wedge \text{InitiallyTrue}(f)]$.
- $\text{InitiallyTrue}(f) \equiv \text{not}(\text{InitiallyFalse}(f))$.
- $\text{Clipped}(t1, f, t2) \equiv \text{Happens}(e, t) \wedge t1 < t \wedge t < t2 \wedge \text{Terminates}(e, f, t)$.

Having as an input the formula that expresses inconsistent system state, the abductive proof procedure defines a sequence of events, that may lead the system to this state.

Verification consists in the system behavior modeling. For this purpose the state of network elements, the security system and inconsistencies should be formalized as fluents, and events that initiate and terminate these fluents should be defined. Fluents and events are related through domain-depended EC axiomatics.

The event Calculus module generalized representation is depicted in Fig.2.

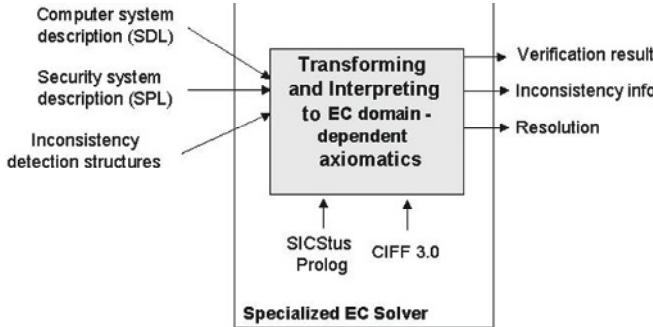


Fig. 2. Generalized representation of Event Calculus module

The input parameters are the computer system description, the policy description and inconsistencies description. The module outputs the “YES/NO” verification result, inconsistency information, including type and conflicting rules, and modified policy.

The module uses the CIFF [11] abductive proof procedure implemented in CIFF 3.0 [10]. This implementation uses SICStus Prolog [38].

The main theoretical problem is the development of domain-dependent axiomatics. This axiomatics includes formalization of all input data: SDL and SPL descriptions, and the definition of inconsistencies. SPL rules are translated to an axiom set that shows the security system reaction to events occurred in the network. Therefore, additionally to policy rules, there are events that activate one or another security system action.

For example, *RequestAuthentication* event with three parameters (subject, authentication type, object) and *Authenticated* fluent with the same parameters are used for authentication rule representation. Axiom general form looks like this:

Initiates(RequestAuthentication (Identity, AuthenticationMethod, Target), Authenticated (Identity, AuthenticationMethod, Target), ?t).

Thus, *RequestAuthentication* event, occurred in time point $?t$, initiates *Authenticated* fluent for the corresponding triple subject-method-object.

Similarly, event, that represents user exit from the system, deactivates authentication fluent:

Terminates(Logout(Identity,Target),Authenticated(Identity,?Authenticatemethod,Target),?t).

For the authorization rules representation, four main types of events are used: user joining the role, privilege assignment for the role and two events of opposite actions. Fluents are relations “user-role” and “role-privilege”.

Similarly, events, fluents, and axioms are defined for other policy rule categories.

The following main verification principle is implemented in this module. System behavior can be expressed in domain-dependent axioms based on fluents whose states are changed by event occurrences. Abductive proof procedure takes as input the axiomatics and query that represents an inconsistent system state and the outputs number of events that can lead to this state. If the scenario is found, rules participated in this scenario are analyzed and one of resolution strategies is applied.

5 Model Checking Verification Module

The generalized representation of the Model Checking verification module is presented in Fig.3. The input parameters and the output for the module are similar to the Event Calculus module.

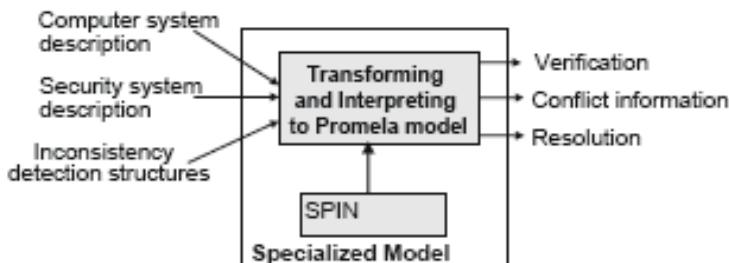


Fig.3. Generalized representation of Model Checking verification module

The main theoretical problem of the module implementation is the construction of a model that simulates behavior of network, security system and subjects, making requests for network resources.

Verification of security policy consists in the model checking of constructed model and further analysis of output. Inconsistencies are expressed in linear temporal logic and added to the model as conditions for consistent states.

The module provides for an inconsistency type and the sequence of events that lead to inconsistent state. Then this trace is analyzed in order to define contradictory rules involved, and one of appropriate resolution strategy is applied.

Fig.4 represents the abstract scheme of Model Checking module operation.

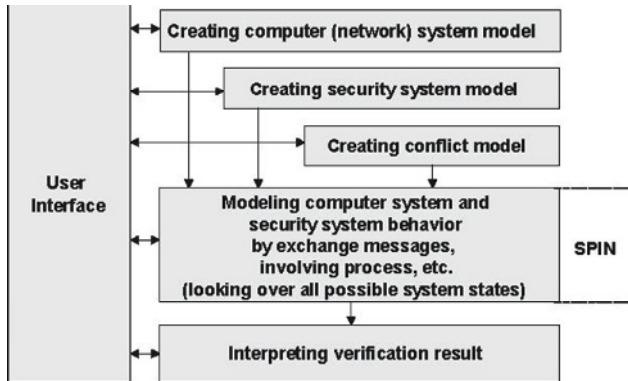


Fig. 4. Abstract scheme of Model Checking verification module operation

In accordance with the scheme given in Fig. 4, operation includes the following steps. Computer network model described in SDL language is built in Promela language. Then security system is represented in Promela and the syntactical constructions are added in the form of formal statements for inconsistencies detection. The built model is initiated and SPIN simulation is started. The system switches to the new state with the new set of values at each step and the formal violation statement is checked. All detected violations are interpreted as inconsistencies.

SPIN is a tool for verification of dynamic systems models. So its use is appropriate when inconsistencies cannot be detected via static analysis of SPL/SDL, i.e. an inconsistency occurs only if some dynamic pre-condition is satisfied.

For example, SPIN verification module detects authorization conflict. It occurs if the following conditions are met: (1) there are two rules with the same “object-action” pair; (2) roles used in these rules are different; (3) privileges defined for the roles are opposite.

There introduced a potential conflict with describe conflict situation that can occur if a user assigns to two roles having opposite privileges. Potential conflicts are detected statically while parsing SPL.

Let us consider the example of authorization conflict modeling.

The following Promela listing introduces potential conflict and some additional data structures:

```

typedef PotentialConflict {
    mtype roles[2]
};

PotentialConflict conflicts[1];

```

```

chan req = [0] of {mtype};
mtype = {Student, Administrator};
bool userRoles[2];

```

PotentialConflict type describes potentially conflicting roles. The conflicts array is to be initialized by potentially conflicting pairs. Its size is a number of such pairs. The variable req (channel type) is used for communication between process describing user and process describing virtual RoleAssignerservice. The type mtype defines roles involved into potential conflicts. Array userRoles[2] represents roles, which user can be assigned to.

The following listing describes the process of user request for role assignment:

```

proctype user () {
    printf("user process started!\n");
    if
        :: req ! Student;
        :: req ! Administrator;
    fi;
    if
        :: req !Student;
        :: req ! Administrator;
    fi
}

```

User process randomly tries to assign two roles (Student or Administrator) from list of potentially conflicting ones. Success of such attempt depends on behavior of RoleAssigner process. If there are no restrictions on role assignment, its behavior is trivial. Any request to assign role is satisfied. If there are some restrictions on role assignment specified in SPL, the behavior of RoleAssigner should be more complicated. There may be a lot of different types of restrictions on role assignment: prioritized roles, deactivating the role upon assigning another role, prohibition to assign some roles simultaneously.

The procedure of role assignment is represented in the following listing:

```

proctype RoleAssigner () {
    mtype r;
    printf("RoleAssigner process started!\n");
end:
    do

```

```

    :: req ? r ->
        printf("received: %d\n", r);
        userRoles[r-Administrator] = 1;
    od
}

```

This part of model initializes conflicts array and starts RoleAssigner() and user() processes. Conflicts array is assigned roles involved into potential conflicts.

The following listing describes the process of model initialization:

```

init {
printf("STARTED\n");
d_step {
    conflicts[0].roles[0] = Administrator;
    conflicts[0].roles[1] = Student;
}
    run RoleAssigner();
    run user();
}

```

In the init procedure the potentially conflicting roles (Administrator и Student) are set and the processes of role assignment and role request of user (run RoleAssigner() and run user()) are initialized.

The last part of the example is used for inconsistency detection:

```

#define p0 (userRoles[conflicts[0].roles[0]-Administrator] &&
userRoles[conflicts[0].roles[1]-Administrator])
never { /* <>p */
    skip;
T0_init:
    if
        :: ((p0)) ->           printf("# of condition:
0\n");
        goto accept_all
        :: (1) -> goto T0_init
        fi;
accept_all:
    skip;
}

```

The clause `Never` checks for conflict occurrence.

6 Specialized Modules

Specialized modules deal with inconsistencies within a particular security rules category. Let us describe two specialized modules.

The *authorization specialized module* detects and resolves authorization conflicts, i.e. if there exist two rules that define allowing and denial privileges for the same "subject-action-object" triple. The module is based on the Flexible Authorization Framework (FAF) [22] approach for access control analysis. The advantage of FAF is a deep consideration of object and subject hierarchies. SEC uses FAF privilege propagation algorithms and conflict resolution strategies.

The *filtering and channel protection specialized module* searches for filtering and IPSec anomalies [1, 15]. It detects inter-firewall and intra-firewall anomalies. Inter-firewall anomalies are ones that appear within access control list of one firewall. They are: shadowing (rules have different actions and higher priority rule matches all the packets that lower priority one), generalization (rules have different actions and lower priority rule matches all the packets that higher priority one does), redundancy (one rule can be removed and security policy will not be affected), correlation (rules have different actions and there are some packets that first and second rules match).

Intra-firewall anomalies may appear between firewalls that filter the same traffic. They are: shadowing (upstream firewall blocks the network traffic accepted by a downstream firewall), spuriousness (upstream firewall permits the network traffic denied by downstream firewall), correlation (there are two correlated rules in the upstream and downstream firewalls), and redundancy (downstream firewall denies the network traffic already blocked by an upstream firewall).

The module also detects IPSec tunnel and multi-transform anomalies. Incorrectly overlapping tunnels produce unsecured traffic on some channel part. The multi-transform anomaly appears when two rules match a common flow and a weaker transform is on the top of a stronger one.

7 Functioning of Security Checker and Graphical User Interface

The example of SEC main form is represented in Fig.5. It allows choosing the security policy for verification and verification modules loading. Verification modules are listed in configuration file. The verification can be started after choosing verified policy and corresponding modules. The verification modules generate logs that contain information about verification process and results. Logs can be viewed by pressing the “View log” button. The verification result form is given in Fig.6.



Fig.5. SEC Main form

All detected inconsistencies are represented due to this form. It depicts for every inconsistency: the name of a verification module that detected the inconsistency; the type of inconsistency; conflicting rules; the status of inconsistency solvability and resolution strategies if any is available. One can choose the resolution strategy for every inconsistency and try to solve it.

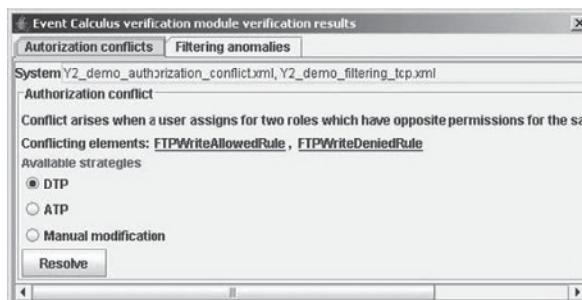


Fig. 6. Verification result form

Every verification module produces a log containing information about verification process duration, constructed model in the module's language (Prolog for Event Calculus, Promela for SPIN), and verification results including inconsistencies and involved rules.

8 Conclusion

The paper considered the developed *approach, models and software for the verification of computer network security policies*. The developed models and software tool “Security Checker” are aimed at eliminating possible inconsistencies in the specifications of policy rules, and also at determining a possibility of implementing these rules in the predefined computer network. The approach is based on multi-modular architecture, and the implementation of conflict detection, and resolution mechanisms. It uses as (1) general purpose modules (based on event calculus and abductive reasoning, and also on model checking) allowing to detect the majority of contradictions (including dynamic), as well as (2) specialized modules working effectively on the particular types of contradictions.

Analysis of current approaches to security policy verification allowed to distinguish general propose methods applicable to the detection and resolving of most inconsistencies and specialized techniques specially intended for operating with particular inconsistencies.

All inconsistencies between the rules of security policy as well as ones between security policy and system description were classified into two types. One type of inconsistencies is conflicts preventing protection system against application of specified policy. Another type of inconsistencies is anomalies causing policy rules to never be activated by a protection system. It should be noted that from user's standpoint anomalies are not less important because in such case security system does not implement all requirements specified.

The classification of inconsistencies by security categories allows unambiguously describing contradictions that can appear during policy creation and modification. Based on this classification the SEC architecture is developed and implemented.

Due to the fact that policy verification precedes a stage of its distribution, Security Checker considerably decreases number of possible inconsistencies in a policy distributed via network. Security Checker is a debugger of security policy in one's way, allowing an administrator to avoid conflicts and anomalies in interactive mode.

Security Checker architecture is multi-modular. Verification manager controls which modules and in which sequence will be used for verification. Exclusively at inconsistency presence detection, verification modules determine inconsistency type and suggest a set of commands to add or remove rules. General purpose modules based on event calculus, model checking and algebraic approach for some conflict types are less effective than suggested specialized modules. System administrator is able to activate the modules needed for verification.

EC Verification is implemented in Java programming language. Module based on event calculus and abductive inference uses software library CIFF 3.0 working on the basis of SICStus Prolog. Model checking verification module works on the basis of SPIN model checker. SEC components work on both Windows and UNIX platforms.

Acknowledgments

This research is being supported by grant of Russian Foundation of Basic Research, program of basic research of the Department for Informational Technologies and Computation Systems of the Russian Academy of Sciences (contract No 3.2/03), Russian Science Support Foundation and partially funded by the EU within the POSITIF project (contract No. IST-2002-002314) and RE-TRUST (contract No. 021186-2).

References

1. E.Al-Shaer, H.Hamed, R.Boutaba, M.Hasan: Conflict classification and analysis of distributed firewall policies. IEEE Journal on Selected Areas in Communications, Vol.23 (10) (2005)
2. Alur, R., Henzinger, T.A., Mang, F.Y.C., Qadeer, S., Rajamani, S.K., Tasiran, S.: Mocha: Modularity in model checking. Proceedings of the Tenth International Conference on Computer-aided Verification (CAV 1998), Lecture Notes in Computer Science, Springer-Verlag, Vol.1427 (1998)
3. Alur, R., Anand, H., Grosu, R., Ivancic, F., Kang, M., McDougall, M., Wang, B.-Y., de Alfaro, L., Henzinger, T.A., Horowitz, B., Majumdar, R., Mang, F.Y.C., Meyer-Kirsch, C., Minea, M., Qadeer, S., Rajamani, S.K., Raskin, J.-F.: Mocha User Manual. JMocha Version 2.0.
<http://embedded.eecs.berkeley.edu/research/mocha/doc/j-doc/>
4. Avizienis, A.: Dependability and Its Threats.
<http://citeseer.ist.psu.edu/705929.html>

5. Bandara, A.K., Lupu, E.C., Russo, A.: Using Event Calculus to Formalise Policy Specification and Analysis. 4th IEEE Workshop on Policies for Distributed Systems and Networks (Policy 2003) (2003)
6. Basile, C., Liou, A.: Towards an algebraic approach to solve policy conflicts. FCS'04, Turku (Finland) (2004)
7. Cambridge University and TU Munich: Isabelle. <http://isabelle.in.tum.de>
8. Cimatti, A., Clarke, E., Giunchiglia, E., Giunchiglia, F., Pistore, M., Roveri, M., Sebastiani, R., Tacchella, A.: NuSMV Version 2: An OpenSource Tool for Symbolic Model Checking. Proceedings of the International Conference on Computer-Aided Verification (CAV 2002), LNCS, Springer-Verlag, V.2404 (2002)
9. Clarke, E.M., Wing, J.: Formal methods: state of the art and future directions. ACM Computing Surveys: Special ACM 50th anniversary issue: strategic directions in computing research. Vol.28, No.4 (1996)
10. The CIFF Proof Procedure for Abductive Logic Programming. <http://www.doc.ic.ac.uk/~ue/ciff>
11. Endriss, U., Mancarella, P., Sadri, F., Terreni, G., Toni, F.: The CIFF Proof Procedure: Definition and Soundness Results. Technical Report 2004/2, Department of Computing, Imperial College London (2004)
12. Fung, T.H., Kowalski, R.A.: The IFF Proof Procedure for Abductive Logic. Programming Journal of Logic Programming, Vol.33, No.2 (1997)
13. Gordon, M., Melham, T.: Introduction to HOL: A theorem proving environment for higher order logic. Cambridge University Press (1993)
14. Group, T.V.: VIS: A system for Verification and Synthesis. Proceedings of the 8th International Conference on Computer Aided Verification, Lecture Notes in Computer Science, Springer-Verlag, Vol.1102 (1996)
15. Hamed, H., Al-Shaer, E., Marrero, W.: Modeling and verification of IPSec and VPN security policies. IEEE ICNP'05 (2005)
16. Hartel, P.H., van Eck, P., Etalle, S., Wieringa, R.: Modelling Mobility Aspects of Security Policies. CASSIS 2004 (2004)
17. Holzmann, G.J.: Design and Validation of Computer Protocols. Englewood Cliffs, N.J.: Prentice Hall (1991)
18. Holzmann, G.J.: The Model Checker SPIN. IEEE Transaction in Software Engineering, Vol.23, No.5 (1997)
19. IBM Formal Methods Research Group: IBM Formal Methods Home Page. <http://www.haifa.il.ibm.com/projects/verification/FormalMethods-Home>
20. Intel Strategic CAD Labs. <http://www.intel.com/research/scl>
21. Jackson, D., Rinard, M: Software Analysis: a Roadmap. Proceedings of 2000 ICSE (2000)
22. Jajodia, S., Samarati, P., Sapino, M.L., Subrahmanian, V.S.: Flexible support for multiple access control policies. ACM Transaction Database Systems, Vol.26, No.2 (2001)
23. Kakas, A.C., Kowalski, R.A., Toni, F.: Abductive Logic Programming. Journal of Logic and Computation, Vol.2, No.6. (2003)

-
- 24. Kaufmann, M., Moore, J.: An Industrial Strength Theorem Prover for a Logic Based on Common Lisp. *IEEE Transactions on Software Engineering*, Vol.23, No.4 (1997)
 - 25. Kowalski, R.A., Sergot, M.J.: A logic-based calculus of events. *New Generation Computing*, Vol.4 (1986)
 - 26. Lupu, E., Sloman, M.: Conflicts in Policy-based Distributed Systems Management. *IEEE Transactions on Software Engineering*, Vol.25, No.6, 1999.
 - 27. Madhavapeddy, A., Mycroft, A., Scott, D., Sharp, R.: The case for abstracting security policies. *International Conference on Security and Management (SAM)*, CSREA Press, Vol.1 (2003)
 - 28. McMillan, K.: The SMV System.
http://www.cs.cmu.edu/_modelcheck/smv.html
 - 29. Mitchell, J.C., Mitchell, M., Stern, U.: Automated analysis of cryptographic protocols using Murphi. *Proceedings of IEEE Symposium on Security and Privacy* (1997)
 - 30. Mitchell, J.C., Shmatikov, V., Stern, U.: Finite-State Analysis of SSL 3.0. *Proceedings of 7th USENIX Security Symposium* (1998)
 - 31. Nalumasu, R., Gopalakrishnan, G.: PV: an Explicit Enumeration Modelchecker. *Formal Methods in Computer Aided Design FMCAD'98*. Lecture Notes in Computer Science, Springer-Verlag, Vol.1522 (1998)
 - 32. On-The-Fly, LTL Model Checking with SPIN. <http://netlib.bell-labs.com/netlib/spin/whatispin.html>
 - 33. Powell, D., Deswarte, Y.: On Dependability Concepts with respect to Deliberately Malicious Faults. <http://citeseer.ist.psu.edu/480547.html>
 - 34. Randell, B.: Dependability-a unifying concept. *Computer Security, Dependability, and Assurance: From Needs to Solutions*. IEEE Computer Society (1999)
 - 35. Schneider, F.B.: Enforceable security policies. *ACM Transactions on Information and System Security*, Vol.3, No.1 (2000)
 - 36. SICStus Prolog. <http://www.sics.se/isl/sicstuswww/site/index.html>
 - 37. Westerinen, A., Strassner, J., Scherling, M., Quinn, B., Herzog, S., Huynh, A., Carlson, M., Perry, J., Waldbusser, S.: Terminology for Policy-Based Management (RFC 3198). www.rfc-archive.org/getrfc.php?rfc=3198

Architecture Types of the Bit Permutation Instruction for General Purpose Processors

Moldovyan A. A., Moldovyan N. A., Moldovyanu P. A.

Specialized Center of Program Systems, SPECTR,
Kantemirovskaya Str. 10, St. Petersburg 197342, Russia
E-mail: nmold@cobra.ru
Web: www.cobra.ru

Abstract. In large information systems different data transform algorithms including the bit permutation operations requiring an execution of great number of cycles are used. To increase significantly the software performance of such algorithm a controlled bit permutation instruction (BPI) is desirable. Here a question of justification of embedding a new command, controlled BPI, into the standard set of instructions of general-purpose processor for increasing the efficiency of different types algorithms implemented in software is studied. In a variety of applications two different types of bit permutation operations are required: arbitrary fixed permutations and variable permutations. The last are used in a new fast cipher designs based on data-dependent permutations. Accounting for an expediency of embedding the controlled permutation command into the set of elementary processor operations the cryptographic applications form only one of the motivation elements. Another strong motivation is BPI's use for solving variety of non-cryptographic problems. The multipurpose architecture of the BPI operation oriented to the efficient execution of both the cryptographic functions based on data-driven permutations and the algorithms including arbitrary bit permutations is proposed.

Keywords: Cryptography, fast software encryption, controlled bit permutation, new CPU instruction

1 Introduction

In Geographical Information Systems (GIS) and many other large information systems different algorithms of data transform use the bit permutation operations that, in general case, require an execution of great number of cycles. A prominent example is the bit-reversal permutation used in number of Fast Fourier Transform (FFT) algorithms. In a general purpose microprocessor a bit-reversal permutation operation requires an execution of 50 or more cycles. Due to use of bit permutation operations in different types of algorithms the last have comparatively low performance at their

software implementation. This defines an actual problem of embedding a universal bit permutation instruction (BPI) in the general purpose processors. One can consider two types of bit permutations: variable (data-driven) permutations and arbitrary fixed permutations. The first type of permutations is the basic primitive in the new class of ciphers based on data-dependent permutations. The second type is used both in cryptographic and the non-cryptographic algorithms. In papers [1-4] several design versions of the controlled permutation command for embedding in the general-purpose processor are introduced. The proposed BPI architectures can be attributed to the following two types: 1) for non-cryptographic applications requiring performing arbitrary bit permutations that are fixed, while executing a specified algorithm and 2) for cryptographic applications requiring performing data-dependent permutations that are variable, i. e. the permutations vary, while executing a specified encryption algorithm. However, in the first case the bit permutations operation is performed in several cycles and in the second case the probability that a required arbitrary fixed permutation can be executed is rather small. Therefore the practice needs a universal BPI architecture combining benefits of the two considered cases. Such BPI architecture is attractive for processors' manufacturers due to its expediency, since such new instruction becomes a multi-purpose instruction imparting to the processors new essential properties that are very attractive for users.

The present work proposes a universal architecture of the BPI operation combining the universality and the efficiency in cryptographic applications.

The paper is organized as follows. In Section 2 the notion of switchable controlled permutations is introduced and a structure of the BPI for cryptographic applications is presented. Section 3 discusses several approaches to the BPI design and Section 4 introduces several designs including the advanced BPI architecture that provides performing in one cycle both the data-driven and arbitrary fixed permutations. Section 5 concludes the paper.

2 Structure of BPI for Cryptographic Applications

The development of controlled permutation command both for the general-purpose processor and the microcontrollers of special purpose attracts practical interest. In the first case the unquestionable demand is the implementation of bit permutations variants, which are used for solving the sufficiently wide range of problems. In the second case comparatively

special permutation variants can be implemented. Considering a construction of encryption algorithms and hash-functions one can digress from the special type permutations. It is sufficient to specify not specific permutations but the ones, providing for good avalanche effect and sufficient uniformity of a probability distribution of bits transition from one position to another. The example of such command is $P^{(e)}_{32/96}$ operation used in the development of COBRA-F64a and COBRA-F64b algorithms [3]. The $P^{(e)}_{32/96}$ operation architecture is described as follows.

The boxes performing controlled bit permutations are denoted as $P_{n/m}$, where n denotes the bit-size of its input and m denotes the bit-size of the control input. Such boxes are usually constructed using elementary switching elements $P_{2/1}$ as elementary building blocks. The $P_{2/1}$ box is controlled by one bit v and forms two-bit output (y_1, y_2) , where $y_1 = y_{1+v}$ and $y_2 = y_{2-v}$. A $P_{n/m}$ box can be represented as a superposition

$$P_{n/m} = L^{(V_1)} \circ \pi_1 \circ L^{(V_2)} \circ \pi_2 \circ \dots \circ \pi_{s-1} \circ L^{(V_s)}$$

where L is an active layer composed of $n/2$ switching elements, V_1, V_2, \dots, V_s are controlling vectors of the active layers from 1 to s , and $\pi_1, \pi_2, \dots, \pi_{s-1}$ are fixed permutations. The inverse CP box has the following structure:

$$P^{-1}_{n/m} = L^{(V_s)} \circ \pi_{s-1}^{-1} \circ L^{(V_{s-1})} \circ \pi_{s-2}^{-1} \circ \dots \circ \pi_1^{-1} \circ L^{(V_1)}$$

Let $\{0,1\}^{(s)}$ denote the set of all binary vectors $U = (u_1, u_2, \dots, u_s)$, where $\forall i \in \{1, \dots, s\} \quad u_i \in \{0,1\}$. While being concatenated the components V_1, V_2, \dots, V_s compose the controlling vector of the $P_{n/m}$ box: $V = (V_1, V_2, \dots, V_s) = (v_1, v_2, \dots, v_n)$, where v_i are control bits and $\forall i \in \{1, \dots, n\}$ the v_i bit controls the i -th switching element. We suppose that the switching elements are numbered consecutively from left to right and from top to bottom in the P -boxes. In the P -boxes they are numbered from left to right and from bottom to top. Examples of typical CPB are presented in Fig. 1.

Note, that due to symmetric structure the mutual inverses $P_{32/96}$ and $P^{-1}_{32/96}$ differ only in the distribution of controlling bits over the boxes $P_{2/1}$ in the same topology. When performing DDP operations with $P_{32/96}$ 96-bit controlling vector depending on some 32-bit data sub-block is formed. Let L be a controlling data sub-block. Thus, bits of $L = (l_1, \dots, l_{32})$ are used on the average three times while defining the controlling vector. When de-

signing respective extension box it is reasonable to use the following criteria:

Criterion 1. Let $X = (x_1, \dots, x_{32})$ is the input vector of the $\mathbf{P}^{(V)}_{32/96}$ -box. Then for all L and i the bit x_i should be permuted depending on six different bits of L .

Criterion 2. For all i the bit l_i should define exactly three bits of V .

Below we use the extension box \mathbf{E} providing the following relation between V and L :

$$V_1 = L_l; \quad V_2 = L_l^{>>>6}; \quad V_3 = L_l^{>>>12}; \quad V_4 = L_r; \quad V_5 = L_r^{>>>6}; \quad V_6 = L_r^{>>>12}$$

where $L_l = (l_1, \dots, l_{16})$, $L_r = (l_{17}, \dots, l_{32})$, and $Y = X^{>>>k}$ denotes rotation of the n -bit word X by k bits, where we have $y_i = x_{i+k}$ for $1 \leq i \leq n-k$ and $y_i = x_{i+k-n}$ for $n-k+1 \leq i \leq n$. Due to symmetric structure of $\mathbf{P}_{32/96}$ its modifications $\mathbf{P}^{(V)}_{32/96}$, where $V = (V_1, V_2, \dots, V_6)$, and $\mathbf{P}^{(V')}_{32/96}$, where $V' = (V_6, V_5, \dots, V_1)$ are mutually inverse. Such symmetry can be used in order to construct switchable CP boxes. This idea can be realized using a simple transposition box $\mathbf{P}^{(e)}_{96/1}$ implemented as some single layer CP box consisting of three parallel single-layer boxes $\mathbf{P}^{(e)}_{2 \times 16/1}$ (Fig. 2). Input of each $\mathbf{P}^{(e)}_{2 \times 16/1}$ -box is divided into 16-bit left and 16-bit right inputs. The box $\mathbf{P}^{(e)}_{2 \times 16/1}$ contains 16 parallel $\mathbf{P}^{(e)}_{2/1}$ -boxes controlled with the same bit e . For example, $\mathbf{P}^{(0)}_{2 \times 16/1}(U) = U$ and $\mathbf{P}^{(0)}_{2 \times 16/1}(U) = U' = (U_r, U_l)$, where $U = (U_l, U_r) \in \{0, 1\}^{32}$. The left (right) inputs of the $\mathbf{P}^{(e)}_{2/1}$ -boxes correspond to the left (right) 16-bit input of the box $\mathbf{P}^{(e)}_{2 \times 16/1}$. If the input vector of the box $\mathbf{P}^{(e)}_{96/1}$ is (V_1, V_2, \dots, V_6) , then at the output of $\mathbf{P}^{(e)}_{96/1}$ we have $V' = (V_1, V_2, \dots, V_6)$ (if $e = 0$) or $V' = (V_6, V_5, \dots, V_1)$ (if $e = 1$). Structure of the switchable CP box $\mathbf{P}^{(L, e)}_{32/32}$ is shown in Fig. 2.

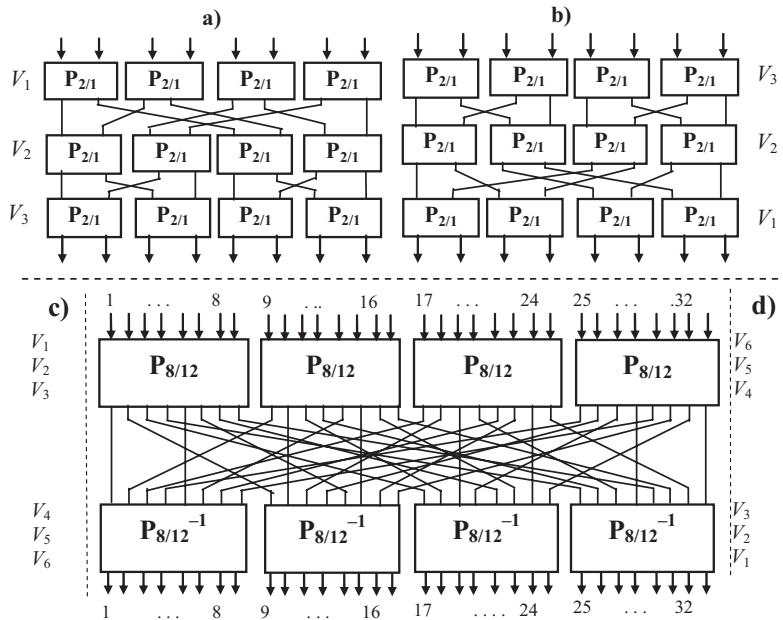


Fig. 1. The CP boxes $P_{8/12}$ (a), $P_{8/12}^{-1}$ (b), $P_{32/96}$ (c), and $P_{132/96}^{-1}$ (d)

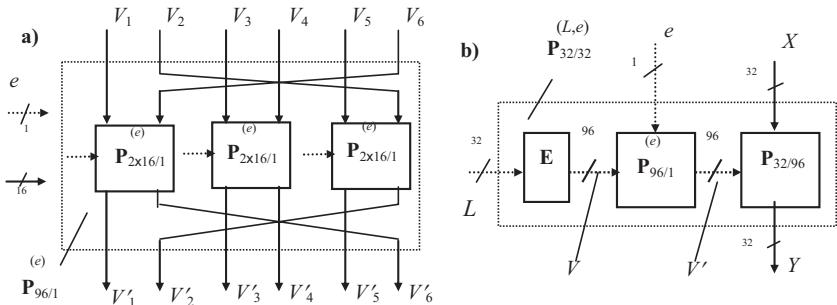


Fig. 2. Switchable CP boxes $P_{96/1}^{(e)}$ (a) and $P_{32/32}^{(e)}$ (b)

3 Approaches to Construction of the Controlled Permutation Command

While designing an architecture of BPI oriented to embedding into the general-purpose processors, it appears expedient to develop such controlled permutation command, that can be effectively used in implementation of cryptographic algorithms on its base and also for solving the wide range of non-cryptographic problems. The most important variants of controlled permutation blocks (CPB), including switchable CPB, have rather low cost of hardware implementation and relatively low latency. Therefore it is interesting to discuss the variant of implementing such command that could implement all possible permutations. For this case appears to be important the fact that for CPB of maximum order constructed according to the recursive scheme of order doubling one can easily calculate the value of control vector [4], allowing any given permutation be implemented.

The number of all possible permutations is great (is equal to $n!$, where n is the size of CPB input in bits), thus, it is practically impossible to compose the tables for all of them in practically important cases $n = 32$ and $n = 64$. However, for most frequently used permutations the tables of corresponding control vectors can be composed. Then indicating one or another value of control vector as a parameter of controlled permutation command it would be possible to implement any required permutations. At that it will not be difficult to elaborate the program to calculate the necessary value of control vector for arbitrary permutations, not indicated in the table of widely used permutation types.

CPB of maximum order are also preferable for cryptographic applications. However, the control vector in such applications repeatedly changes during the transformation process of one input block. In case of large length of control vector this brings about the necessity of several machine cycles for loading the value of control vector till the current modification of permutation will be executed. This fact can significantly reduce the speed of cryptographic transformations executed using a controlled permutation command. For cryptographic use the command not requiring a successive loading of control value is preferable.

Simultaneous achievement of these two objectives can be attained with the use of a configurable architecture of controlled permutation command. Since two main variants of such command use are presumed: 1) cryptographic applications and 2) applications that require the execution of large set of preset permutations, it is sufficient to introduce one additional control bit according to which «cryptographic» or «multi-purpose configura-

tion» will be adjusted. There exists one more important aspect, which consists in the fact that for «cryptographic» configuration the use of switching mode from direct controlled permutation to the inverse one via changing the value of certain other bit e , which determines the mode of operation, is presumed.

For «multi-purpose» configuration such switching in principle is not required, since the control vector implementing the inverse permutation can be calculated in the same way as for direct permutation. Nevertheless the reloading of control parameter will lead to delay equal to several machine cycles. In case of algorithms including the periodic usage of mutually inverse permutation pairs this delay will accumulate, thus, reducing the efficiency of algorithm. In such cases the use of possibility of switching from direct permutation to inverse one via inverting e bit can significantly increase the speed of mentioned algorithms' operation.

4 Main Types of Architectures of the Controlled Bit Permutation Command

Let us examine main variants of construction schemes of controlled permutation command for a general-purpose processor. The most important cases relate to the use of switchable CPB $\mathbf{P}^{(e)}_{32/96}$, $\mathbf{P}^{(e)}_{32/144}$, $\mathbf{P}^{(e)}_{64/192}$ and $\mathbf{P}^{(e)}_{64/352}$, which have symmetric topology (see [3] and [6], where the construction of switchable CPB with symmetric structure is described in detail). The scheme of switchable unit of controlled permutations for these two cases is shown in Fig. 3. The base of this unit is CPB with symmetric topology $\mathbf{P}_{n/m}$, which is represented in the form of superposition of direct permutation $\mathbf{P}_{n/m}$, fixed permutation involution and inverse permutation $\mathbf{P}_{n/m}^{-1}$:

$$\mathbf{P}_{n/m} = \mathbf{P}_{n/m'} \bullet \mathbf{P}_{n/m'}^{-1}$$

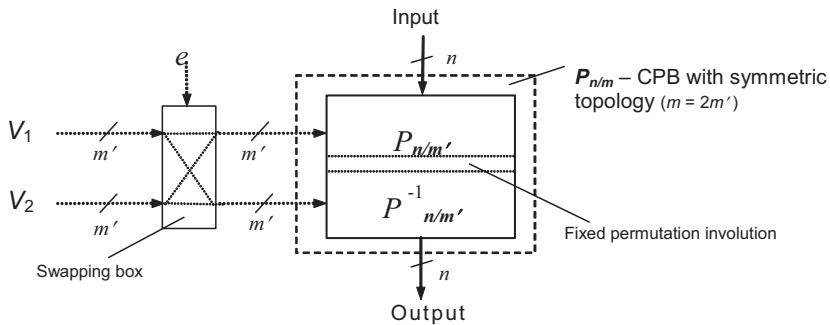


Fig. 3. Schematic representation of switchable controlled permutations unit on the basis of CPB with symmetric topology

P_{32/96} and **P**_{64/192} blocks are not the blocks of maximum order, therefore the unit under consideration can be applied in case of construction of controlled permutation command of «cryptographic» orientation, for example, in case of application in microcontrollers. In this case the unit under consideration should be supplemented with expansion block implemented in the form of simple interconnection of conductors. Expansion block input is the input for one of the operands of controlled permutation operation (see Fig. 4). The value of the control n -bit data sub-block can be input there and will be transformed into control vector $V = (V_1, V_2)$ by the expansion block.

This scheme of construction of controlled permutation command can be expanded in case of application of symmetric block of maximum order as **P** _{n/m} block. Indeed the superposition of two mutually inverse blocks of first order controlled by independent binary vectors forms the resultant block of maximum order [3]. Therefore **P** _{n/m'} and **P** _{n/m'} ⁻¹ blocks in the indicated above scheme must be the blocks of the first order and the fixed permutation involution must be the identity substitution. In this case we shall have the controlled permutation command based on the switchable CPB of maximum order.

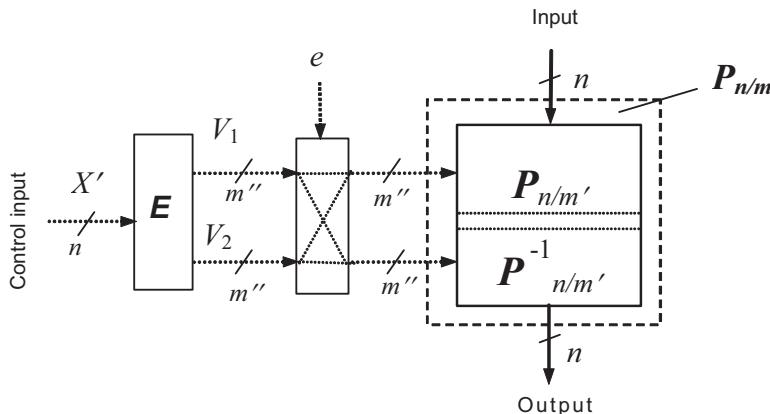


Fig. 4. Scheme of construction of controlled permutation command for cryptographic applications

In this variant the input signal passes the number of active layers equal to $2\log_2 n$. The implementation of the block of maximum order with the number of layers equal to $2\log_2 n - 1$ has the symmetric topology as well, therefore there exist more economical variant of implementation of universal command, in which $P_{n/m'}$ and $P_{n/m}$ blocks are not the blocks of first order, but the additional active layer L is arranged between them. Between the indicated active layer and $P_{n/m'}$ and $P_{n/m}^{-1}$ CPB the fixed permutations π_4 and π_4^{-1} of such sort are arranged, and two following superpositions $P_{n/m'} \bullet \pi_4 \bullet L$ and $L \bullet \pi_4^{-1} \bullet P_{n/m}^{-1}$ represent mutually inverse blocks of the first order. Such topology of the CPB is well known (see [5]). In this case the transposition block of control vectors V_1 and V_2 is simplified, since V_{add} component of control vector corresponding the central active layer L , remains invariable during the execution of direct and inverse permutation of maximum order. In case of construction of controlled permutation command for multi-purpose applications we shall assume, that more economical implementation of switchable CPB of maximum order is used.

The usage of such block of maximum order is of principal importance for constructing the controlled permutation command of universal type. In this case it is required to use instead of the expansion block the shift register designated for input and storage of control vectors with arbitrary values (in the previous scheme not all values of control vectors can be input on the control input of $P_{n/m}$ block). In view of fact that bit capacity of data bus ($n = 32$ or 64) is significantly less than the size of control vector ($m = 144$ or 352), it is required to execute five or six steps of writing of sufficiently long control value for input of the control vector in the register R .

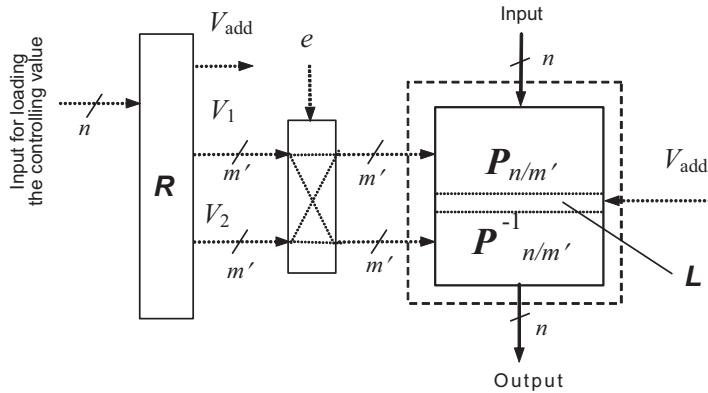


Fig. 5. Scheme of construction of controlled permutation command for multi-purpose applications

In view of fact that one can calculate the value of control vector for inverse permutations, in principle it is possible avoid the transposition block of V_1 and V_2 .vectors. Such variant of command structure is shown in Fig. 6.

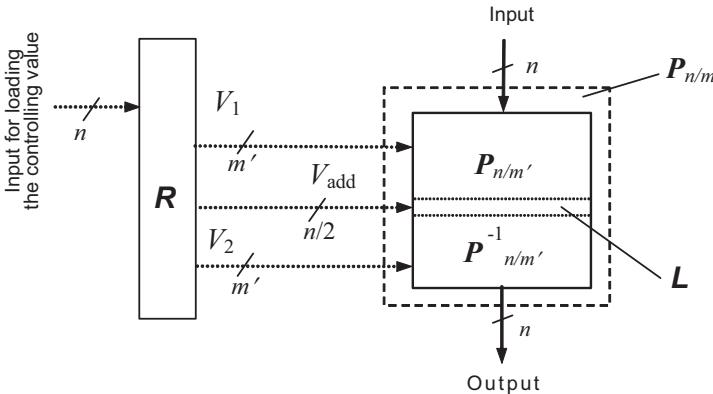


Fig. 6. Simplified scheme of construction of controlled permutation command for «multi-purpose» applications

It is possible to combine the advantages of two previous variants of construction of controlled permutation command using the unit setting the configuration features of a given command. The structure of configurable command is shown in Fig. 7. Both the register for input and storage of control vector of arbitrary type and the expansion block, which forms at its

output only the subclass of control vectors according to the value of n -bit control data sub-block, are used there.

Transposition block \mathbf{T}' controlled by the bit of configuration choice (e' bit) inputs on the transposition block of control vector components (block \mathbf{T}) either the output value of register \mathbf{R} (for example, if $e'=1$) or the output value of expansion block \mathbf{E}' (if $e'=0$). In the first case the arbitrary permutation can be implemented, at that the preliminary loading of corresponding control value in the register \mathbf{R} is required. In the second case the preliminary loading of register is not required, but not the all variants of bit permutations of input transformable vector are implemented.

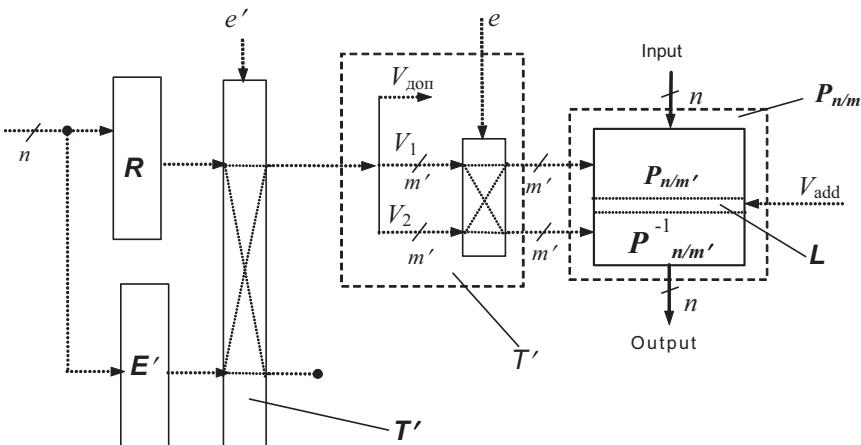


Fig. 7. Scheme of construction of universal controlled permutation command

In the universal BPI architecture it is supposed to use the $\mathbf{P}_{n/m}$ box of maximum order, i. e. the BPB that can specify arbitrary bit permutation. Examples of such boxes are presented in [5] as the $\mathbf{P}_{32/144}$ box for the $n=32$ case and as the $\mathbf{P}_{64/352}$ box for the $n=32$ case. The $\mathbf{P}_{32/144}$ box contains nine active layers and the \mathbf{E}' extension box should satisfy criteria 1 and 2 mentioned in section 2. For this purpose one can propose the following distribution of the 32-bit control data block bits:

$$\begin{aligned}
 V_1 &= (l_1, l_2, l_1, l_2, l_3, l_4, l_3, l_4, l_3, l_4, l_1, l_2, l_1, l_2); \\
 V_2 &= (l_5, l_6, l_6, l_5, l_7, l_7, l_8, l_8, l_8, l_8, l_7, l_7, l_5, l_5, l_6, l_6); \\
 V_3 &= (l_9, l_{10}, l_{13}, l_{14}, l_9, l_{10}, l_{13}, l_{14}, l_{10}, l_9, l_{13}, l_{14}, l_{10}, l_9, l_{13}, l_{14}); \\
 V_4 &= (l_{11}, l_{12}, l_{11}, l_{12}, l_{15}, l_{16}, l_{15}, l_{16}, l_{15}, l_{16}, l_{15}, l_{16}, l_{11}, l_{12}, l_{11}, l_{12}); \\
 V_5 &= (l_{15}, l_{15}, l_{16}, l_{16}, l_{27}, l_{28}, l_{27}, l_{28}, l_{27}, l_{28}, l_{27}, l_{15}, l_{16}, l_{15}, l_{16}); \\
 V_6 &= (l_{27}, l_{28}, l_{27}, l_{28}, l_{31}, l_{32}, l_{31}, l_{32}, l_{31}, l_{32}, l_{27}, l_{28}, l_{27}, l_{28}, l_{27});
 \end{aligned}$$

$$\begin{aligned}
 V_7 &= (l_{25}, l_{26}, l_{29}, l_{30}, l_{25}, l_{26}, l_{29}, l_{30}, l_{26}, l_{25}, l_{29}, l_{30}, l_{26}, l_{25}, l_{29}, l_{30}); \\
 V_8 &= (l_{21}, l_{22}, l_{22}, l_{21}, l_{23}, l_{23}, l_{24}, l_{24}, l_{24}, l_{24}, l_{23}, l_{23}, l_{21}, l_{21}, l_{22}, l_{22}); \\
 V_9 &= (l_{17}, l_{18}, l_{17}, l_{18}, l_{19}, l_{20}, l_{19}, l_{20}, l_{19}, l_{20}, l_{19}, l_{17}, l_{18}, l_{17}, l_{18}),
 \end{aligned}$$

where the control data block is represented as $V_1 = (l_1, l_2, \dots, l_{32})$. It is easily to see that this distribution is symmetric relatively the fifth active layer, therefore it can be used in the BPI designs presented in Fig. 3 and Fig. 4 for the case of the maximum order box $\mathbf{P}_{n/m} = \mathbf{P}_{32/144}$. The BPI architecture presented in Fig. 7 provides for a possibility to change in one cycle an arbitrary bit permutation to its inverse.

It is reasonable to evaluate the Implementation cost of various BPI architecture variants that can be evaluated i) as a number of standard switching elements $\mathbf{P}_{2/1}$ or ii) as a number of NAND gates used for forming the combinational circuit that implements the BPI operation. Evaluations of the hardware implementation cost for different variants of the BPI operation are shown in Tables 1 and 2.

Table 1. Evaluation of the implementation cost for various variants of controlled permutation command (the number of used elementary switches P2/1 is indicated)

Topology	Application	$\mathbf{P}_{n/m}$	\mathbf{T}	\mathbf{T}'	Total
$\mathbf{P}^{(e)}_{32/32}$	Cryptographic	96	48	—	144
$\mathbf{P}^{(e, e')}_{32/144}$	Multi-purpose	144	64	144	352
$\mathbf{P}^{(e)}_{64/64}$	Cryptographic	192	96	—	288
$\mathbf{P}^{(e, e')}_{64/352}$	Multi-purpose	352	160	352	864

Table 2. Evaluation of the implementation cost in NAND gates

Topology	Application	$\mathbf{P}_{n/m}$	\mathbf{T}	\mathbf{T}'	Total
$\mathbf{P}^{(e)}_{32/32}$	Cryptographic	576	288	—	864
$\mathbf{P}^{(e, e')}_{32/144}$	Multi-purpose	864	384	864	2112
$\mathbf{P}^{(e)}_{64/64}$	Cryptographic	1152	576	—	1728
$\mathbf{P}^{(e, e')}_{64/352}$	Multi-purpose	2112	960	2112	5184

5 Conclusion

All considered variants of architectures of controlled permutation command have relatively low cost of the hardware implementation and can be used for embedding in microcontrollers and general-purpose processors. The most interesting variant is the architecture combining the universality of application in the non-cryptographic algorithms using fixed bit permutations and the effectiveness in cryptographic algorithms based on data-

driven permutations. The advanced BPI architecture provides for a possibility of execution in one cycle both the variable and arbitrary fixed permutations. This variant seems to be the most attractive one for the manufacturers of processors, since it gives new qualities and capabilities to processors that facilitate the enhancement of possibilities of their application. Applying a fast multi-purpose BPI to a variety of data transforms used in GIS will increase significantly the performance of data processing algorithms.

References

1. R.B. Lee, Z.J. Shi, X. Yang, Efficient permutation instructions for Fast Software Cryptography. IEEE Micro. 2001, vol. 21, no 6, pp. 56-69.
2. R.B. Lee, Z.J. Shi, R.L. Rivesr, M.J.B. Robshaw, On permutation operations in Cipher Design. Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04), Las Vegas, Nevada, April 5-7, 2004, vol. 2, p. 569-579.
3. N.A. Moldovyan, A.A. Moldovyan. Innovative cryptography.- Charles River Media, Boston, Massachusetts, 2006.- 386 pp.
4. N.A. Moldovyan, P.A. Moldovyanu, D.H. Summerville. On Software Implementation of Fast DDP-Based Ciphers. International Journal of Network Security. 2007. vol. 4, no. 1. P.81-89 (<http://isrc.nchu.edu.tw/ijns/>)
5. N.A. Moldovyan, A.A. Moldovyan, Sklavos N. Controlled Elements for Designing Ciphers Suitable to Efficient VLSI Implementation. // Telecommunication Systems. 2006. vol. 32, no 2/3. P. 149-163.

New Public Key Cryptosystems Based on Difficulty of Factorization and Discrete Logarithm Problems

Moldovyan N. A.

Specialized Center of Program Systems, SPECTR, Kantemirovskaya Str. 10, St. Petersburg 197342, Russia
E-mail: nmold@cobra.ru
Web: www.cobra.ru

Abstract. Electronic messages authentication issue is of significant importance for geographical information systems. A number of public key cryptosystems base on RSA modulus (n) has been proposed and none of them provides a possibility to use its public key to perform the RSA encryption. Present paper introduces a new cryptoscheme whose public key can be also used to perform the RSA encryption and signing procedures, its own procedures are different though. The signature (S) verification formula in new scheme is $\alpha \Rightarrow S^M \bmod n$, where M is the signed document and pair (n, α) is a public key. Requirements imposed on parameters n and α are discussed. Public encryption, signature generation, and key agreement protocol based on new cryptoscheme are considered. Using the (n, α) public key a new signature formation mechanism was proposed and it was used to design the digital signature scheme (DSS) described by the following verification equation: $g - k \not\in \alpha^{kgM} \bmod n \bmod \delta$, where (g, k) is signature and δ is a prime number. Finally the new signature formation mechanism to design a set of new short signature DSS based on difficulty of finding discrete logarithm are used. Also it is shown the Schnorr's DSS is incorporated as a particular case.

Keywords: Digital signature, public encryption, public key cryptosystem, key agreement protocol

1 Introduction

Systems designed for electronic documents authentication in large information systems such as geographical information systems (GIS) are based on public key cryptosystems. The RSA public key cryptosystem introduced by R.L. Rivest, A. Shamir, and L.M Adleman in 1978 [1] has become the first world wide used digital signature and public-key encryption system. In RSA the public key is represented by pair of numbers (n, e) , where $n = pq$ is the product of two randomly chosen distinct prime num-

bers and e is a random number that is relatively prime with Euler phi function $\phi(n) \neq p - 1)(q - 1)$. The triple (p, q, d) , where $d = e^{-1} \bmod \phi(n)$, is secret. Data ciphering with RSA is described as follows:

$$C = M^e \bmod n \quad (\text{public-key encryption}) \quad \text{and} \quad M = C^d \bmod n \quad (\text{decryption}),$$

where $M < n$ is a plaintext, C is ciphertext. RSA signature (S) generation and verification are performed, correspondingly, as decryption and encryption:

$$S = M^d \bmod n \quad (\text{generation}) \quad \text{and} \quad M = S^e \bmod n \quad (\text{verification}).$$

Usually the signed documents are comparatively long, and the hash function values $H = F_H(M)$ corresponding to the document: $S = H^d \bmod n$ are signed. Signing the H values prevents some attacks exploiting the RSA multiplicative property [2]. Some other attacks against the RSA are connected with the case of small encryption [3, 4] and small decryption [5] exponents. Cycling attack against RSA [6] is not practical, but it can be used by an attacker engaged in process of developing software implementing RSA. The listed attacks are theoretical ones, and RSA is successfully used in practice. The RSA security is based on difficulty of factoring modulus n , which depends on the structure of primes p and q . At present the requirements to the generating primes p and q procedure are well clarified [6, 7].

Some other public-key cryptosystems based on the RSA modulus n [8, 9] has been also proposed, however none of them gained the RSA's fame. The keys of those cryptosystems are "incompatible" with RSA, though it is highly desirable to get a wider use and recognition of the possibility to apply the public key of a new cryptosystem to perform the RSA encryption, decryption, and signing procedures.

In this paper a new RSA-modulus-based public-key cryptosystem is introduced, which is the RSA "compatible".

This work is organized as follows: in Section 2 a new digital signature scheme is described. In Section 3 the public key encryption based on new scheme is presented. In Section 4 the key agreement protocol based on new scheme is proposed. Section 5 compares new cryptosystem with RSA and shows its "compatibility" with RSA, i. e. that its public key can also be used to perform correctly the RSA signing and ciphering algorithms. Section 6 discusses the requirements imposed on the parameters of the new cryptoscheme. In Section 7 another new digital signature scheme (DSS) with the RSA modulus is presented. In the second DSS a new signa-

ture formation mechanism is used that provides for a possibility to reduce significantly the signature size. Also an example of utilizing the new mechanism in the design of the DSS based on difficulty of finding discrete logarithm is considered. Finally, conclusions are given in Section 8.

2 New Digital Signature System

2.1 Basic Scheme

New digital signature scheme uses a modulus of the form $n = pq$, where p and q are strong primes easy to generate using Gordon's algorithm [2, 10]. The primes p and q are supposed to be of large size $|p| \approx |q| \geq 512$ bits, where $|W|$ denotes the binary representation length of the integer W . Gordon's algorithm allows to generate strong primes p and q for which the numbers $p - 1$ and $q - 1$ contain different prime divisors γ' and γ'' , respectively. The γ' and γ'' values should be of some required length, for example from 100 to 160 bits (see Section 6 for details). The secret key is the triple (p, q, γ) .

The public key is a pair of numbers (n, α) , where α is generated as follows. Select random number β that is simultaneously primitive element modulo p and primitive element modulo q , compute $t = \gamma'^{-1} \gamma''^{-1} \phi(n) = \gamma'^{-1} \gamma''^{-1} (p-1)(q-1)$ and $\alpha = \beta^t \pmod{n}$. Number α is a generator of the γ -order group $\{\alpha, \alpha^2 \pmod{n}, \dots, \alpha^{\gamma'-1} \pmod{n}, \alpha^{\gamma'} \pmod{n}\}$ i. e. $\alpha^{\gamma'} \pmod{n} \neq 1$. Generation of α parameter can be also performed in the following way:

1. Choose random $\beta < n$ and calculate $\sigma = \beta^t \pmod{n}$.
2. If $\sigma \neq 1$ and $\gcd(\sigma - 1, n) \neq 1$, then $\alpha \leftarrow \sigma$, otherwise go to step 1.

A document or hash value corresponding to it is interpreted as integers M and H , correspondingly. Thus, where required the g -bit sequence $\mathbf{W} = (w_{g-1}, w_{g-2}, \dots, w_1, w_0)$ is taken as number

$$W = w_{g-1}2^{g-1} + w_{g-2}2^{g-2} + \dots + w_22^2 + w_12^1 + w_0.$$

Signature generation. To sign a document M the following signature generation procedure is to be performed:

1. Using the document as input of some hash function generating hash value H of the size $160 \leq |H| \leq 256$ bits calculate the hash value $H = F_H(M)$ where F_H is a hash function.
2. Check whether $H \neq 0$, $H \neq 1$, and $\gcd(H, \gamma) \neq 1$. If $H = 0$, $H = 1$, or $\gcd(H, \gamma) \neq 1$, then modify document M and go back to step 1.
3. Calculate: $U = H^{-1} \pmod{\gamma}$.
4. Calculate signature S : $S = \alpha^U \pmod{n}$.

Signature verification. To verify signature S do the following:

1. Calculate hash function $H = F_H(M)$.
2. Check whether the following signature verification equation $\alpha = S^H \pmod{n}$ is satisfied. If $S^H \pmod{n} \neq \alpha$, then reject the signature.

Proof that signature verification works:

$$S^H \equiv (\alpha^U)^H \equiv (\alpha^{H^{-1}})^H \equiv \alpha^{H^{-1}H \pmod{\gamma}} \equiv \alpha \pmod{n}.$$

If size of document is small enough, i. e. $M < \gamma$, then M can be used instead of H : $S = \alpha^{M^{-1}} \pmod{n}$ and $\alpha = S^M \pmod{n}$.

2.2 Possible Attacks and Modified Signature Schemes

The signature scheme described in Section 2 can be attacked as follows. Attacker presents for signing the hash value $H = H_1H_2$. If he gets the signature

$$S = \alpha^{H^{-1}} = \alpha^{H_1^{-1}H_2^{-1}} \pmod{n}$$

corresponding to H , then he is able to derive signatures corresponding to the hash values H_1 and H_2 : $S_1 = S^{H_2} \pmod{n}$ и $S_2 = S^{H_1} \pmod{n}$. Signatures

S_1 and S_2 satisfy the verification equation: $S_1^{H_1} = S^{H_2H_1} = S^H = \alpha \pmod{n}$

and $S_2^{H_2} = S^{H_1H_2} = S^H = \alpha \pmod{n}$. Thus, one should never sign a random document or hash function submitted by a stranger. Hash value corresponding to the document should be calculated just before signature generation.

Above mentioned theoretical attack can be thwarted modifying the signature verification equation as follows: $(S + H)^H = \alpha \pmod{n}$ (1) or

$S^{H + H^{\leftarrow 7}} = \alpha \bmod n$ (2). In the last equation the operation H^{\leftarrow} is defined as follows: interpret the binary representation of number H as concatenation of 16-bit binary vectors, rotate to the left by 7 bits each 16-bit vector, interpret the result as binary number H^{\leftarrow} . The respective signature generation equation is $S \not\in \alpha^U - H \bmod n$, where $U = H^{-1} \bmod n$, (1) or $S = \alpha^U \bmod n$, where $U \not\in (H + H^{\leftarrow})^{-1} \bmod n$ (2). These modifications introduce negligible change in the time required for generating and verifying a signature. However initial version of the cryptoscheme is preferable for practical use, since the attacks considered above are theoretical ones.

3 Public-Key Encryption Algorithm

To encrypt a message $M \leq n - 1$ (if the message is long, then split it into data blocks M_i satisfying condition $|M_i| \neq 0$) perform the following algorithm.

1. Generate a random k , $1 \leq k \leq \gamma - 1$, and calculate data enciphering key $K = \alpha^k \bmod n$.
2. Encrypt the message M (or current data block M_i): $C = KM \bmod n$.
3. Encrypt the data encryption key K : $R = K^t \bmod n$, where t is a specified number $t \geq 2$.
4. Send the ciphertext (R, C) .

To decrypt ciphertext (R, C) do the following.

1. Calculate data enciphering key: $K = R^{1/t \bmod q'} \bmod n$.
2. Decrypt the cryptogram C : $K^{-1}C \bmod n = M$.

Proof that decryption is correct:

Since $K = \alpha^k \bmod n$, where $\gcd(\alpha, n) = 1$, then n and K are co-prime, i. e. $\gcd(K, n) = 1$. Therefore integer $K^{-1} \bmod n$ exists and can be easily calculated using extended Euclidean algorithm. Thus,

$$K^{-1}C \bmod n = K^{-1}(KM \bmod n) \bmod n \not\in K^{-1}KM \bmod n = M.$$

The public-key encryption described above is a probabilistic one, like ElGamal public-key encryption [2].

4 Key Agreement Protocol

To generate a common data encryption key users Alice and Bob do the following steps.

1. Alice generates a random integer k_A , $1 \leq k_A \leq \gamma$, calculates key $K_A = \alpha_B^{k_A} \bmod n_B$, transforms the key: $R_A = K_A^t \bmod n_B$, where t is a specified integer, and sends R_A to Bob.
2. Bob generates a random integer k_B , $1 \leq k_B \leq \gamma$, and calculates key $K_B = \alpha_A^{k_B} \bmod n_A$, transforms the key: $R_B = K_B^t \bmod n_A$, and sends R_B to Alice.
3. Alice deciphers Bob's key K_B : $K_B = R_B^{1/t \bmod \gamma_A} \bmod n_A$ and calculates the common data encryption key: $K_{AB} = K_A K_B$.
4. Bob deciphers Alice's key K_A : $K_A = R_A^{1/t \bmod \gamma_B} \bmod n_B$ and calculates the common data encryption key: $K_{AB} = K_A K_B$.
5. Bob generates a random integer R' and sends R' to Alice.
6. Using the key K_{AB} and a symmetric encryption algorithm E Alice encrypts R' : $C' = E_{K_{AB}}(R')$. Then she generates a random integer R'' and sends C' and R'' to Bob.
7. Using the key K_{AB} and a symmetric encryption algorithm E Bob encrypts R'' : $C'' = E_{K_{AB}}(R'')$ and sends C'' to Alice.
8. Using corresponding symmetric decryption algorithm D Alice checks whether $C'' = D_{K_{AB}}(R'')$. Bob also checks whether $C' = D_{K_{AB}}(R')$.

Bob's key K_B can be recovered from R_B only by Alice who knows the secret value γ_A . Alice's key K_A can be recovered from R_A only by Bob who knows the secret value γ_B . Thus, both parties can correctly perform steps 5-8 providing authentication of the data encryption key K_{AB} . To perform symmetric encryption the AES algorithm, for example, can be used. For this purpose one can split the bit string corresponding to $K_A K_B$ into respective number of 128-bit keys, for example, K_1, K_2, \dots, K_z . Then encryption E can be defined as z consecutive encryptions with AES: "Set $R_0 = R$; for $i = 1, 2, \dots, z$ do $R_i \leftarrow AES_{K_i}(R_{i-1})$. Output is $E(R) = R_z$ ".

One can also apply the following protocol to send a secret key via a public channel:

1. Alice generates a random integer k , $1 \leq k \leq \gamma$, calculates key $K = \alpha_B^k \bmod n_B$, transforms the key K as follows: $R = K^t \bmod n_B$,

generates the signature S_K corresponding to key K (or signature S_R corresponding to R), and sends values R and S_K (or sends R and S_R) to Bob.

2. Using R Bob recover the key K : $K_A = R_A^{-1/t \bmod \gamma_B} \bmod n_B$. Then, using Alice's public key (n_A, α_A) Bob verifies signature S_K (or signature S_R).

Thus, Bob gets the authenticated Alice's key K .

5 Compatibility with RSA

In the digital signature scheme described above the users can generate public keys (n, α) for which the α value satisfy the following two conditions: $\alpha^\gamma \bmod n \neq 1$ and $\gcd(\alpha, \phi(n)) \neq 1$. Trying several different values β (see Section 2) it is easy to find the required value α . Using such public key one can perform the RSA public-key encryption: $C = M^\alpha \bmod n$ and decryption $M = C^\delta \bmod n$, where $\delta = \alpha^{-1} \bmod \phi(n)$ is secret exponent.

Thus, one can combine new digital signature scheme with the RSA encryption/decryption. Besides, the public key (n, α) can be used to perform the RSA-based blind signing [11].

The blind signature protocol based on RSA uses multiplicative property of RSA. The protocol can be transformed into an attack, therefore RSA should be never used to sign a random hash value presented by a stranger [3]. New signature scheme is secured against such attacks. It is also problems free with small exponents, since it uses no small public key or small secret one.

The performance comparison results of the new scheme with RSA are presented in Table 1 for the modulus size $|n| = 2048$ bits. It is slower at a signature verifying but faster at signing than RSA. New scheme has better integral performance

Table 1. Average time required for signing and verifying (in arbitrary units)

RSA $(e =6)$		Proposed signature scheme	
		$ \gamma =160$	$ \gamma =256$
Sign	≈ 2048	≈ 160	≈ 256
Verify	≈ 16	≈ 160	≈ 256
Sign + Verify	≈ 2064	≈ 320	≈ 512

6 Requirements to Parameters n and α

To explain requirements imposed on parameters n and α it is useful to consider the case of prime value γ (for example: $\gamma \mid p - 1$ and $\gamma \nmid q - 1$) for which

$$\begin{aligned} \alpha &= \beta^{\frac{\phi(n)}{\gamma}} \not\equiv \beta^{\frac{(q-1)}{\gamma}} \mod n \Rightarrow \\ \Rightarrow \quad \alpha &\equiv (\beta^{\frac{(q-1)}{\gamma}})^{\frac{(p-1)}{\gamma}} \equiv 1^{\frac{(p-1)}{\gamma}} \equiv 1 \mod q \Rightarrow \\ \Rightarrow \quad \alpha - 1 &\equiv 0 \mod q \Rightarrow q \mid \alpha - 1 \Rightarrow \gcd(\alpha - 1, n) \neq 1. \end{aligned}$$

Thus, in the considered case it is possible to factorize modulus using extended Euclidean algorithm. Therefore some restrictions imposed on generating the public key are necessary. We can prevent this attack using a prime γ that divides both $p - 1$ and $q - 1$, but γ^2 does not divide $p - 1$ nor $q - 1$. In this case:

$$\alpha \equiv \beta^{\frac{(p-1)(q-1)}{\gamma^2}} \equiv \beta^{u'u''} \pmod{n},$$

where γ does not divide each of the numbers $u' = (p - 1)/\gamma$ and $u'' = (q - 1)/\gamma$. If β is simultaneously primitive element modulo p and primitive element modulo q , then $\alpha \mod p \neq 1$ and $\alpha \mod q \neq 1$, i. e. $\gcd(\alpha - 1, n) \neq 1$.

Unfortunately, in the case of prime secret element γ it can be calculated factorizing the $n - 1$ value. Indeed, we have: $p = u'\gamma + 1$, $q = u''\gamma + 1$, and $n = u'u''\gamma^2 + u'(u'' + 1)\gamma + 1$, hence $\gamma \mid n - 1$. Therefore the composite value $\gamma = \gamma'\gamma''$, where γ' and γ'' are different divisors of $p - 1$ and $q - 1$, should be used. If β is “double primitive element”, then

$$\alpha \equiv \beta^{\frac{(p-1)(q-1)}{\gamma'\gamma''}} \equiv \beta^{u'u''} \pmod{n},$$

where $u' = (p - 1)/\gamma'$ and $u'' = (q - 1)/\gamma''$. Thus, in such way of the public key formation $\gcd(\alpha - 1, n) \neq 1$. If one of the primes γ' and γ'' , for example, γ' , is small, then one can factorize n trying different values γ' and veri-

fying relation $\gcd(\alpha^{\gamma'} - 1, n) \neq 1$. Therefore both of the values γ' and γ'' should be sufficiently large.

7 DSS with Small Signature Length

7.1 Scheme based on factorization problem

The (α, n) public key can be used in another DDS based on a new signature formation mechanism that provides for a possibility to reduce significantly the signature size. This scheme is described by the following verification equation:

$$g + k \neq \alpha^{kgH} \pmod{n} \pmod{\delta},$$

where (g, k) is signature and $\delta > \gamma$ is a prime number having, for example, the length $|\delta| = |\gamma| + 8$ bits.

Signature generation is performed as follows:

1. Given the M document calculate the hash value $H = F_H(M)$.
2. Check whether $H \neq 0$ and $\gcd(H, \gamma) \neq 1$. If $H = 0$ or $\gcd(H, \gamma) \neq 1$, then modify document M and go back to step 1.
3. Select random $U < \gamma$ and calculate $Z \neq \alpha^U \pmod{n} \pmod{\delta}$ and $D = (Z^2 / 4 - U / H) \pmod{\gamma}$.
4. Check whether D is quadratic residue modulo γ . If not, then go back to step 3.
5. Solve the following system containing one congruence and one equation relatively unknowns g and k :
- 6.

$$\begin{cases} kgH \equiv U \pmod{\gamma} \\ g + k = Z \end{cases}$$

The solution gives the following signature generation formulas:

1. Calculate the signature using the formulas $g = Z / 2 \pm \sqrt{D} \pmod{\gamma}$ and $k = Z - g$.

Signature verification is performed as follows:

1. Calculate hash function $H = F_H(M)$.
2. Check whether the following signature verification equation $g + k \not\equiv \alpha^{kgH} \pmod{n}$ mod δ is satisfied.
If $g + k \not\equiv (\alpha^{kgH} \pmod{n}) \pmod{\delta}$, then reject the signature.

Proof that signature verification works:

The left part of the signature verification equation is equal to:

$$g + k = Z \not\equiv \alpha^U \pmod{n} \text{ mod } \delta.$$

The right part of the signature verification equation is equal to:

$$(\alpha^{kgH} \pmod{n}) \pmod{\delta} \not\equiv \alpha^U \pmod{n} \text{ mod } \delta = g + k,$$

$$\begin{aligned} \text{since } kgH &\equiv (Z - Z/2 \mp \sqrt{D})(Z/2 \pm \sqrt{D})H \equiv (Z^2/4 - D)H \equiv \\ &\equiv [Z^2/4 - (Z^2/4 - U/H)]H \equiv U(\pmod{\gamma}). \end{aligned}$$

In this DSS the signature length is significantly less than signature length in RSA, Rabin's scheme, and other known DSS based on factorization problem difficulty [2, 3, 7].

7.2 Scheme Based on Discrete Logarithm

The signature formation mechanism described above is very attractive to design DSS based on difficulty of finding discrete logarithm. Suppos, the public key $y = \alpha^x \pmod{p}$, where p is a large prime, α is a γ -order element modulo p , and x is the secret key. In this subsection it is assumed that γ is a not secret prime number and it has the length $160 \leq |\gamma| \leq 256$. The following verification congruence can be used to define a new DSS with small signature size:

$$k - g \equiv (y^k \alpha^{gH} \pmod{p}) \pmod{\gamma}.$$

The corresponding signature formation procedure consists in calculation using the following formulas:

$$k = \frac{ZH - U}{H - x} \bmod \gamma \text{ and } g = \frac{xZ - U}{H - x} \bmod \gamma,$$

where the $U < \gamma$ value is selected at random and then the Z value is calculated as $Z = \alpha^U \bmod p$. The last two formulae correspond to the solution of the following system of two congruences:

$$\begin{cases} kgH \equiv U \bmod \gamma \\ g + k = Z \end{cases}$$

derived from the verification congruence.

One can propose a modification of the last DSS by two verification equations in the following form:

$$R = y^k \alpha^g \bmod p \quad \text{and} \quad k - \lambda g = F_H(M | R),$$

where $\lambda \in \{0, 1\}$. For particular case $\lambda = 0$ we have well known Schnorr's signature scheme [12]. Thus, the proposed DSS is some extension of the Schnorr's scheme. For many practical applications of the DSS the hash function is calculated directly upon receiving the message. To prevent repeated message processing, while generating or verifying the signature it is efficient to define the following three variants of the DSS with short signature:

- i) $R = y^k \alpha^{gH} \bmod p \quad \text{and} \quad k \pm \lambda g = R \bmod \gamma,$
- ii) $R = y^k \alpha^g \bmod p \quad \text{and} \quad k = gR^H \bmod \gamma,$
- iii) $R = y^{kg} \bmod p \quad \text{and} \quad k = R^k \alpha^{gH} \bmod \gamma.$

The reader can easily deduce the corresponding signature generation procedures for cases i) and ii). In case three we have the following steps in the signature generation algorithm:

1. Select random $U < \gamma$ and calculate $k = Z \not\equiv \alpha^U \bmod p \bmod \gamma$.
2. Solve the following system containing two congruences relatively unknown g and t :

$$\begin{cases} t \equiv xgZ \pmod{\gamma} \\ Zt + gH \equiv U \pmod{\gamma} \end{cases}$$

3. The solution gives the following formula for calculating g :

$$g = \frac{U}{xZ^2 + H} \pmod{\gamma}.$$

8 Conclusion

This paper introduces a new public key cryptosystem based on the RSA modulus. Analogously to RSA, new cryptosystem gets its security from a difficulty of factorizing modulus and from difficulty of finding H -th root ($H \geq 2$) modulo a composite number n and discrete logarithm. Solving one of these difficult problems means breaking the cryptosystem. In new scheme the α element of public key (n, α) is a generator of the γ -order group and γ is secret element corresponding to α , whereas in RSA public exponent e is inverse modulo $\varphi(n)$ of the secret exponent d . Use of the α parameter as a part of the public key introduces new restrictions imposed on generating modulus n and a new security problem related to the α parameter generation.

New cryptosystem uses its own procedures for signature verification/signing, public key encryption, and key agreement. Due to its compatibility with RSA it can be combined in different ways with RSA using the same public key. Compatibility of new digital signature scheme with RSA cryptosystem is an attractive feature from practical point of view, since while combining RSA with new signature generation/verification scheme provides for more flexibility for security systems designers and users. To thwart some theoretic attacks a modification of the source digital signature scheme has been proposed.

Another signature formation mechanism based on solving a system containing one congruence and one equation was proposed at using the (n, α) public key. The new mechanism provides reducing the signature length to the value that is significantly less than the n modulus length. It also was shown that this mechanism could be efficiently applied to design DSS based on difficulty of finding discrete logarithm. A set of the short signature DSS is proposed, incorporating Schnorr's DSS as a particular case. All such DSS can be used as analogues to design new DSS implemented using elliptic curves. Such implementations will provide for significant

complexity reduction of the verification and signature generation procedures. The proposed DSS with short signature length are attractive to perform the authentication of electronic documents in GIS and other large information systems.

References

1. Rivest R.L., Shamir A., and Adleman L.M. A Method for Obtaining Digital Signatures and Public Key Cryptosystems, Communications of the ACM, 1978, vol. 21, n. 2, pp. 120-126
2. Menezes A. J., Vanstone S.A. Handbook of Applied Cryptography. CRC Press, 1996. -780 p.
3. Schneier B., Applied Cryptography: Protocols, Algorithms, and Source Code (Second Edition) -New York.: John Wiley & Sons. 1996. -758 p.
4. Hastad J. Solving simultaneous modular equations of low degree. SIAM Journal on Computing. 1988. vol. 17 pp. 336-404.
5. Wiener M.J. Cryptanalysis of short RSA secret exponents. IEEE Transactions on Information Theory. 1990. vol. 36. pp. 553-558.
6. Simmons G.J., Norris M.J. Preliminary comments on the M.I.T. public-key cryptosystem. Cryptologia. 1977. n. 1, pp. 406-414.
7. Pieprzyk J., Hardjono Th., Seberry J. Fundamentals of Computer Security. Springer-verlag. Berlin, 2003. - 677 p.
8. Rabin M.O. Digitalized signatures and public key functions as intractable as factorization. –Technical report MIT/LCS/TR-212, MIT Laboratory for Computer Science, 1979.
9. Fiat A., Shamir A. How to prove yourself: Practical solutions to identification and signature problems. Advances in cryptology –CRYPTO'86, Springer-Verlag LNCS, 1987, vol. 263, pp. 186-194
10. Gordon J. Strong primes are easy to find, Advances in cryptology –EUROCRYPT'84, Springer-Verlag LNCS, 1985, vol. 209, pp. 216-223.
11. Chaum D. Security without identification: Transaction systems to make big brother obsolete, Communications of the AMS, 1985, vol. 28, n. 10, pp. 1030-1044.
12. Schnorr C.P. Efficient signature generation by smart cards, J. Cryptology. 1991. vol. 4. pp. 161-174.

Application of a Dynamic Recurrent Neural Network in Spatio-Temporal Forecasting

Tao Cheng¹ and Jiaqiu Wang²

¹Department of Geomatic Engineering, University College London,
Gower Street, WC1E 6BT London, United Kingdom

²School of Geography and Urban Planning
Sun Yat-sun University, Guangzhou, P R China
{tao.cheng@ucl.ac.uk.;cafes123@163.com}

Abstract. Spatio-temporal data mining is the extraction of unknown and implicit knowledge, structures, spatio-temporal relationships, or patterns not explicitly stored in spatio-temporal databases. As one of data mining techniques, forecasting is widely used to predict the unknown future based upon the patterns hidden in the current and past data. In order to achieve spatio-temporal forecasting, some mature analysis tools, e.g., time series or spatial statistics, are extended to spatial or temporal aspect, respectively.

Among other methods, neural network is widely used for spatial forecasting. Normally a static forward neural network is employed to discover the hidden and deeply entangled spatial relationships. However, such approach is insufficient in forecasting dynamic process developing over space (such as forest fire). Elman is a kind of dynamic recurrent neural network (RNN) which allows the network to detect and generate time-varying patterns as well as spatial-varying patterns. Therefore, we use the Elman network for spatio-temporal forecasting. Experimental results collected from real case of forest fire area prediction confirm the viability and effectiveness of the proposed methodology.

1 Introduction

As one of data mining techniques, forecasting is widely used to predict the unknown future based upon the patterns hidden in the current and past data [1]. Spatio-temporal forecasting has been developed from individual spatial or temporal forecasting by modifying some mature analysis tools. For example, time series and spatial statistics are extended to spatial and temporal aspects, respectively. Spatio-temporal forecasting is gaining heavy attention for its promising performance in handling complex data in which not only spatial but also temporal characteristics must be taken into account. The challenge of spatio-temporal forecasting is how to integrate space and time seamlessly and simultaneously.

Among other methods, neural networks are widely used for forecasting [2], [3], [4]. Normally a static feedforward neural network based on back-propagation (BP) algorithm is employed to discover the hidden and deeply entangled spatial relationships [5]. However, such approach is insufficient in forecasting dynamic processes developing over space (such as forest fire), which usually change nonlinearly. Recurrent neural networks are more powerful than feedforward networks and they are appropriate for simulating dynamic systems since a recurrent neural network can consider both the parametric and structural learning (that is, the value of weights and appropriate topology of nodes and links) [6]. The recurrent neural network is suitable for a dynamic system with a form $y_t = f(y_{t-1}, y_{t-2}, \dots, y_{t-k}, x_t)$ where y_t and x_t are the output and the input at time t . They are also very appropriate when the input vector includes groups of components specified at several consecutive periods, $t, t-1, t-2, \dots, t-k$, as often occurs in time-series forecasting scenarios [7]. Therefore, they are called dynamic recurrent neural networks - DRNNs.

Several researchers have confirmed the superiority of DRNNs over static feedforward networks when performing nonlinear time series prediction [8], [9], [10], [11]. The main difference between a static feedforward back-propagation network and a dynamic recurrent network is the existence of a feedback mechanism in the nodes of the recurrent network. This feedback mechanism facilitates the process of using the information from the previous patterns along with the present inputs [12].

This paper presents an integrated spatial-temporal forecasting framework based upon a dynamic recurrent neural network. Section 2 presents the principle and structure of a dynamic recurrent neural network (DRNN) – Elman network. Section 3 proposes a framework of an integrated spatio-temporal forecasting by using the Elman network. It discusses the principle of the framework, problem definition and the algorithm of spatio-temporal forecasting, and the evaluation criteria of the prediction accuracy. The methodology proposed in Section 3 is illustrated by a case study of landscape-scale fire impacts in Canada in Section 4. The major finding and directions for future research are summarized in the last section.

2 A Dynamic Recurrent Neural Network – Elman Network

A dynamic recurrent neural network (DRNN) is a neural network with feedback connections. In a DRNN, the output depends not only on the current input to the network, but also on the current or previous input, out-

put or the state of the network. A DRNN can learn to map input sequences to output sequences. In principle DRNN can implement almost arbitrary sequential behavior. DRNNs are biologically more plausible and computationally more powerful than other adaptive models such as Hidden Markov Models (no continuous internal states), feedforward networks and Support Vector Machines (no internal states at all).

The topological structure of a dynamic recurrent neural network is shown in Figure 1. The synaptic connections of a DRNN unit in Figure 1 contain a self-recurrent connection that represents a weighted feedback signal of its state and lateral feedback connections. In terms of information processing, the feedback signals involved in a DRNN deal with some processing of the past knowledge and store current information for future usage. A DRNN unit has its own internal potential or internal state that is used to describe the dynamic characteristic of the network.

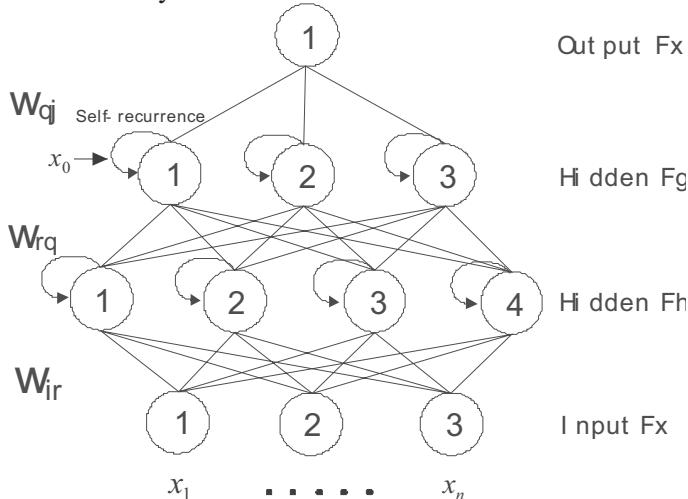


Fig. 1. Topological structure of a DRNN

Various works have been presented showing that DRNNs are quite effective in modelling non-linear dynamical systems [13], [14], [15]. The critical issue in the application of DRNN is the choice of the network architecture, i.e., the number and type of neurons, the location of feedback loops and the development of a suitable training algorithm.

Despite the great potential that dynamic recurrent neural networks hold, a successful practical application is conditioned by several drawbacks [16]. Drawbacks are: (1) the computational efficiency of the learning stage, which depends on the initial weights; (2) the information content of what is learned, which depends on the data set; (3) since they are black-boxes

models, the real structure of the network is not captured and hence not accessible [17].

Considering the number of DRNN topologies and training algorithms available, the choice of an appropriate pair (architecture and learning) is intimately dependent on the purposes and can be decisive for its success. The Elman network [18], one type of the DRNN, which is different from the feedforward neural network, is a globally feed-forward locally recurrent network [18]. The conceptual structure of Elman network is shown in Figure 2.

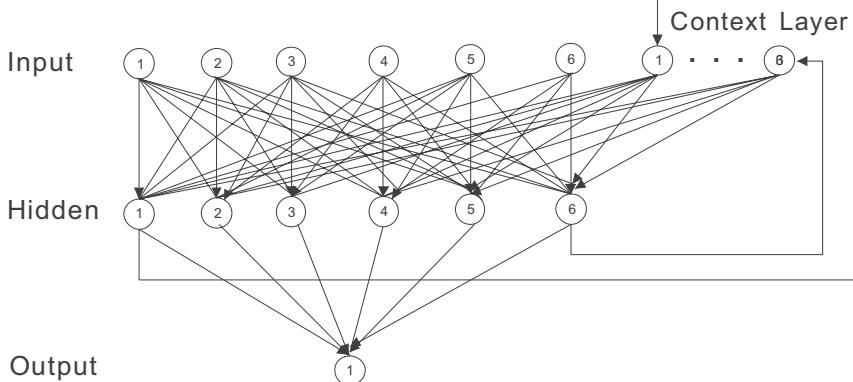


Fig. 2. Topological structure of the Elman network

Elman network consists of four layers: input, hidden, context, and output layer. z^{-1} is a one-step delay unit here. Especially, the context layer can store internal states. Moreover, Elman neural network has a backward connection between the hidden and context layer. This feedback path allows it to learn temporal and spatial patterns of input data, which is essentially useful in geographic spatial-temporal relationship prediction [19]. Thus, Elman neural network has certain dynamic characteristics over the feedforward neural network. On the other hand, the memory nodes and feedback connections increase its computational complexity, and generally result in a time-consuming training process. This is very important for Elman network and it is also our motivation to select this type of DRNN in this research.

The dynamic equation of Elman network can be described as below:

$$Y(k) = W_1 \sigma(W_2 X(k) + W_3 S(k-1) + W_o) \quad (1)$$

where $Y(k)$ is the network output, $X(k)$ is the network input, $S(k-1)$ is the output of hidden neurons at time $k-1$. W_1 is the output weighting and W_2 is the input weighting. W_3 is a weighting for hidden neurons' feedback and W_o is a network threshold. σ is a sigmoidal function.

The main difficulty related to the recursive training of recurrent networks arises from the fact that the output of the network and its partial derivatives with respect to the weights depend on the inputs since the beginning of the training process and on the initial state of the network. Several training algorithms have been proposed to adjust the weight values in DRNNs. Examples of these methods are the dynamic backpropagation from Narendra and Parthasarathy recurrent networks [20], the real time recurrent algorithm (RTRL) from Williams and Zipser [22] and the backpropagation through time (BPTT) from Werbos [21].

BPTT algorithm for training a recurrent neural network stems from the standard backpropagation algorithm [21]. It extends standard backpropagation algorithm so that it applies to dynamic systems. The central idea to BPTT is the unfolding of the discrete-time recurrent neural network into a static feedforward neural network (SFNN) each time a sequence is processed. Detailed BPTT algorithm can be seen in Werbos (1990). In this study, BPTT algorithm is considered.

3 An integrated Spatio-Temporal Forecasting Framework - STIFF

3.1 Principle

Recently, Li and Dunham proposed a spatio-temporal integrated framework - STIFF, which is applied to forecast the water flow rate at gauging station in the catchments [5]. In STIFF, time series analysis strategy is incorporated to capture the temporal correlations and the artificial neural network technique is employed to discover the hidden and deeply entangled spatial relationships, then the two mechanisms are combined via regression to generate the overall forecasting. It overcomes deficiency of previous works by loosening their stringent assumptions and excessive simplification.

However, STIFF approach is insufficient in forecasting dynamic processes developing over space, which cannot be handled by a static forward

neural network based on BP algorithm that STIFF employed. Therefore, the Elman network is adopted in our approach. Its recurrent connection allows the network to detect and generate time-varying patterns as well as spatial-varying patterns. Because the network can store information for future reference, it is able to learn temporal patterns as well as spatial patterns. To differentiate from STIFF, we call our approach as ISTIFF, i.e. improved STIFF, due to its improved forecasting accuracy.

The key idea of spatio-temporal forecasting is as follows: constructing a stochastic time series models to capture the temporal characteristics of each spatially independent subcomponent, then building a dynamic recurrent neural network (DRNN) to discover the hidden spatial correlation, finally combining the previous individual temporal and spatial forecasts based upon statistical regression to procure the final forecasting result.

3.2 Problem Definition

The spatio-temporal forecasting can be formally defined as follows [5]:

1. The research area Δ is composed of $n+1$ subcomponents denoted by $\alpha_0, \alpha_1, \dots, \alpha_n$, which can be spatially separated from each other. Without loss of generality, α_0 is assumed to be the only target subcomponent where the spatio-temporal forecasting will be conducted.
2. For each $\alpha_i \in \Delta$, there are j time series observations $\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{ij}$, that are recorded as Π_i for convenience.
3. Given the collection of subcomponents $\Delta = \{\alpha_0, \alpha_1, \dots, \alpha_n\}$, the whole available dataset $\Pi = \{\Pi_0, \Pi_1, \dots, \Pi_n\}$ and the look-ahead steps of s , the problem asks to find a mapping relationship f , defined as

$$f : \{\Delta, \Pi, l, s\} = f \left\{ \sigma_{0(l+1)}, \sigma_{0(l+2)}, \dots, \sigma_{0(l+s)} \right\} \quad (2)$$

which should be as precise as possible, where $l (=|a_i|)$, σ_{ij} ($i=0, 1, \dots, n$; $j=0, 1, \dots, l$) is the j^{th} observation in time series data σ_i .

3.3 Algorithm

The problem defined above can be solved by the algorithm with the following steps:

- *Step 1:* Define the forecasting problem in terms of the specification by determining the target subcomponent α_0 and its spatially-correlated siblings $\alpha_1, \alpha_2, \dots, \alpha_n$.
- *Step 2:* For each subcomponent $\alpha_i \in \Delta$ build a ARIMA (autoregressive integrated moving average) model TS_i . ARIMA is the most general model for time series forecasting which can be stationarized by transformations such as differencing and logging. It is calibrated by three parameters: p (the number of autoregressive terms), d (the number of non-seasonal differences), and q (the number of lagged forecast errors in the prediction equation). Specifically, temporal forecasting for the target subcomponent α_0 is denoted as f_T instead of TS_0 to differentiate from other subcomponents.
- *Step 3:* Spatial autocorrelation is an important factor in spatial forecasting. Applying Moran' I index to analyze the spatial differentiation of all non-target subcomponents α_i ($i = 1, \dots, n$). Based upon the spatial correlation of all non-target subcomponents α_i ($i = 1, \dots, n$), a modified Elman recurrent neural network with a backpropagation through time algorithm (BPTT) is built to capture the spatial influence of all non-target subcomponents over the target subcomponent. The network first gets trained and adjusted accordingly. Then forecasts from each time series model TS_i ($i \neq 0$) are fed into the network. The spatial forecasts at α_0 , identified as f_S , can be finally obtained from the network output.
- *Step 4:* In order to keep the combination easily understood and as simple as possible, linear regression combination of forecasting method is applied to the final spatio-temporal forecast. A linear regression is employed to merge the time series model f_T and the Elman neural network model f_S . The linear regression is suggested as follow

$$f_{overall} = x_1 \times f_T + x_2 \times f_S + \text{Regression_Const}\tan t \quad (3)$$

where both the regression coefficients, include x_1 and x_2 , and regression constant, t , have to be estimated beforehand.

The novelty in our approach lies at Step 3, i.e. a modified Elman recurrent neural network is applied (please refer to Section 4.2 for detailed implementation), which overcomes the shortcomings in STIFF.

3.4 Evaluation Criteria

In order to validate the forecasting performance of different approaches, NMSE (Normalised Mean Square Error) is used to evaluate the forecasting accuracy. The NMSE is an estimator of the overall deviations between predicted and actual values. Given N pairs of the actual values (or targets y_t) and predicted values (\hat{y}_t), the NMSE which normalizes the MSE by dividing it through the variance of respective series can be defined as [22]

$$NMSE = \frac{\sum_{t=1}^N (y_t - \hat{y}_t)^2}{\sum_{t=1}^N (y_t - \bar{y}_t)^2} = \frac{1}{\sigma^2} \cdot \frac{1}{N} \cdot \sum_{t=1}^N (y_t - \hat{y}_t)^2 \quad (4)$$

where σ^2 is the estimated variance of the data and \bar{y} is the mean.

If a model has a very low NMSE, then it is well performing both in space and time. On the other hand, high NMSE values mean that the model is not well.

4 Case Study – Forest Fire Area Prediction

The experiment is based upon the data kindly provided by Canada Forest Service – CFS [23]. Canada Large Fire Database records large forest fires of which areas exceed two hundred hectare from year 1959 to year 2000 and covers every province, region and forest park in Canada. Figure 3 shows the spatial relationship of the provinces for the study area and province Alberta (AB) is chosen as the target location for forecasting, i.e. the target subcomponent. The neighbouring (spatially correlated) provinces, i.e. the non-target subcomponents, are British Columbia (BC), Saskatchewan

wan (SK), Manitoba (MB), Ontario (ON), Northwest Territories (NWT) and Quebec (QC). We take monthly data from January 1959 to December 1989 as in-sample (training periods) data sets (502 observations including 48 samples for validation). We also take the data from January 1990 to December 2000 as out-of-sample (testing period) data sets (132 observations) which is used to evaluate the performance of prediction based on evaluation criteria proposed in Section 3, 4.



Fig. 3. Spatial distributions of provinces in Canada

In the experiment, time series analysis for the temporal forecasting is implemented via SPSS software package, ARIMA (auto-regression integrated moving average), which is produced by Apache software Foundation. The Elman network is built using the Matlab software package, which is produced by Mathworks Laboratory Corporation. Firstly, the time series mode for temporal forecasting is built for capturing temporal characteristic after data transformation, then parameters of ARIMA model will be estimated through examining the autocorrelation and partial autocorrelation function (ACF and PACF) plots. After the estimation, constructed ARIMA model will be used to temporal forecasting, finally the values of temporal forecasting individually will be fed to Elman RNN for spatial forecasting. The structure of the Elman RNN is designed based upon the spatial neighboring relationship. For example, according to Figure 3, connectivity matrix based on area adjacency is shown in Figure 4.

	BC	SK	MB	ON	QC	NWT
BC	1					1
SK		1	1			
MB		1	1	1		
ON			1	1	1	
QC				1	1	
NWT	1					1

Fig. 4. Connectivity matrix based on area adjacency for forest fire forecasting

According to Figure 4, the network would be in 6- X -1 structure. That is, there are 6 input neurons, an unknown number (X) of neurons in the hidden layer, and one neuron in the output layer. To determine X, we vary the number of neurons in the hidden layer from the most sparse 3 to the highly dense 12 in order to find an optimal one. It turns out 6, 7 and 8 almost have the same best performance during the training stage. As a result 6 is picked up for its most simple structure. Thus, according to the Figure 4, the condensed network with 6 neurons in the hidden layer is chosen as the one used to find the spatial forecasting. The network structure is shown in Figure 5.

So far there are two forecasted values associated with target location. One carries the temporal forecasting from time series model f_T , and the other bears the spatial one from Elman neural network f_S . Our current goal is to generate an optimal overall spatial-temporal forecasting. The linear regression was applied to combine the spatial forecasting f_S and the temporal forecasting f_T . Due to the limit of the length of the paper, the computational process are omitted but can be obtained from the authors if required.

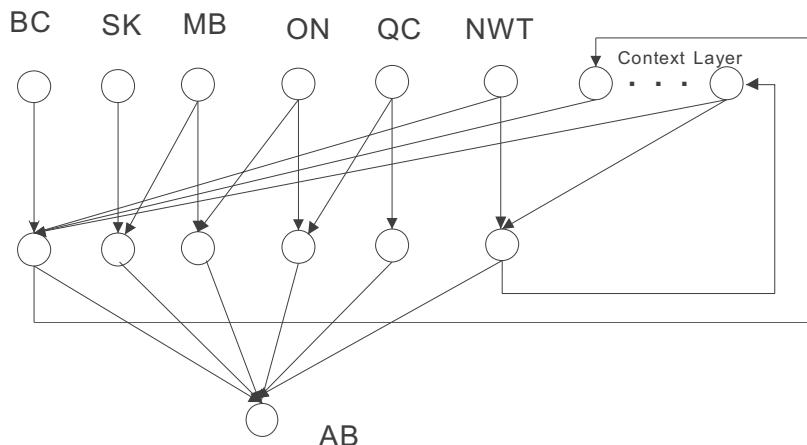


Fig. 5. The topological structure of the Elman network for forest fire prediction

The forecasting results by different models are represented in Figure 6, and Tables 1 shows their accuracies. From the graph and the table, we can generally see that ISTIFF method achieved better forecasting accuracy than STIFF, which is better than ARIMA. It also implies that Elman network in ISTIFF obtains better spatial forecasting than BP network in STIFF.

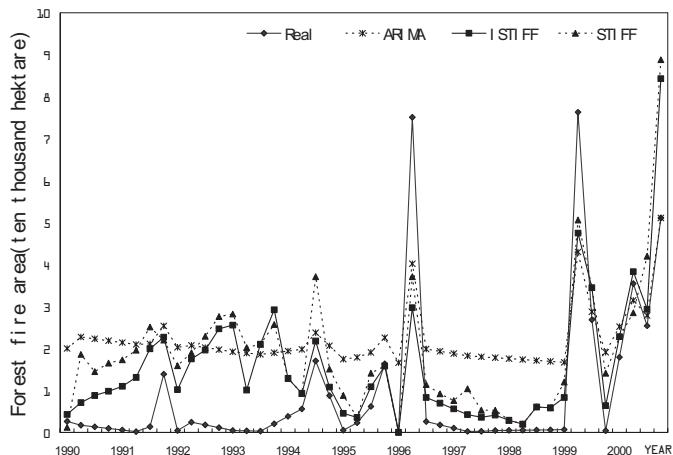


Fig. 6. Predictions of forest fire area by different forecasting methods

Table 1: A comparison of *NMSE* between different methods for *forest fire area* prediction

Method	Data			
	Train data		Test data	
	NMSE	Rank	NMSE	Rank
ARIMA	0.3716	3	0.5969	3
STIFF	0.2453	2	0.4151	2
ISTIFF	0.1872	1	0.2112	1

5 Conclusions

This paper presents an integrated spatio-temporal integrated forecasting framework – ISTIFF by using a dynamic recurrent neural network – Elman Network. The high accuracy of ISTIFF has been illustrated by the cases study of forest fire area predication in Canada and temperature prediction of meteorological stations in China. In future, ISTIFF should be extended to accommodate the nonlinear combination of time series analysis and spatial prediction and multiple spatio-temporal variables.

Acknowledgements

The research is supported by the Major State Basic Research Development Program of China (973 Program, no. 2006CB701306) and the Ministry of Education of China (985 Project, No. 105203200400006).

References

1. H.D. Margaret. “Data Mining: Introductory and Advanced Topics,” Prentice Hall, 2003.
2. R. Drossu, and Z. Obradovic. “Rapid Design of Neural Networks for Time Series Prediction,” IEEE Computational Science & Engineering, Vol. 3(2), 78-89, 1996.
3. P. Tenti. “Forecasting foreign exchange rates using recurrent neural networks,” Applied Artificial Intelligence, Vol. 10(6), 567-81, 1996.
4. S. Walczak. “Artificial neural network medical decision support tool: predicting transfusion requirements of ER patients,” IEEE Transactions on Information Technology in Biomedicine, Vol. 9(3), 468-474, 2005.

5. Z. Li, and M.H. Dunham. "STIFF: A forecasting framework for spatio-temporal data," *Mining Multimedia and Complex Data*, Berlin:Springer-Verlag, 183-198, 2002.
6. Ciucă, and E. Jitaru. "On the recurrent neural network induction by evolutionary computations with applications in forecasting," at http://www.ici.ro/ici/revista/sic1998_2/art08.html, 1998.
7. J.R. McDonnell, and D. Waagen. "Evolving Recurrent Perceptrons for Time-Series Modeling," *IEEE Transactions on Neural Networks*, 5(1), 24-38, 1994.
8. Adam, L. Zarader, and M. Milgram. "Identification and prediction of non-linear models with recurrent neural network," *Laboratoire de Robotique de Paris*, 1994.
9. J. Connor, and L. Atlas. "Recurrent neural networks and time series prediction," In *Proceedings of the International Joint Conference on Neural Networks*, Seattle, 301-306, 1993.
10. A.M. Logar, E.M Corwin, and W.J.B. Oldham. "A comparison of recurrent neural network learning algorithms," In *Proceedings of the International Joint Conference on Neural Networks*, Piscataway, 1129-1134, 1993.
11. K. Kamijo, and T. Tanigawa. "Stock price pattern recognition: A recurrent neural network approach," In *Proceedings of the International Joint Conference on Neural Networks*, Kosaka, 1215-1221, 1990.
12. J. Roman, and A. Jameel. "Backpropagation and recurrent neural networks in financial analysis of multiple stock market returns," *Proceedings of the Twenty-Ninth Hawaii International Conference on System Sciences*, Vol. 2, 454-460, 1996.
13. Parløs, K. Chong, and A. Atiya. "Application of the recurrent multilayer perceptron in modeling complex process dynamics," *IEEE Transactions on Neural Networks*, Vol.5 (2), 255-266, 1994.
14. J. Draye, D. Pavsic, and G. Libert. "Dynamic recurrent neural networks: a dynamical analysis," *IEEE Transactions on Systems Man and Cybernetics*, Vol. 26(5), 692-706, 1996.
15. E. Kosmatopoulos, M. Polycarpou, and A. Iannou. "High-order neural network structures for identification of dynamical systems," *IEEE Transactions on Neural Networks*, Vol. 6(2), 422-431, 1995.
16. L. Jin, P. Nikiforuk, and M. Gupta. "Approximation of discrete time state space trajectories using dynamic recurrent networks," *IEEE Transactions on Automatic Control*, Vol. 40(7), 1266-1270, 1995.
17. B. Filho, E. Cabral, and A. Soares. "A new approach to artificial neural networks," *IEEE Transactions on Neural Networks*, Vol. 9(6), 1167-1179, 1988.
18. Elman. "TLEARN - simulator program for neural networks," Center for Research in J-aiwxe, C-008, University of California, San Diego, La Jolla, CA, 92093-0108, 1990.
19. T. Cheng, and J. Wang. "Applications of spatio-temporal data mining and knowledge for forest fire," *ISPRS Technical Commission VII Mid Term Symposium*, Enschede, 148- 153, 2006.

20. K.S. Narendra, and K. Parthasarathy. "Gradient methods for the optimization of dynamical systems containing neural networks," IEEE Transactions on Neural Networks, Vol. 2(2), 252-262, 1991.
21. P.J. Werbos. "Backpropagation through time, what it does and how do it," In Proceedings of the IEEE, USA, 78(10), 1550-1560, 1974.
22. A.S. Weigend, and N.A. Gershenfeld. "Time series prediction: forecasting the future and understanding the past," In Proceedings of the NATO advanced research workshop on comparative time series analysis, Santa Fe, 1-70, 1994.
23. Canada Forest Services. "CFS," Canada Large Fire Database,
http://fire.cfs.nrcan.gc.ca/research/climate_change/lfdb/lfdb_download_e.htm, 2002.

System of Traffic Control on the Basis of Cartographic Databases and Geoinformation Technologies

A.A.Kravtsov, A.N.Kriuchkov, E.E.Sotikova

United Institute of Informatics Problems National Academy of Sciences of Belarus
Surganov str. 6, Minsk, 220012, Belarus
kravtsov@newman.bas-net.by

Abstract. Software and information tools for cartographic support of transport dispatching service on the basis of electronic regional maps and geographic information systems (GIS) are examined.

Keywords: Electronic map, geoinformation systems, navigation systems, terrain analysis.

Nowadays the integration of contemporary communication and navigation systems for various applications is a widespread tendency. Examples of such an integration are the systems of dispatching control of vehicles location and state.

Development of software and technologies of geographic information systems, digital cartographic data bases, widespread implementation of global position systems like GPS NAVSTAR and GLONASS gave the possibility to create full-scale traffic navigation systems, including on-board analysis, navigation situation visualization, and dispatcher stations for traffic control.

Traffic navigation systems (TNS) are being actively developed to give the drivers and control services the possibilities: to chose, analyze and show on the monitor the geographic, navigation and special information of concrete situation, to define automatically the traffic location, to chart and plan the routes, to create navigation databases, to control the vehicles routes changes, to calculate the route period and petrol consumption, to get supplemental information about traffic road condition, gas stations, stops,

hotels and traffic police points. The information could be presented in graphic and text mode.

The realization of routing problems can be considered in compliance with various limiting factors: statistic (road characteristics and road constructions, path profile), dynamic (traffic activity on given route). Also various types of route optimizations can be used (the sequence of stops control, petrol consumption optimizations).

The United Institute of Informatics Problems (National Academy of Sciences of Belarus) in the frame of the Russian-Belarusian Union State program “COSMOS-BR” has developed the program-information facilities for dispatcher stations of traffic navigation systems (TNS).

Software of dispatcher stations consist of following components:

- the package for visualization of electronic maps, navigation situation, organizing the requests about navigation situation,
- the data bases of electronic terrain maps (ETM),
- the package for navigation information processing.

The software structure chart of dispatcher stations is on the figure 1.

The hardware and some software modules for navigation information processing and control are developed by the Special Design Enterprise “Kamerton” (Minsk) in the frame of the Russian-Belarusian Union State program “COSMOS-BR”.

The software control module realizes the interfaces of the program of navigation information processing and the program for information interchanges with GSM-terminal.

The software module of information interchange with GSM-terminal is designed for providing the connection with controllable vehicle and alert signal waiting. When the command from the module of control arrives the configuration of inquired objects is defined, the priority of inquiries is fixed (to be transmitted via radio and digital GSM channels) and the inquiry interval (for SMS channel) is determined.

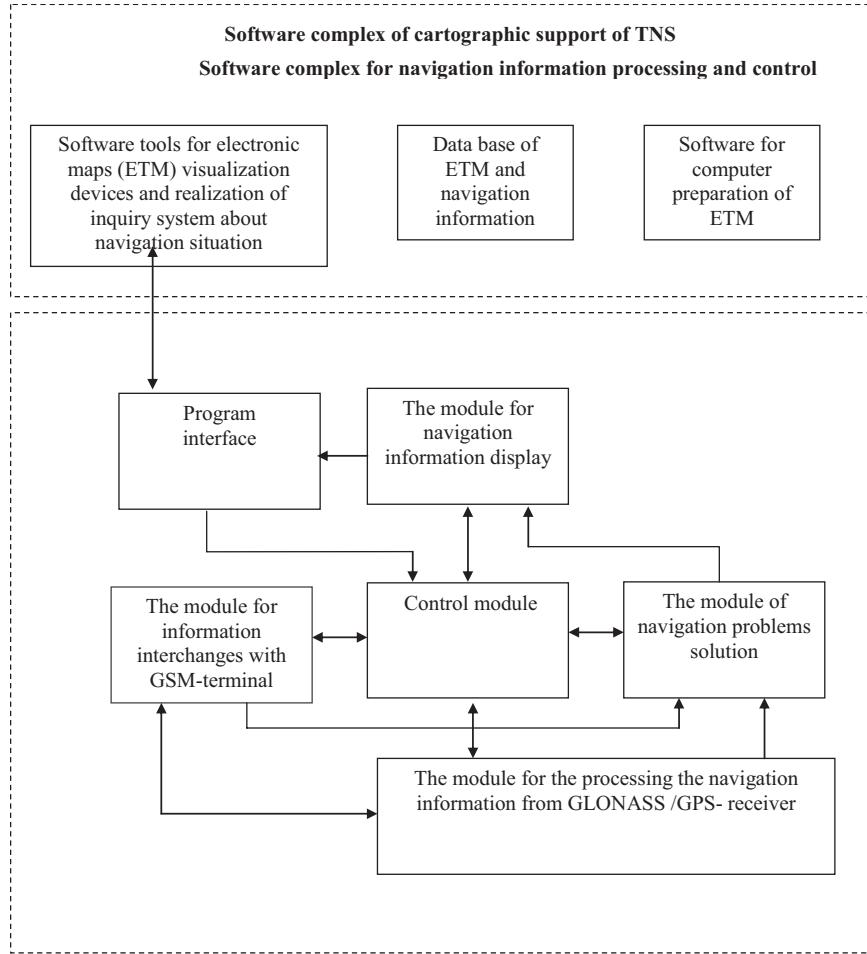


Fig. 1. The structure chart of dispatcher stations software

The priority index (P) determines the period between certain vehicle inquiries depending the circle. It can vary in the range from 0 till 99. The inquiry cycle is the time to inquire all controllable vehicles:

Inquiry cycle:

$$(IC) = T_i + n \times T_r,$$

where T_i – inquiry time;

T_r – response time;

n – number of the objects under control.

The priority index determines how many inquiry cycles would be missed between each inquiry of this vehicle. For example, if the priority is 0, the inquiry will be in every inquiry cycle; if the priority is 1 - every second inquiry cycle. In this case:

$$IP = P + 1,$$

where IP – inquiry period accounted in inquiry cycles.

The controllable objects configuration is the set of individual numbers of the vehicle under control with the priority index for every vehicle.

The information from GLONASS /GPS- receiver and the module of exchange with GSM-terminal is coming through the module of navigation information processing to the module of navigation tasks solution. The last one calculates the coordinates, speed and course of controllable vehicles, makes if necessary the correction of the coordinates to find the errors with use of differential amendments, coming from the module of navigation information processing. A routine of work (with coordinates correction or not) is assigned in the control module.

Calculated navigation information comes to the navigation information display module.

Navigation information with re-counted coordinates comes to the program interface which serves to match the data and to provide the data exchange between the program of navigation information processing and GMS-terminal control, GSM-receiver and the program tools of ETM and navigation situation visualization. The data comes through the program of visualization toward the control module where analysis and processing are realized.

Program-information package for cartographic support of vehicle control is designed to reflect in on-line mode the navigation situation on the basis of electronic terrain maps and also to process the set of inquiries on the displayable objects.

The program-information facilities consist of:

- the tools for digital maps data base creation for the terrains of navigation system activity (creation of digital position zone over chosen terrain, extraction of the objects from the model, special objects mapping, etc);

- the tools for cartographic and navigation information input (matching, displaying, editing, etc);
- the tools for receiving the supplemental information of route and vehicle conditions;
- the tools for planning and registration of the routes and vehicle depot;
- the tools for the monitoring the chosen vehicle provided that the vehicle is always on the screen regardless its location on the route;
- the tools for the control of information quality of vehicle (information absence during long period, route deflection for more than fixed distance);
- the tools for the registration of information accepted from the vehicle to accomplish the analysis.

The possibility of recorded information “playback” in real, slow and speeded up modes.

The tools for receiving the supplemental information about the route condition and route planning are based on cartographic functions of surface analysis:

- altitude determination on the terrain relief or relief altitude matrix;
- building of relief profile along the selected object;
- route length calculation taking into account the relief;
- the relief attitude matrixes on arbitrary region;
- vehicle location reflection on the relief profile;
- slope angle calculating.

All the functions were realized on the basis of vector model of ETM within one sheet of topographic map of any scale or within the bounds of digital position zone, received by the joining of any numbers of maps. ETM is represented in the F20S format with the structure included:

- the file of matrix description of the terrain objects (coordinates, orientation and form);
- the file of semantic description of the objects (qualitative and quantitative characteristics, e.g. the road covering);
- the file of supplementary information (service information of the map or terrain, special objects information, etc).

The functions of analysis of the surfaces are connected with the system of visualization and provide the information about the status of the relief on the route in graphic mode.

The function of creation of relief altitudes matrix (RAM) is realized as separate module and is used as preliminary facility to make the data base of the maps in the process of creation of additional file of relief altitudes in ETM model.

An important characteristic of navigation situation control is the functions of creation, maintenance and edition of the vehicle routes and special conventional signs data base. These functions with the use of functions of visualization and interpretation of topographic and special information can provide the displaying the navigation situation on the ETM. These program tools include:

- the tools for planning and creation of the vehicle routes data base (input, matching, displaying, editing);
- the tools for monitoring and displaying on the electronic map the special navigation working load.

The software includes the following modules:

- the database route-information input module;
- mapping on ETM special navigation working load;
- editing the route information (input, deleting, amendments of the objects);
- creation of graphic image of the conventional signs of special working load.

All modules are created as a separate classes with the set of functions and are included to the system of visualization of cartographic and navigation data and provide the information of the route status on the way of the vehicles in graphic mode.

For special customer needs interpretation and visualization the system includes the following functions:

- functions of road route calculating taking into account region relief;
- profile road route building and reflection;
- reflection of moving vehicle on map route or profile;
- creation, reflection and correction of special working load layer;
- creation, reflection and correction of vehicle routes database.

All the functions are used in navigation data interpretation and are included to the query system. The system realizes the navigation data interpretation by messages exchange with the program complex of navigation information processing and control.

Conclusion

The system of traffic control lets realize all controllable vehicle monitoring. It's assigned to raise the traffic safety and effectiveness of vehicles use, to estimate vehicle drivers actions automatically and also to provide the telematics services for different categories of users.

References

1. Harisov V.N., Perov A.I., Boldin V.A. The global satellite radio navigation system GLONASS. – M., 1999. - 560 p.
2. Kriuchkov A.N. Sotikova E.E. Tarasov S.N. Cartographic supply for transport navigation systems. First Belarussian Space congress; October 28-30, 2003. Congress proceedings. – Minsk: UIIP NASB. 2003. – p.259-261.
3. USNO NAVSTAR Global Positioning System
<http://tycho.usno.navy.mil/gpsinfo.html> (11.08.03).
4. Fugawi UK DigitalMaps
http://fugawi.com/web/products/fugawi_uk_digital_maps_v2 (11.03.05).
5. NaviMap - <http://www.kurshin.orc.ru> (11.03.05).

Information Fusion of Land Laser Scanning for Geographic Information Systems

Ilya S. Tarasov and Nikita A.Pikhtin

A.F.Ioffe Physico-Technical Institute Russian Academy of Sciences,
26, Polytekhnicheskaya, 194021, St.Petersburg, Russia, nike@hpld.ioffe.rssi.ru

Abstract. There is a wide variety of data sources for geographic information systems which acquisition technology is currently improved. Presently earth remote sensing and land laser scanning techniques are the most perspective ones. They are rapidly developed and offer strong resources for acquisition and analysis of spatial information with shortened outgoing in comparison with traditional techniques. Using of high power laser diodes in such systems open new opportunities in improvement of the whole system efficiency. This becomes possible due to the great improvement of output characteristics of diode lasers which will be described in the present work.

Keywords: Geographic information systems, land laser scanning, earth remote sensing, information integration, information presentation of spatial environment, diode lasers, asymmetric separate confinement quantum well heterostructures.

1 Introduction

Geographic information systems (GIS) have found wide application as program software widely available for users. Also GIS is a subject of serious attention from scientists of both natural sciences and computer specialization.

GIS entity is determined by combination of technical, program and information resources providing not only acquisition, storage, processing, access, representation and spreading, but also mathematic-cartographical modeling and integral data representation for solving problems of territorial planning and control. Therefore GIS combine technical resources, software, established regulations and rules of acquisition, storage, analysis and transfer of information about processes and phenomenon having spatial binding and propagation. As a result they are developed at the turn of many branches of science and used in different fields of management activity.

The environment in which a GIS operates (machines, people, networks) is called a "spatial information system", and is designed and created to respond to the strategic spatial information needs of people or organizations.

Special direction in the development of engineering and methods of spatial data acquisition is the use of modern laser equipment. The use of laser technology for earth remote sensing (land scanning) is connected with several technical and information problems. Integration of earth remote sensing data and existing GIS cartographic information is the most specific aspect of given scientific-technological direction.

2 Diode Lasers for Earth Remote Sensing Application

Presently laser locator known as a lidar is more and more often used for earth remote sensing technique. Utilization of lidar in topography means the appearance of principally new laser location method of land scanning which has several advantages in comparison with topographic mapping using aerophotographic camera. Most of all it is high efficiency and accuracy of topography-geodetic survey.

Laser diodes emitting in near infrared wavelength region are used as the light source of probe radiation. The main function of laser diode is the generation of pulse or continuous wave (CW) radiation which reflecting from the earth surface or objects measures the distance between the light source and the object which has caused the reflection. Due to that fact an optoelectronic part of the lidar sometimes is named range-finder. Thus diode laser is of the main optoelectronic component of the lidar.

In recent decade an enormous progress in the improvement of output characteristics of semiconductor lasers has been made. Above all it is an increase of output optical power and wallplug efficiency, an improvement of beam divergence and reliability of diode lasers. All mentioned characteristics are the main issues in the applications of diode lasers particularly as light source in lidars. Below main output characteristics of high power diode lasers are presented more in detail.

Strained quantum well (QW) separate confinement heterostructure (SCH) [1] is considered already classical for high power laser diode development. Utilization of QW SCH (emitted at $1.03\text{ }\mu\text{m}$ wavelength) with $0.4\text{ }\mu\text{m}$ waveguide thickness allowed to reach 1 cm^{-1} internal optical loss, 9 W continuous wave (CW) room temperature (RT) output power and 65% wallplug efficiency [2]. Minimization of internal optical loss in such structure is one of a major condition in the development of high power semiconductor laser [3].

A minimal internal optical loss as low as 0.34 cm^{-1} are reached in asymmetric SCH with QW displaced from the centre of waveguide with more than $1 \mu\text{m}$ thickness. An achievement of such ultra low internal loss in given laser heterostructure design permits to enlarge laser diode cavity length without notable drop of external quantum efficiency. It gives an opportunity to use the higher injection current and therefore to enhance the emitting output power without losing the wallplug efficiency. Such type of heterostructures designed for laser diodes emitting around $1 \mu\text{m}$ wavelength are grown by low-pressure metalorganic chemical vapor deposition (MOCVD) method. Laser heterostructures consist of the following layers: highly doped N-Al_{0.3}Ga_{0.7}As ($N=1018 \text{ cm}^{-3}$) and P- Al_{0.3}Ga_{0.7}As ($P=3,5 \text{ } 1018 \text{ cm}^{-3}$) cladding layers (Si was used as a donor impurity, Mg as an acceptor); an undoped $1,7 \mu\text{m}$ - thick GaAs waveguide and an active region formed by one strained 90\AA In_{0.24}Ga_{0.76}As QW displaced from the waveguide centre on $0,2 \mu\text{m}$. Schematic energy band diagram of such laser heterostructure is presented in Fig.1.

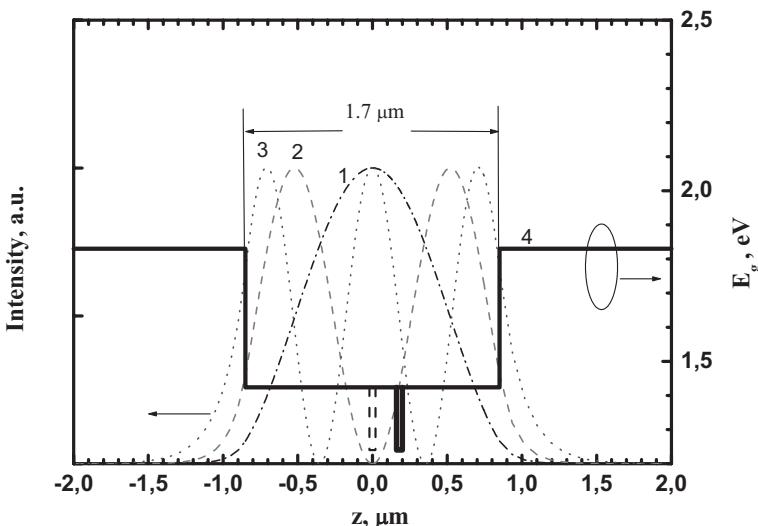


Fig.1. Schematic energy band diagram of QW SCH InGaAs/GaAs/AlGaAs and transverse mode distribution. 1 - fundamental mode, 2 – first mode, 3 – second mode. Z is an axis of heterostructure growth.

The grown wafers are processed into ridge construction lasers with meastripe width $W=100 \mu\text{m}$ formed by dry and wet chemical etching using the conventional post-growth process [4]. Cleaved laser diode chips with

cavity lengths $L = 1 \div 5$ mm are mounted junction down on copper heatsink using indium solder. Anti-reflective/high-reflective (AR/HR=5%/95%) coatings were deposited on the front/rear facets of laser diodes.

RT CW light-current characteristic of coated laser diode with 3 mm cavity length is shown in Fig.2. It was the typical dependence for laser diodes with cavity length 2 \div 4 mm. CW maximum output power reached record-high value of 16W at constant heatsink temperature and was limited by thermal roll-over. No catastrophic optical mirror damage was observed. In studied laser diode threshold current density was 90 A/cm². Slope efficiency $\sim 1,06$ W/A remained stable up to 8W output power. A maximum value of the wallplug efficiency reached 72% at 4A drive current (Fig.2). Thermal and series resistances were 4 K/W and 24 m Ω , respectively. 2 mm long devices exhibited 12W maximum optical output power with 1,1 W/A slope efficiency and record-high 74% wallplug efficiency.

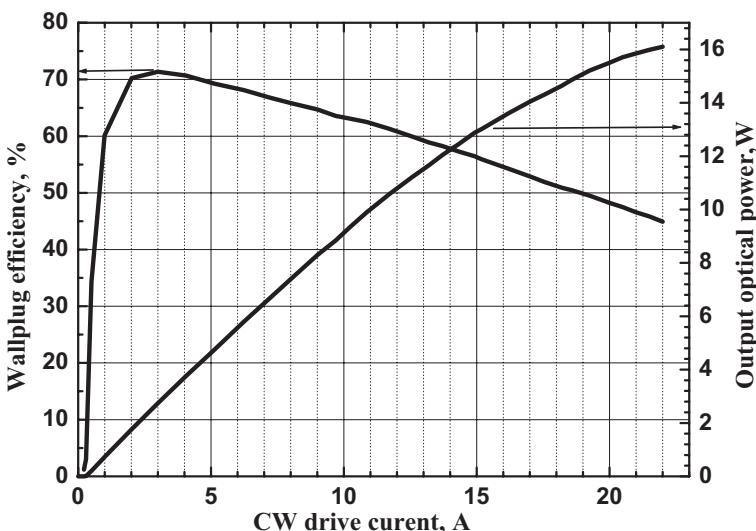


Fig. 2. CW RT light-current and wallplug efficiency characteristics of 100 μ m aperture AR/HR coated laser diodes based on asymmetric SCH InGaAs/GaAs/AlGaAs with waveguide thickness 1,7 μ m.

In pulse operation (pulse duration 100 ns, repetition rate 1 kHz) output optical power in laser diodes described above reaches record value of 145 W. RT pulse light-current characteristic is presented in Fig. 3.

Emitting spectra at pulse operation is presented in Fig.4. It had around 1060 nm peak wavelength and 3 nm and 10 nm full width at half maximum (FWHM) at 20A and 100A pulse drive currents, respectively. Far-field pattern in the planes parallel and perpendicular to p-n junction is presented in Fig.5. The transverse angle value $\Theta_{\perp} = 30^{\circ}$ measured at FWHM nearly doesn't change with drive current. Such stable behavior of beam divergence indicates the fundamental transverse mode character of radiation.

Reliability tests of 2 ÷ 3 mm cavity coated laser diodes were performed at 3 ÷ 4 W operating output powers at 65 °C heatsink temperature. Stable operation was observed for 1000h. The decrease of operating output optical power was within 3%.

Application of land laser scanning information in geographic information systems offers an opportunity to increase efficient functioning of different systems, including systems related to decision making, responsible for safe development of production processes (Fig.6.).

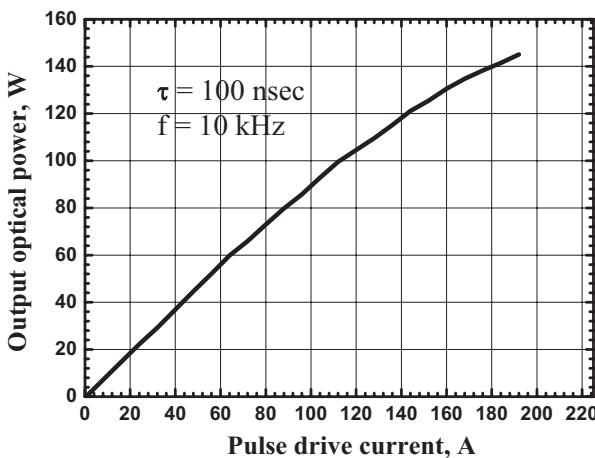


Fig. 3. Pulse RT light-current characteristic of 100 μm aperture AR/HR coated laser diode based on asymmetric QW SCH InGaAs/GaAs/AlGaAs with waveguide thickness 1.7 μm .

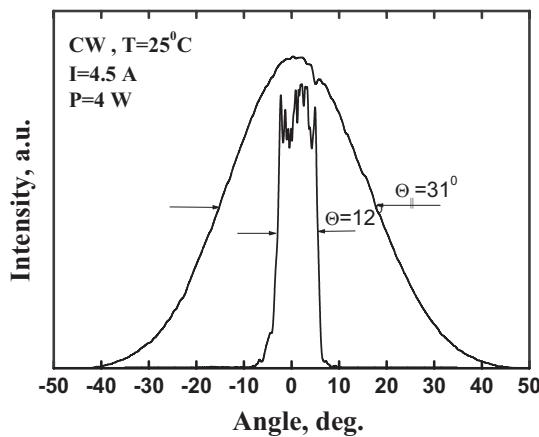


Fig.4. Beam divergence in both planes of high power diode laser with 100 μm aperture based on asymmetric QW SCH InGaAs/GaAs/AlGaAs with waveguide thickness 1.7 μm .

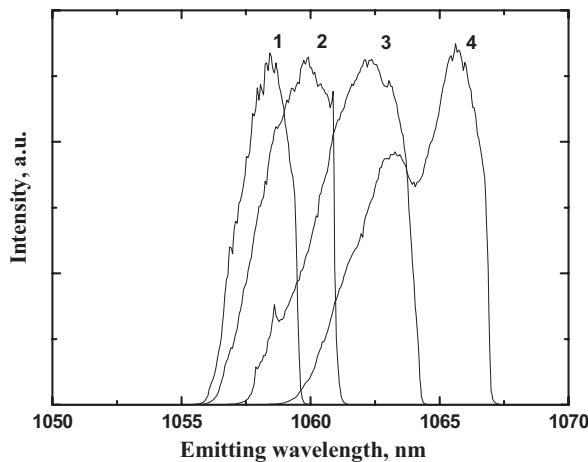


Fig.5. Emitting spectra of laser diode based on asymmetric QW SCH In-GaAs/GaAs/AlGaAs with waveguide thickness 1.7 μm at different pulse (100 nsec, 10 kHz) drive currents

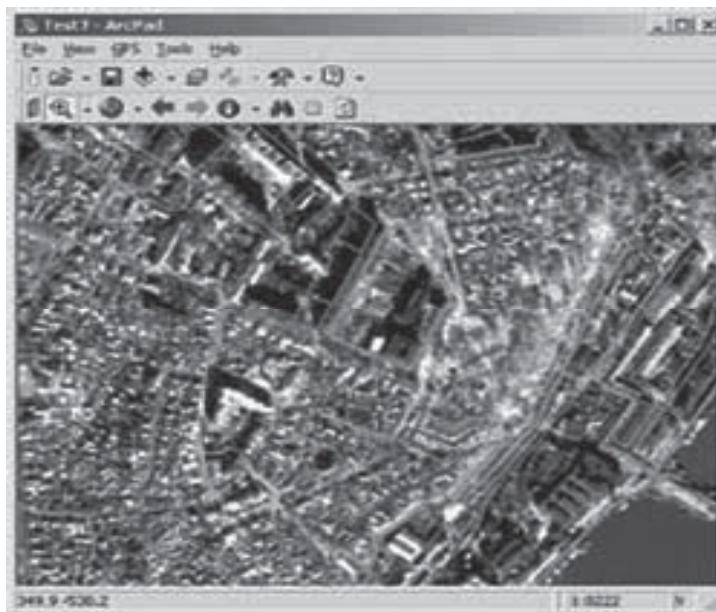


Fig 6. Example of information fusion of land laser scanning for geographic information systems

3 Conclusion

Application of modern laser technique based on high power laser diodes for acquisition of spatial data gives a possibility to increase the efficiency of geoinformation systems and make cheaper the process of geodata actualization. Great improvement of output characteristics of laser diodes, mainly output optical power in CW operation (16W from 100 μm aperture device) and pulse operation (145 W), as well as record high value of wall-plug efficiency which reaches 74%, allow to consider application of high power high efficiency high brightness diode lasers for earth remote sensing technique.

The approach to information fusion of land laser scanning methods and tools application in geographic information systems offered in this article envisages opportunities for different users of parametrization and increase of nomenclature of scanning methods, sequence change of location analy-

sis on the electronic map and allows speaking about its universalism and wide practical applicability.

References

1. E. G. Golikova, V. A. Kureshov, A. Yu. Leshko, A. V. Lyutetski, N. A. Pikhtin, Yu. A. Ryaboshtan, G. V. Skrynnikov, I. S. Tarasov, and Zh. I. Alferov, "Properties of wide - mesastripe InGaAsP/InP lasers", Semiconductors, vol. 34(7), pp. 853 – 856, 2000.
2. D. A. Livshits, I. V. Kochnev, V. M. Lantratov, N. N. Ledentsov, T. A. Na-lyot, I. S. Tarasov, and Zh. I. Alferov, "Improved catastrophic optical mirror damage level in InGaAs/AlGaAs laser diodes," Electron. Lett., vol. 36 (22), pp. 1848-1849, 2000.
3. N.A. Pikhtin, S.O. Slipchenko, Z.N. Sokolova, A.L. Stankevich, D.A. Vinokurov, I.S. Tarasov and Zh.I. Alferov "16 W continuous-wave output power from 100- μ m-aperture laser with quantum well asymmetric heterostructure", Electronics Letters vol. 40, no. 22, 28th October, 2004, p. 1413-1414
4. A. Yu. Leshko, A. V. Lyutetsky, N. A. Pikhtin, S.O. Slipchenko, Z. N. Sokolova, N. V. Fetisova, E. G. Golikova, Yu. A. Ryaboshtan, and I. S. Tarasov, "High power single-mode ($\lambda = 1.3\text{--}1.6 \mu\text{m}$) laser diodes based on quantum well InGaAsP/InP heterostructures," Semiconductors. v.36, pp. 1308-1314, 2002.

Using GIS to Analyze Acute Myocardial Infarction in Turkey

Mehmet Zeki Coskun¹, H. Can Ünen¹, Cevat Kırma², Ercument Yılmaz³

¹ Istanbul Technical University, Civil Engineering Faculty, Department of Geodesy and Photogrammetry Engineering,

34469 / Maslak / Istanbul / Turkey, coskun@itu.edu.tr, unen@itu.edu.tr

² Koşuyolu Training and Research Hospital of Heart, Division of Cardiology, Kartal / Istanbul / Turkey, ckirma@hotmail.com

³ Istanbul University, Medical Science Faculty, Division of Cardiology, Çapa / İstanbul / Turkey, ercuyilmaz@superonline.com

Abstract. Approximately 3.5 million people have cardiologic problems and the number increases by 100 thousand every year in Turkey. Annually, approximately 400 thousand myocardial infarctions (MI) occur and fifty percent of deaths occur due to myocardial infarctions. These deaths occur due to lack of coordination and unconsciousness. It is well known that eighty percent of these incidents can be prevented if they can reach the hospital or health care providers on time. This study is part of a proposal project submitted to TUBITAK (The Scientific And Technical Council of Turkey) and BAP-ITU (Science and Research Projects - Technical University of Istanbul) in order to develop an emergency management model on Myocardial Infarction and, analyze concentration of MIs with respect to location, proximity to health care providers, public transportation routes, and arrange new locations for ambulances. This project will also help the determination of locations of new hospitals or ambulances and/or restoration centers.

Key words: GIS, Emergency, Management, Myocardial Infarction.

1 Introduction

Acute myocardial infarction (AMI or MI), commonly known as a heart attack, is a disease that occurs when the blood supply to a part of the heart is interrupted. The resulting oxygen shortage causes damage and potential death of heart tissue. It is a medical emergency, and the leading cause of death for both men and women all over the world [1]. Important risk factors are history of vascular diseases such as atherosclerotic coronary heart disease and/or angina, a previous heart attack or stroke, any previous episodes of abnormal heart rhythms or syncope, older age- especially men over 40 and women over 50, smoking, excessive alcohol consumption, the

abuse of certain illicit drugs, high triglyceride levels, elevated creatine kinase or troponin levels in the blood, high LDL ("bad cholesterol") and low HDL ("good cholesterol"), diabetes, high blood pressure, obesity, and chronically high levels of stress in certain persons [1]. Risk factors for atherosclerosis are generally risk factors for myocardial infarction: older age, male gender, cigarette smoking, hypercholesterolemia (more accurately hyperlipoproteinemia, especially high low density lipoprotein and low high density lipoprotein), diabetes (with or without insulin resistance), high blood pressure, obesity [2], [3], [4]. Myocardial infarction is a common presentation of ischemic heart disease. The WHO estimated that in 2002, 12.6% of deaths worldwide were from ischemic heart disease [1]. In the most countries, diseases of the heart are the leading cause of death, causing a higher mortality than cancer [2].

In Turkey, the half of deaths is caused by heart and vascular diseases. People say "100,000 people will die in Istanbul earthquake", but approximately 400,000 people die because of heart diseases, and doctors say about 300,000 of these deaths can be prevented. The majority of deaths occur due to lack of coordination. All of the deaths can't be prevented but, some portion of these incidents could be unlethal [5], [6].

Additionally, based on previous research, some trends and patterns are known about myocardial infarction. For example, it is known that myocardial infarction occurs mostly in the morning, between 7 am to 12 am. But a recent study shows that this result does not indicate the true death density. The research indicates that the peak density of deaths moved to another time zone (from 13 pm to 19 pm) because of changes in lifestyle and differences of demographic structure [7]. Thus, the study points that new studies should be done.

Since there is no reliable statistical research made for Turkey yet, unfortunately, we can not give any information or some charts showing the rates of death caused by heart diseases or attacks.

As we all know, Geographic Information Systems (GIS) are moving from isolated, standalone, monolithic, proprietary systems working in client-server architecture to smaller web-based applications and components offering specific geo-processing functionality and transparently exchanging data among them. Web-GIS allows the development of spatial information solutions that are low-cost, simple to implement, compatible with existing information technology infrastructure, and have the potential of interoperating with other systems and applications in the future [8], [9]. As a result, project team plans to complete the project by Web-GIS or by open systems.

2 Data Storage

The defined and necessary data will be determined and a plan will be prepared for gathering it. The term data in the geographical context refers to both the attribute (listed below) and the location identifier for the attributes (e.g., address, a postcode or a census enumeration district). Some of the attributes are listed below:

- geographical maps of pilot area,
- roads,
- day population,
- night population,
- location of existing ambulances,
- location of hospitals,
- location of restoration center if exists,
- demographic data,
- Data about the patients such as location.,,
- socio-economic levels,
- age of person,
- gender of person,
- time of attack,
- response if exists,
- response type and time,
- type of attack,
- dead or alive,
- death place (home/other, hospitals, emergency room/dead on arrival) if dead

3 Steps of Project

To utilize the above objectives, a feasibility study is required to determine the steps. The steps of the project are defined after this feasibility study. The first step of the project is to define and collect data. Data above, may be detailed diversified and enlarged during the research advances. Some of data such as geographical maps of pilot area, roads, day and night population, location of existing ambulances, location of hospitals are already available.

The second step is to utilize a Geographic Information System and define the analysis steps. At present, a database for the system has been de-

signed for further studies but not completed yet. To understand the spatial distribution of survival rates among people discharged alive after their myocardial infarction, the role of contextual socio-economic factors for understanding that spatial distribution, the proximity relationships between hospitals (or health provide centers) and survival systems will be defined and analyzing tools will be developed.

The third step is finding new and the best possible distributions of locations of ambulances and restoration centers. These centers have first started to settle up in United States of America. This is a coverage problem which finds applications in the location of emergency facilities, such as hospitals or ambulances, where it is desirable that every possible emergency must be covered within a fixed suggested minutes of response time, or, when the objective is to minimize the worst-case response time, to reach the furthest possible point. These problems are referred as location-allocation problems, because two types of decisions are made: where to locate, and how to allocate demand for service to the central facilities. This typical location-allocation problem might involve the selection of location for fire stations. Here GIS can be used to find optimum locations for hospitals and restoration centers that result in better response times to emergencies [10].

Additionally, the problem of finding intensive care places can be solved using web-gis. Hospitals, which can not help the person who suffers heart attack, have to send their patients to the specialist hospitals when a myocardial infarction case is responded. This problem can be solved if a central database is used to spot nearest vacant intensive care units which can be of help to MIs.

Finally, as there is no reliable and permanent database and statistical system in Turkey the last and the final step is to develop a location based database for MIs.

4 Study Area

The project will first start with a pilot area is Istanbul metropolitan city located on 41°01"-N / 28°10"-E at left-bottom and 41°08"-N / 29°51"-E at right-top. Istanbul is chosen for research activities because Istanbul is one of the largest and the most populated cities of Europe and located on both Europe and Asia. The city is really big with an area of 11350 square kilometers, a population over 15 million inhabitants and with struggle with major problems of high-density commuter traffic.

Urban and rural areas are shown in Fig. 1. The yellow and the green (from light to dark) parts of picture show rural areas and the other colors (i.e. white, red, brown,) show the urban areas. Densely populated areas are in both red and white colors.



Fig. 1. Urban and Rural Areas of Istanbul

In Fig. 2, the locations of hospitals are shown with the district boundaries. Although there are some private ambulance services, unfortunately, there is not much information about the ambulances and where they are located. While the project studies advance, those data will be included to the system.

It is well known that there is a relationship between survival and the amount of time passed before response and treatment. After the onset of symptoms, patients must be taken to the nearest hospital in a shortest possible time. After restoration, patients are mostly transferred to hospitals which have intensive care units especially for cardiologic diseases. The importance of location of health care providers is figured out when the response time is taken in to account. As seen in the figures 1 and 2, it is obvious that more hospitals or restoration centers are necessary for a big portion of Istanbul.



Fig. 2. Locations of Hospitals

5 Conclusion

For Turkey there are no (or very few) GIS studies and designed databases which would provide better decisions for health care even though the popular technology of GIS is used widely for management and planning.

This study is the start of a project that will conduct the analyses for health care research using GIS covering firstly the region of the city of Istanbul and finally for whole Republic of Turkey. The study is a proposal project suggesting developing an emergency management model of Myocardial Infarction to show importance of the restoration centers, analyze concentration of MIs with respect to location, proximity to health care providers, public transportation routes, and arrange new locations for ambulances. In summary, the goals of the project are listed below.

- Analyzing geographic patterns of acute myocardial infarction.
- Investigation of the spatial distribution of MI incident rates
- Analyzing the relationship between hospital distance and survival.
- Determining new locations for ambulances.
- Determining the location of new restoration centers

- Establishing a new, reliable, permanent database and statistical system for MIs.
- Helping the hospitals via web-GIS which do not have vacant intensive care rooms.

Acknowledgments: Please see following web pages for further information about participants of the project: <http://www.jfm.itu.edu.tr/>, www.ift.istanbul.edu.tr/, <http://www.kosuyolu.gov.tr>

References

1. World Health Organization. Annex Table 2: Deaths by cause, sex and mortality stratum in WHO regions, estimates for 2002 The world health report (2004).
2. Wilson P.W., D'Agostino R.B., Levy D., Belanger A.M., Silbershatz H., Kannel W.B.: "Prediction of coronary heart disease using risk factor categories". *Circulation*. 1998;97:1837-1847 (1998)
3. Yusuf S., Hawken S., Ounpuu S., Bautista L., Franzosi M.G., Commerford P., Lang CC, Rumboldt Z, Onen CL, Lisheng L, Tanomsup S, Wangai P Jr, Razak F, Sharma A.M., Anand S.S.: INTERHEART Study Investigators. "Obesity and the risk of myocardial infarction in 27,000 participants from 52 countries: a case-control study." *Lancet* (2005)
4. Jensen G, Nyboe J, Appleyard M, Schnohr P. "Risk factors for acute myocardial infarction in Copenhagen, II: Smoking, alcohol intake, physical activity, obesity, oral contraception, diabetes, lipids, and blood pressure." *Eur Heart J*, PMID 2040311 (1991) 298-308.
5. Turkish Heart Foundation : <http://www.tkv.org.tr/index.php> (accessed 2006).
6. Heart Surgery Foundation of Turkey: http://www.kalpsagliji.org/asil_fay_hatti.htm, (accessed 2006).
7. Sarı, İ., Davutoglu, V., Soydiç, Sucu, M., Erer, B., Tekbaş, E., Uyarel, H., Aksoy, M.: Time Distribution of Myocardial Infarction in Turkish society, XXII national Cardiology Conference, 24-26 October 2006, Antalya, Volume 34, Turkey, (2006) 36
8. Anderson G; Moreno-Sanchez R.: Building Web-Based Spatial Information Solutions around Open Specifications and Open Source Software, Transaction In GIS, Volume 7 Issue 4, (2003) 447
9. Zhong-Ren Peng, Ming-hsiang Tsou, Hoboken, N.J. : Wiley, Internet GIS: distributed geographic information services for the internet and wireless networks, Wiley, USA, (2003)
10. Erden T., Coşkun Z.: "Searching The New Locaitons And Determining The Latest Status Of Fire Stations With GIS In Istanbul", Proceedings of 13th International Emergency Management Society Annual Conference, May 23-26, 2006, Seoul, Korea, (2006) 43-51.

Obtaining Semantic Descriptions Based on Conceptual Schemas Embedded into a Geographic Context

Miguel Torres and Serguei Levachkine

Geoprocessing Laboratory – Centre for Computing Research – IPN, Mexico City,
MEXICO
`{mtorres, sergei}@cic.ipn.mx`

Abstract. Information integration and semantic heterogeneity are not trivial tasks. An integrated view must be able to describe various heterogeneous data sources and its interrelation to obtain shared conceptualizations. In this paper, we propose an approach to make a conceptualization of the real world based on conceptual schemas, which are used to generate a semantic description of the geospatial domain. This description attempts to provide the guidelines to formalize the geographic domain in form of geospatial ontologies according to specific contexts. We propose conceptual schemas in order to abstract specific and essential *parts* of the geospatial domain and to represent schematically which geospatial entities should be collected and how they must be organized. Moreover, we perceive that geographic data modeling requires models more specific and capable of capturing the *semantics* of geospatial data, offering higher abstraction mechanisms and implementation independence. Therefore, we approach conceptual schemas to describe the contents of the real world abstraction to specify the behavior of the geospatial entities, in which the context plays an important role to guarantee shared and explicit conceptualizations. Our research is mainly oriented to propose an approach related to conceptual issues concerning what would be required to establish ontologies of the geospatial domain. In addition, the work is led to formalize appropriate methods basing on conceptual schemas to represent ontologies of the geospatial domain.

1 Introduction

Ontology has gained increased attention among researchers in geographic information science in recent years. Up-to-date, the ontology notion plays an important role in establishing robust theoretical foundations for geographic information science [1]. Under this umbrella, it is possible to unify several interrelated research subfields, each of which deals with different perspectives on geospatial ontologies and their roles in geographic information science. Three broad sets of foundational issues need to be resolved: (1) conceptual issues concerning what would be required to establish an exhaustive ontology of the geospatial domain, (2) representational

and logical issues relating to the choice of appropriate methods for formalizing ontologies, and (3) issues of implementation regarding the ways in which ontology ought to influence the design of information systems.

Nowadays, there are diverse institutions that use geospatial data to make a decision in different environments. The use of geographic databases through geographic information systems (GIS) provides tools for managing, analyzing and processing geospatial data. However, information can not be sometimes represented in “adequate” way, since it contains ambiguities that do not allow the appropriate use and analysis. These ambiguities are originated by imprecision of information, heterogeneity and isolation sources. Whereby, it is difficult to develop interoperable applications that allow us to share, integrate and represent geospatial information.

These facts bear with searching solutions oriented to geospatial data representation and integration, semantic heterogeneity and imprecise geographic objects issues. Consequently, commercial GISs do not have tools to extensibly explore the essential properties and relations of geographic objects. Therefore, it is difficult to explore the *semantics* of a set of geographic objects.

According to [2] and [3], the ontologies and the knowledge representation are essential for the creation and the use of standards to exchange data, as well as for the design of human computer interaction. Whereby, an ontology allows us to solve problems associated to heterogeneity, interoperability, representation, integration and exchange of geospatial data. These problems imply incompatibility between diverse geographic objects, as well as a different spatial conceptualization according to a specific context. For example, we engage with the world day by day in a variety of different ways: we use maps, specialized languages, and scientific instruments; we also engage in making rituals and telling stories; we use information systems, databases, different machines and other software-driven devices of various types. Each of these ways of engaging with the world involves a certain *conceptualization*. It involves a system of concepts and categories, which divide up the corresponding universe of discourse into objects, processes and relations in different sorts of ways. Thus, in a religious ritual setting, we might use concepts such as *God*, *salvation* and *sin*; in a scientific setting, we might use concepts such as *micron*, *force* and *nitrous oxide*; in a story-telling we might use concepts such as *magic spell*, *dungeon* and *witch*. These conceptualizations are often tacit, that is, they are often invisible components of our cognitive apparatus, which are not specified or thematized in any systematic way [4].

On the other hand, the conceptualization of geospatial domain is diverse, because the geospatial data are often imprecise or many subjects have different points of view. Thereby, it is important to consider alterna-

tive object representations, which are independent of the imprecise nature of the geospatial data [5].

Our research is mainly oriented to propose an approach related to conceptual issues concerning what would be required to establish ontologies of the geospatial domain, according to the point (1) described in [1].

In this paper, we propose an approach to make a conceptualization of the real world based on conceptual schemas, which are used to generate a semantic description of the geospatial domain. This description can provide the guidelines to formalize this domain in form of geospatial ontologies according to specific contexts.

The rest of the paper is organized as follows. Section 2 describes related works with these issues. Section 3 sketches out our approach to conceptualize the geospatial domain. Section 4 depicts a case study to build a semantic description based on conceptual schemas. Our conclusions and future works are outlined in Section 5.

2 Related Works

Some works related to ontologies and semantics in geospatial information science to be mentioned are as follows. Guarino [6] coined the term “ontology-driven information systems” and provided a broad discussion on their place in the computer and information science. Gruber, one of the pioneers of the use of ontological methods in information science, defines an ontology as “a specification of a conceptualization” in [7].

Smith *et al.* [4] reported the results of a series of experiments designed to establish how non-expert subjects conceptualize geospatial phenomena. Subjects were asked to give examples of geographical categories in response to a series of differently phrased elicitations. The results yield an ontology of geographical categories – a catalogue of the prime geospatial concepts and categories shared in common by human subjects independently of their exposure to scientific geography.

Bishr *et al.* [8] argued that information modeling requires to be controlled to allow successful sharing of information. Also, they suggest that any coherent information model need to be based on accepted ontological foundation to guarantee unambiguous interpretation. In addition, their work attempts to show that ontology based information modeling provides more cognitive foundation for information systems models and therefore minimizes the problem of semantic heterogeneity.

Smith *et al.* [2] designed an ontology of geographic kinds to yield a better understanding of the structure of the geographic world, and to support

the development of GIS that are conceptually sound. This work first demonstrated that geographical objects and kinds are not only larger versions of the everyday objects and kinds previously studied in cognitive science.

Fonseca *et al.* [9] proposed a framework to link the formal representation of semantics (i.e., ontologies) to conceptual schemas describing information stored in databases. The main result is a formal framework that explains the mapping between a spatial ontology and a geographic conceptual schema. The mapping of ontologies to conceptual schemas is made using three different levels of abstraction: formal, domain and application levels. At the formal level, highly abstract concepts are used to express the schema and the ontologies. At the domain level, the schema is regarded as an instance of a generic data model. At the application level, authors focus on the particular case of geographic applications. Additionally, they discuss the influence of ontologies in both the traditional and the geographic systems methodologies, with an emphasis on the conceptual design phase.

According to this works, it is important to distinguish that our research is concentrated to use conceptual schemas to describe the semantic contents of the real world abstraction to specify the behavior of geospatial entities, in which the context plays an important role to guarantee shared and explicit conceptualizations.

3 Geospatial Domain Conceptualization

This section gives the guidelines to build conceptual schemas in order to conceptualize the geospatial domain. Thus conceptual schemas are used to generate a semantic description, which can provide the framework to formalize the geospatial domain, according to specific contexts. In this section, we point out the main terms involved in our approach such as conceptual schema and context.

3.1 Design of Conceptual schemas for Geospatial Domain

In the modeling approach, the modeler is required to capture a user's view of the real world in a formal conceptual model. Such an approach forces the modeler to mentally map concepts acquired from the real world to instances of abstractions available in his paradigm choice. On the other hand, the consolidation of concepts and knowledge represented by a conceptual schema can be useful in the initial steps of ontology construction. To adequately represent the geographic world, we must have computer representations capable not only of capturing descriptive attributes about

its concepts, but also of describing the relations and properties of these concepts.

We propose conceptual schemas to describe the contents of the real world abstraction in order to specify the behavior of the geospatial entities. In this case, conceptual schemas certainly correspond to a level of knowledge formalization. Conceptual schemas are built to abstract specific parts of the geospatial domain and to represent schematically which geographic entities should be collected and how they must be organized. We perceive that geographic data modeling requires models more specific and capable of capturing the *semantics* of geospatial data, offering higher abstraction mechanisms and implementation independence.

The proposed conceptual schemas are composed of two types of concepts (C): *terminal* (C_T) and *non-terminal* (C_N). The first ones are concepts that do not use other concepts to define their meaning (they are defined by “simple values”). The meaning of non-terminal concepts is conceived by other concepts, which can be terminal or non-terminal concepts (see Eqn. 1).

$$C = C_N \bigcup C_T \quad (1)$$

Each concept has a set of *aspects*. They are characteristics that describe the properties, relations and instances that involve the geospatial objects. From-now-on, we shall use the term “relation” to denote unary relations/properties as Berendt *et al.* [10]. From this point of view, all aspects of a terminal concept are simple, e.g. the type of all aspects that belong to the set of primitive types (punctual, linear and areal objects) is denoted by (T_P), as shown in Eqn. 2.

$$\begin{aligned} T_P &= \{number, character, string, enumeration, struct\}, \\ A &= \{a_i \mid type(a_i) \in T_P\}, \end{aligned} \quad (2)$$

where T_P is the set of primitive types; A is the set of aspects.

Then, the set of *terminal concepts* is defined by Eqn. 3.

$$C_T = \{c(a_1, a_2, \dots, a_n) \mid a_i \in A, i = 1, \dots, n\} \quad (3)$$

The *non-terminal concepts* have at least one aspect that does not belong to T_P . It is denoted by Eqn. 4.

$$C_N = \{c(a_1, a_2, \dots, a_n) \exists \exists a_i \notin A\}, \text{ where } c \text{ is a concept.} \quad (4)$$

Finally, the set of relations R is defined by the pairs that are associated to Γ and Φ , in which Γ and Φ are non-reflexive, non-symmetric, and transitive relations (Eqn. 5).

$$R = R_\Gamma \cup R_\Phi = \{(a, b) | a\Gamma b, a \in C_N, b \in C\} \cup \{(a, b) | a\Phi b, a \in C_N, b \in C\} \quad (5)$$

According to these definitions, it is necessary to express the semantics that can provide a conceptual schema by means of a description D . Therefore, we consider the concepts C embedded into the conceptual schemas through geospatial objects, which are represented by primitive types as well as the set of relations R involved among geospatial objects (see Eqn. 6)

$$D = \langle C, R \rangle \quad (6)$$

Fig. 1 depicts a conceptual schema, which has been designed for the geospatial domain. Thus, this schema is adaptive for any context. In other words, it attempts to reflect the main features involved in this domain. For instance, if we have topographic, geologic, or tourism contexts, it is possible to describe the entities, characteristics and relations embedded between geographic objects, as an inheritance mechanism. The main features involved into geospatial domain have been abstracted of the real world in order to obtain a conceptualization. This conceptualization provides us explicit vocabulary that represents the ontological commitment of the cognitive and intuitive perception of the subjects.

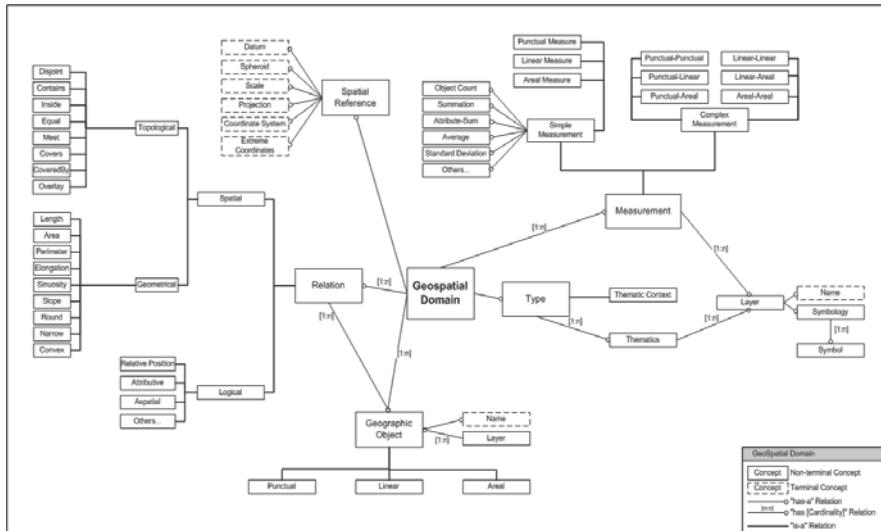


Fig. 1. Conceptual schema proposed to conceptualize the geospatial domain

The conceptual schema shown in Fig. 1 conceptualizes the geospatial domain. This schema represents a raw conceptualization, which contains an optimal number of relations. We are looking for a compact conceptual schema-based abstraction that drives the cognitive process of phenomenon semantic description under specific context. We consider that conceptual schema depicted in Fig. 1, could have more concepts involved in geospatial domain, thereby it can be a certain approximation about the main entities that compose the domain. This could be considered as the first step to collect and organize the concepts contained into the geographic context.

3.2 Context-driven Approach to Restrict the Domain into Conceptual Schemas

The context term is defined as “that which surrounds, and gives meaning to something else”¹ or it is the “discourse that surrounds a language unit and helps to determine its interpretation”². However, to obtain shared conceptualizations and to accomplish with ontological commitments, it is necessary to take into account the context term. Also, it can be used to consistently map different conceptualizations. Due to this, the meaning of a

¹ According to the Free On-line Dictionary of Computing.

² Definition provided by WordNet.

spatial concept may be dependent on a large number of contexts within which the concept is used.

Contexts about a particular use of a spatial concept refer to the knowledge that human uses to constrain the meaning of communication. To reach a common understanding of a vague concept, e.g. *near*, the system and the user require sharing knowledge about the relevant contexts that affect the understanding of the vague concept. We focus on three features: *task*, *spatial contexts*, and *background of the user*. We perceive that context is a key issue in interaction between human and computer, describing the surrounding facts that add meaning.

Context is very useful in geographic information science. For instance, when a user requests geographic information (map) to be displayed by a GIS, the user is trying to perform a *domain task* that has some information needs. The task becomes an important part of the use *context* for *spatial concepts*. Suppose the same request “*show me a map near Cancun*” may be made by a *subject-A*, who is in a task situation of selecting clothing store, and by a *subject-B* who is planning vacation. However, *subject-B* is likely to expect a map showing a larger geographic area comparing with *subject-A*. There are evidences that the meaning of spatial concepts, such as “*near*”, is also dependent on the spatial context. Therefore, the relevant spatial context of an object depends on the purpose of the considered geo-spatial data.

We assume that the context term can be used as a mean to express *exceptions*³ or *constraints*⁴. This use of context is particularly adapted to a rule-based representation of geospatial knowledge, in which exceptions to the rules contain *context-related* terms. Thus, we present a set of intuitive ideas and preliminary definitions that aim better understanding the roles that play the context into the conceptualization based on conceptual schema.

- **Context.** Let a set (of terminal and non-terminal concepts) X , which contains a set of subcontexts Y and $X \subseteq Y$. Then, the set of subcontexts composes the universe of the context denoted by $Y \subseteq C_G$, in which C_G is called *geographic context*. X should be a large set (“large” with respect to cardinality $|C_G|$). Thus,

- A concept C , which can be terminal (C_T) or non-terminal (C_N) concept, belonging to subcontext Y should mentally suggest or bring into our attention Y .

³ Example: “remove all buildings except the isolated one”.

⁴ Example: “the river must be into a valley”.

- $C \in Y$ implies that the name (mention, evocation⁵) of C makes us thinking about Y . In the real world, Y occurs, appears, is produced, is achieved, happens, is used} whenever C {occurs, appears, is produced, is achieved, happens, is used}. For example, concept river \in context H \ddot{O} ROLOGY. H \ddot{O} ROLOGY is a set, but we wrote here just its name, since it is a named set.
 - Context should be obvious, not hidden. It should be evoked by every C belonging to it.
 - Context is the extension of *concept* to sets (to named sets).
 - A concept may belong to several contexts. For example, river \in H \ddot{O} ROLOGY; river \in WATER FLOW. A concept (belonging to a context) could be a context, too. For example, MEXICAN H \ddot{O} ROLOGY \in H \ddot{O} ROLOGY
 - Contexts can overlap and merge.
- *Problem or Objective (P)*. It contains initial state and ending state, in other words, the study object (O_i), a result object (O_r) and a set of constraints (K) that involve the problem or objective (see Eqn. 7).

$$P_m = \{O_i, O_f, K_m\}, \quad K \rightarrow K_m \quad (7)$$

Therefore, we should take into account the context of the problem to obtain a shared conceptualization about the phenomenon of the real world. Then, the context (Ψ) can be denoted by the problems that are defined by itself (see Eqn. 8).

$$\Psi = \bigcup_m P_m \quad (8)$$

We perceive that semantics is always defined by a specific context and it is given by a collection of geospatial entities that involve the context.

4 Case Study

In this section, we describe two scenarios, which are focused on showing how to conceptualize the geospatial domain, by means of conceptual schemas in order to obtain a semantic description regarding specific con-

⁵ Thinking, depicting, imagining.

text. The goal is to depict how these scenarios converge in the same semantic description (see Fig. 4). Although their representations are different, they belong to the same context; thereby their semantic description is the same as well as their conceptualization. Therefore, these kind of conceptualizations can be represented in a conceptual schema and restricted by a context

- **Scenario 1: Imagining the real world.** Suppose that we are looking at a landscape, which depicts several entities such as a forest that has a lake and a river. Moreover, the freeway F25 crosses the highway I37, F25 is used to arrive to Santa Cruz, which is the main town of the surroundings (see Fig. 2)⁶. So, it is important to make a conceptualization about our observations. In other words, we are making and abstraction process that is used to conceptualize the landscape.

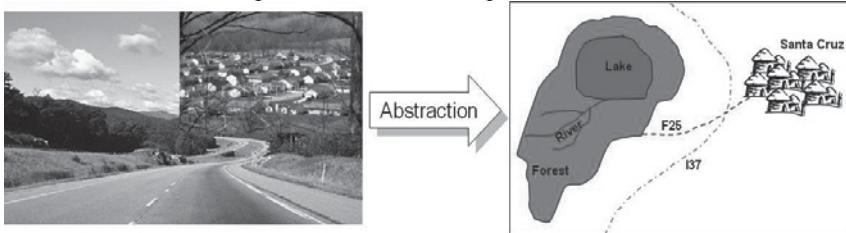


Fig. 2. Scenario 1: Imagining and representing the real world

- **Scenario 2: Vector map.** Suppose that we are looking at a map (Fig. 3), it depicts different thematics that consist of different layers, in which each layer contains geographic objects represented by geospatial primitives. The map has *Populations* (POP), *Hydrologic Features* (HYF), *Roads* (ROD) and *Soils* (SOL). Additionally, each thematic and its layers are denoted by a legend. The map is composed of 2 areal objects, 3 linear objects and 1 punctual object.

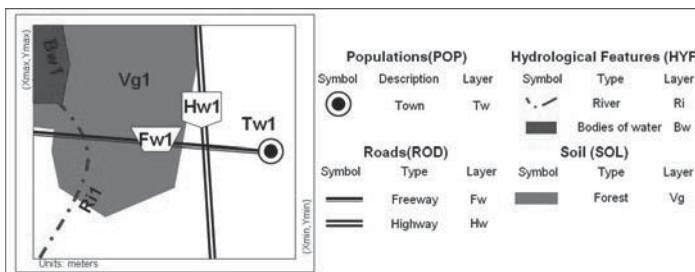


Fig. 3. Thematic map used to illustrate the second scenario

⁶ Fig. 2 is only an approximation or an idea of a general landscape described above, we only select some objects to show an illustrative example.

Thereinafter, we use the conceptual schema shown in Fig. 1 to describe both scenarios. According to Fig. 1, in order to obtain a semantic description from conceptual schema, it is necessary to map the geospatial entities into the conceptual schema. Once concepts have been defined into the conceptual schema, we choose the non-terminal concept to be described (this means to select the aspect to be pointed out). The process continues until we find a terminal concept. When the terminal concept is found, it is necessary to select a pair of geospatial objects, verifying if a relationship between them exists, then a part of description is generated. Terminal concepts are defined by the kind of relation between two objects. In other words, the description starts at the non-terminal concept called “**Geospatial Domain**”. The non-terminal concepts are denoted by means of *rectangles* and the values of the terminal concepts are represented by *ellipses*.

According to the aspect of each non-terminal node, we establish a relation that defines another non-terminal or terminal concept (depending on the objective). This leads to complete the description of geographic objects that compose both scenarios. They converge in the same description according to the context; even though these are represented in different ways (see Fig. 4).

The method is focused on describing the *semantic content* based on conceptual schemas in a geographic context. However, the description depends on a number of relations, properties and measurements⁷ that are considered, whereby it is possible to increase the semantic resolution in the description. The description is made using *tuples* of non-terminal and terminal concepts related among themselves (they are denoted by *Concept relation Concept*). For instance, Fig. 2 and 3 are composed of several spatial objects. The objects in the layer reflect the relation “**is-a**” (i.e., *HW is-a Linear Object*). Moreover, the topological relation “**Intersect**” is related to *Hw1* and *Fw1*, in which both are linear objects. Thus, in description the “**Intersect**” relation is generalized as a spatial relation too.

⁷ A measurement is a procedure for computing values, which are the basis to evaluate characteristics of geospatial phenomena.

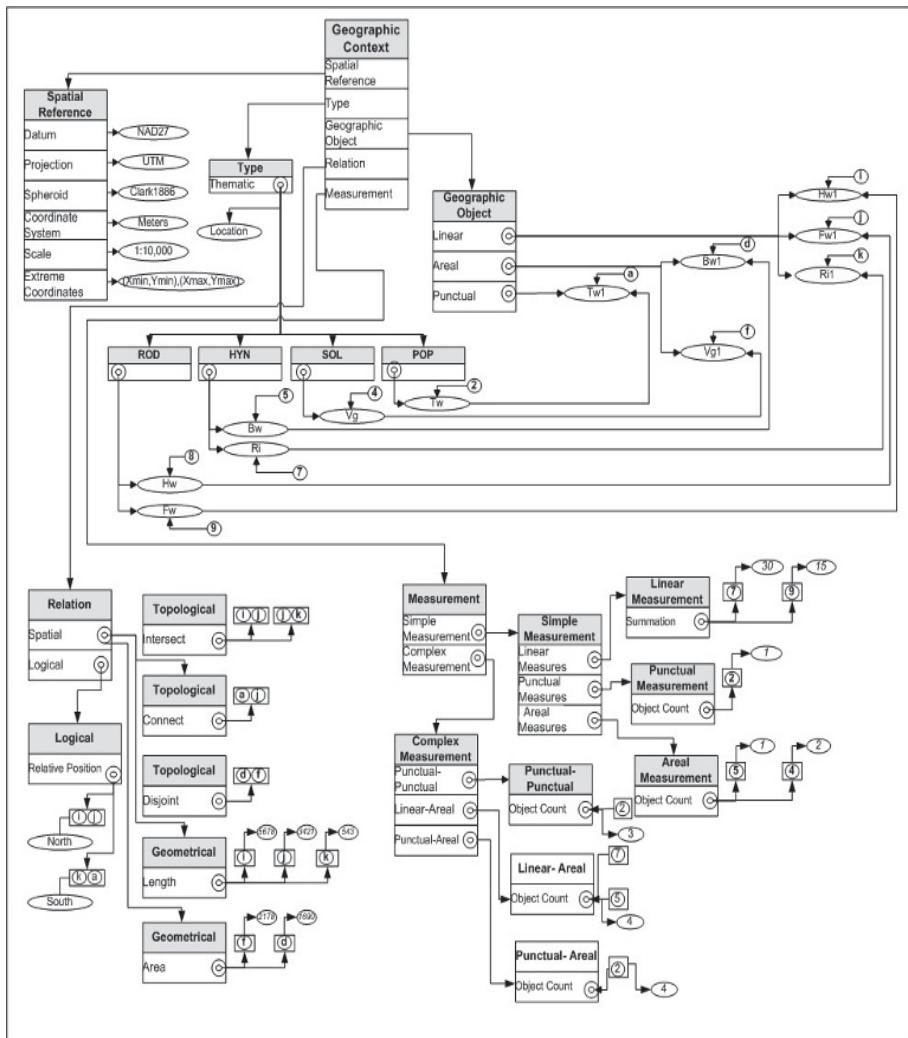


Fig. 4. Semantic description of the scenarios 1 and 2

5 Conclusion and Future Works

We propose an approach to make a conceptualization of the real world based on conceptual schemas, which are used to generate a semantic description of the geospatial domain. This description attempts to provide the guidelines to formalize the geographic domain in form of geospatial ontologies according to specific contexts.

On the other hand, we perceive that geographic data modeling requires models more specific and capable of capturing the semantics of geospatial data, offering higher abstraction mechanisms and implementation independence.

This approach allows us to process imprecise data and aid to information integration and semantic heterogeneity tasks. Thus, the method is focused on describing the semantic content based on conceptual schemas embedded into geographic context. We have introduced two types of concepts: “terminal” and “non-terminal”, as well as two kinds of relations: “*has*” and “*is-a*” to build the conceptual schema. Additionally, we have described a set of intuitive definitions oriented to conceptualize the geospatial domain, referring to conceptual schemas and context.

Therefore, we approach conceptual schemas to describe the contents of the real world abstraction to specify the behavior of the geospatial entities, in which context plays an important role to guarantee shared and explicit conceptualizations.

In addition, several scenarios can converge in the same *semantic description*, although any representation could be more reach than other. This fact essentially depends on the *cognitive* sense of each subject.

Future works are mainly oriented to propose conceptual issues related to translate semantic descriptions into geospatial ontologies, as well as what would be required to establish these kinds of ontologies. In addition, our work is led to formalize appropriate methods to represent ontologies of the geospatial domain and to measure semantic contents between geospatial ontologies.

Acknowledgments

The authors of this paper wish to thank the CIC, SIP, IPN and CONACYT for their support.

References

1. Mark, D., Smith, B., Egenhofer, M. and Hirtle, S.: Ontological Foundations for Geographic Information Science, in McMaster, R. and Usery, L. (Eds.) *A Research Agenda for Geographic Information Science*, CRC Press, Boca Raton, FL (2004) 335-350.
2. Smith, B. and Mark, D.: Ontology and Geographic Kinds. *Proceedings of the 8th International Symposium on Spatial Data Handling*, Vancouver, Canada (1998) 308-320.

3. Minsky, M.: *A Framework for Representing Knowledge*, Technical Report, in MIT-AI Laboratory, AIM-306, USA (1974).
4. Smith, B. and Mark, D.M.: Geographical categories: an ontological investigation. *International Journal of Geographic Information Science*. 15(7) (2001) 591-612.
5. Torres, M., Moreno, M., Quintero, R. and Fonseca F.: Ontology-driven description of spatial data for their semantic processing. *Proceedings of the First International Conference on Geospatial Semantics*, Springer-Verlag, 3799, Mexico City, Mexico (2005) 242-249.
6. Guarino, N.: Formal Ontology and Information Systems. *Proceedings of the International Conference on Formal Ontology in Information Systems*, Kluwer Academic Publishers, IOS Press, Trento, Italy (1998) 3-15.
7. Gruber, T.R.: Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human and Computer Studies*. 43(56) (1995) 907-928.
8. Bishr, Y and Kuhn, W.: Ontology-Based Modelling of Geospatial Information. *Proceedings of 3rd AGILE Conference on Geographic Information Science*, Helsinki, Finland (2000) 24-27.
9. Fonseca, F., Davis, C. and Câmara, G.: Bridging Ontologies and Conceptual Schemas in Geographic Information Integration. *GeoInformatica*. 7(4) (2003) 355-378.
10. Berendt, B., Barkowsky, T., Freksa, C. and Kelter, S. I. E.: Spatial Cognition - An Interdisciplinary Approach to Representing and Processing Spatial Knowledge. *Spatial Representation with Aspect Maps*, Springer-Verlag, (1998) 157-175.

Representing the Semantic Content of Topological Relations into Spatial Databases

Miguel Martinez, Marco Moreno, Miguel Torres and Serguei Levachkine

Geoprocessing Laboratory – Centre for Computing Research – IPN, Mexico City,
MEXICO miguelrosales@sagitario.cic.ipn.mx,
{marcomoreno, mtorres, serguei}@cic.ipn.mx

Abstract. An approach focused on incorporating semantic content into Spatial Databases is proposed. Our methodology is based on a conceptualization of a geospatial domain restricted to retrieve the meaning of topological relations between geographic objects by means of concepts. Indeed, in spatial databases only a small set of topological relations is explicitly represented and they are identified when the geospatial data are displayed or analyzed. While, a semantically enriched set of such relations may be required, sometimes this can be identified at the time when the geospatial data are manipulated by the user. Thus, we define six relations, which are obtained considering the behavior of diverse thematics such as Hydrology, Land Use, Transportation Networks, and Settlements. Geospatial objects are analyzed to identify the topological relationships. We consider two analysis levels: intrinsic and extrinsic. The intrinsic level consists of the analysis between geographic objects that belong to the same thematic. The extrinsic level is composed of relations between pairs of geographic objects that belong to different thematics. Therefore, descriptions are automatically generated in form of tuples $\{O_i, R O_j\}$, where O_i and O_j represent a pair of geospatial objects, and R represents the concept (relation). Each tuple represents the meaning of a topological relation. For example, a highway (O_1) crosses (R) a roadway (O_2). We consider that this method adds a partial semantic content to the geographic databases, because the concepts represent the meaning of topological relations. The conceptual representation has some advantages with respect to the traditional approaches: the conceptualization does not depend on the data scale, geo-reference system, dimension, etc. In addition, we propose a native format that has been designed to appropriately represent and analyze the topological relations.

1 Introduction

Nowadays, the spatial databases commercially used have a little or null semantic content, great part of this content is represented implicitly in the data and requires being extracted analyzing geographic data. In general, the geospatial data have different properties that cover diverse aspect of

geographical data, such as topological, geometrical, thematic and logic properties.

Additionally, the topological relationships between geographical objects are not explicitly represented in the spatial databases. Frequently, these relationships are identified when the data are displayed or analyzed [12].

Up to date it is necessary that the GIS lead the efforts on investigation to describe the spatial relationships explicitly, by means of objects conceptualization and the relationships that maintain with other entities in any case study. In addition, to make use of the semantics in order to solve problems that traditionally are dealt with numerical or classic processing, it is necessary to conceptualize a specific content.

When a topological description of a spatial database is made in explicit way, it is based on concepts that represent the topological relationships. These can be explicitly stored in a spatial database. This approach automatically identifies topological relationships analyzing different themes that compose a spatial database.

Several works have been published to analyze spatial relations [15]:

- Intersection models, developed by [4] [5] [6].
- Schemas based in RCC (Region Connection Calculus), developed by [13], [14], [3], [1] and [2].

Considering these models, we have chosen the intersection model, because it defines the topological component for geographic objects, based on a set point theory, which can be used to analyze and formalize the topological relations between spatial objects. Additionally, the topological components are considered to analyze the relations between objects that are represented with different geometrical primitives of representation, instead of RCC model that does not take into account.

Semantic content can be added to a spatial database by means of tuples. These tuples are composed of a concept that represents the topological relationships. In this way, each tuple represents the meaning of a topological database.

The context of this work is focus on the topology among geographic database. In order to represent semantic content in spatial databases, we use a conceptual representation of the topological relationships. Fig. 1 shows the methodology that we propose to integrate semantic content to databases. Fig. 2 shows the conceptual representation of a topological description.

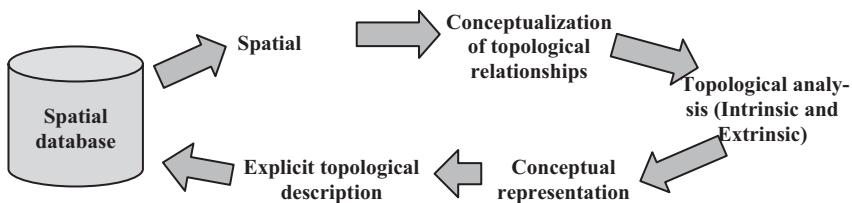


Fig. 1. Methodology to integrate topological semantic content to spatial databases

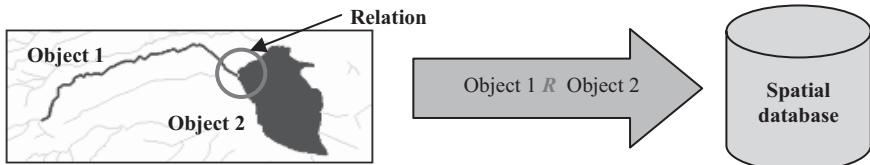


Fig. 2. Conceptual representation of a topological description

This paper is organized as follows: Section 2 sketches out the conceptualization of topological relationships used to describe the relations between two objects. Section 3 describes the levels of topological analysis: intrinsic and extrinsic. Section 4 shows the experimental results. Finally, Section 5 points out our conclusions and future works.

2 Conceptualization of Topological Relationships

The conceptualization is performed by means of the relationships defined in 9-intersection model [6] and INEGI¹ specification [7][8][9]. We have defined a set of 6-relations to describe topological descriptions between two spatial objects. These relations are based on the properties described in the topographic data dictionaries. The objects that were used to compose the spatial database are classified in four themes: Hydrology, Land Use, Communication Network and Settlements (see Table 1).

¹ National Institute of Geography, Statistics and Informatics, National Mapping Agency

Table 1. Set of relations

Relations	Symbol	Description
<i>Connect</i>	C	Valid to relate linear object to another object. The initial or final node of the linear object is connected to a limit [6] of another object. For instance: A river <i>connect</i> to a prey; A road <i>connect</i> to another road; A highway <i>connect</i> to a population.
<i>Share</i>	S	Valid to relate area objects to area or linear objects. Pairs of this objects have common elements, except the boundary. For instance, a river that is part of the boundary of a country.
<i>Share limit</i>	Sl	Valid to relate area objects to another area objects. The only common element is the boundary. For instance, the boundary between two states.
<i>Cross</i>	X	Valid to relate linear objects to areal or linear objects. (7) part of linear object is inside of an area object; (5) two linear objects are intersected, but the flow is not shared. For instance, a highway <i>cross</i> a railroad.
<i>Intersect</i>	Y	Valid to relate pairs of linear object. This relations describe an intersection and their flow is shared. For instance, an street <i>intersect</i> with another street.
<i>Inside</i>	I	Valid to relate any kind of objects, if they are inside of an area object. For instance, a town <i>inside</i> an state.

The description of 6-relations is showed in Table 1, as well as the symbol with which each relation is described.

Connect relation is described in [10]. This relation represents the connection between two objects. There are three variants of this relation: relations between Line/Area, Line/Line and Line/Point object. These possibilities depend on the spatial objects that are related. In other words:

- When the relation is between one line object O_L and one area object O_A , we say that O_L *connect* with O_A ; i.e. if any node of O_L *connect* with the limit of any O_A . Using the topological components [11] to define this relation, we have the following:

$$\partial O_L \cap \partial O_A \neq \emptyset \text{ and } {}^o O_L \cap {}^o O_A = \emptyset \text{ and } {}^\sim O_L \cap {}^\sim O_A \neq \emptyset$$

For example, “The river X *Connect* with the lake Y” (see Fig. 3).

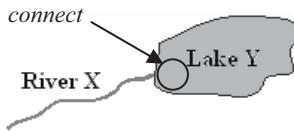


Fig. 3. The river X *connect* with the lake Y

- If the relation between two line objects O_{L1} and O_{L2} , a node of O_{L1} exists inside or it is in the limit of O_{L2} , we say that O_{L1} *connect* with O_{L2} . Using the topological components to define this relation, we have the following:

$$(\partial O_{L1} \cap \partial O_{L2} \neq \emptyset \text{ or } \partial O_{L1} \cap {}^o O_{L2} \neq \emptyset) \text{ and } {}^c O_{L1} \cap {}^c O_{L2} \neq \emptyset$$

For instance, “The river X *connect* with the river Y” (see Fig. 4a) and “The street X *connect* with the street Y” (see Fig. 4b).

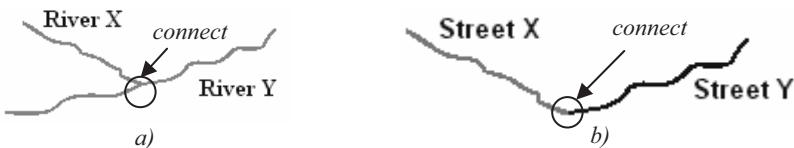


Fig. 4. a) The river X *connect* with the river Y. b) The street X *connect* with the street Y

- When the relation between any line object O_L and any point objects O_P , we say that O_L *Connect* with O_P , if O_P is in the limit of O_L . According with the Egenhofer’s definition, the limit of the point object is itself. Using the topological components to define this relation, we have the following:

$$\partial O_L \cap \partial O_P \neq \emptyset \text{ and } {}^c O_L \cap {}^c O_P \neq \emptyset$$

For example, “The highway X *connect* with the population Y” (see Fig. 5).

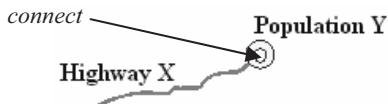


Fig. 5. The highway X *connect* with the population Y

3 Topological Description Based on Intrinsic and Extrinsic Relationships

The topological description consists of two levels. First it is necessary to analyze the intrinsic relations and the second level is used to analyze the extrinsic relations. According to these levels, we generate the explicit topological description.

3.1 Intrinsic Relationships Description

The intrinsic relationships are those that exist between objects that compose a same theme; for example, relations that exists inside Communications Networks theme. To identify these relations, we use a diagram workflow for each theme. Fig. 6 depicts the intrinsic relations of the Communication Network.

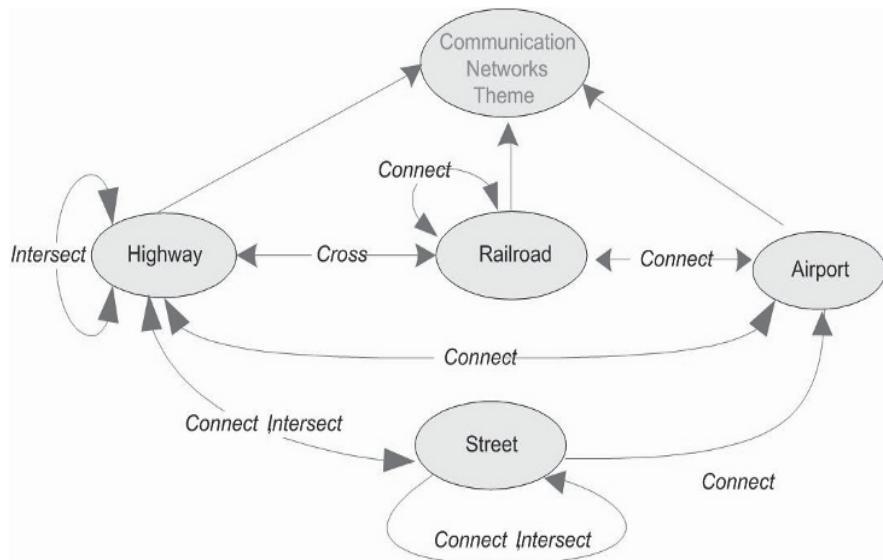


Fig. 6. Intrinsic relations in Communication Networks

3.2 Extrinsic Relationships Description

The extrinsic relationships are those that exist between different thematics, for example, the relations between objects that belong to Hydrology and

Settlement thematics. For instance, in Fig. 7 is shown the relations among Hydrology and Settlement thematics.

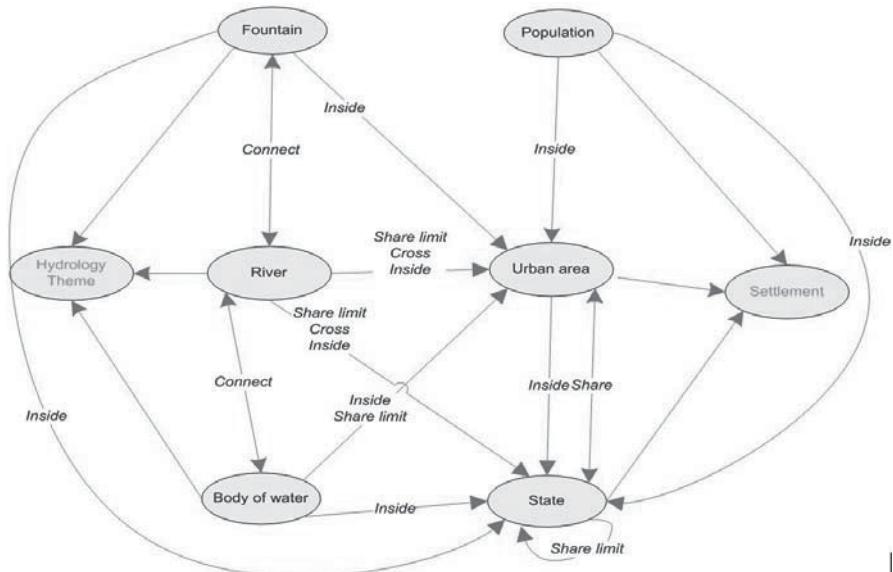


Fig. 7. Extrinsic relations between Hydrology and Settlements

4 Experimental Results

The results that were obtained from the topological descriptor are data stored in a table, in which each table represents the descriptions of the topological relationships. These descriptions are stored in dBASE format (.dbf). Table 2 shows the attributes that compose the table of descriptions.

Table 2. Attributes of description table

Attribute	Description
ID#R	Identifier of relation between two objects.
ID_OBJ_1	Index of the first object. This index corresponds to the index of the attribute table.
LAYER_BELONG_1	Data layer name of the first object.
ID_OBJ_2	Index of the second object. This index corresponds to the index of the attribute table.
LAYER_BELONG_2	Data layer name of the second object.
RELATION	Contain the symbol that identifies the relation that exists between two objects.

To make the topological analysis, we have developed a library of classes that stores and manages the spatial object in a proprietary format. These classes were implemented in Borland C++ Builder. The implementation of the classes is focused on working on vector data.

The functions in C++ to identify the *Inside relation* between polygons are the follows:

```

bool Relations::Inside(Poly *_p1,Poly *_p2) {
    bool flag=false;
    if(_p2->polyInPoly(_p1))
        flag=true;
    return flag;
}

bool Poly::polyInPoly(Poly *_poly) {
    bool flag=true;
    for(int i=0;i<_poly->narcs;i++)
        for(int j=0;j<_poly->arcs[i].npoints;j++)
    {
        if(pointOutPoly(&(_poly->arcs[i].point[j])))
        {
            flag=false;
            break;
        }
    }
    return flag;
}

```

These classes implement methods to obtain the basic (connectivity and adjacency), topological relationships between pairs of spatial objects. In addition, they provide methods for computing geometrical measures such as sinuosity measure of an arc shape.

The topological relationships are obtained using the methods *Connect()*, *Share()*, *Share_Limit()*, *Inside()*, *Cross()*, and *Intersect()*.

Fig. 8 depicts an example in which we show *Inside* relationships between two spatial objects. “Green urban area” and “Urban area” (explicitly represented in the database).

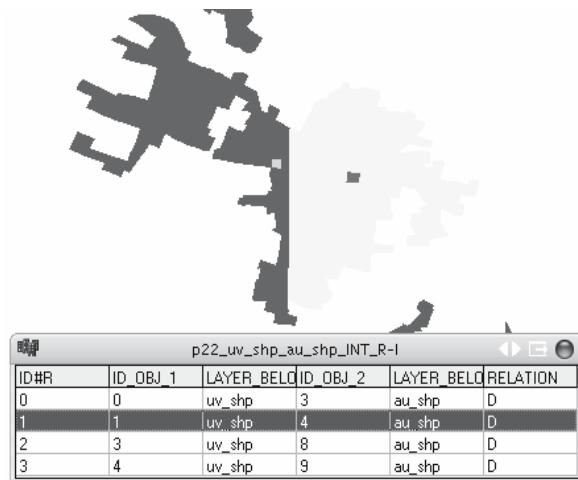


Fig. 8. Green urban area Inside of Urban area

5 Conclusions and Future Works

With this classification in thematic, the analysis of relations in two levels has been carried out. They were analyzed by the existent relations between objects of each theme. This analysis is denominated *analysis of intrinsic relations*. Afterwards, it was analyzed the existent relations between thematics, to which it was called *analysis of extrinsic relations*.

The *content semantic* of relations between data is expressed by *concepts*. The conceptualization of topological relationships and concepts allow us to integrate a partial *semantic* to the GIS applications. The semantic content is obtained relating pairs of objects with one topological relationship.

The concepts are generated using sets of geographic data. The concepts represent the interpretation of spatial data and the meaning of the relations between geospatial objects.

This work, we attempt to catch the semantic content that implicitly contains the spatial data and they do not depend on other factors, like scale or projection.

As a future work, the conceptualizations of these relations can be enhanced incrementing the number of thematics and topological relations, with which can be added new relations between objects. Another aspect could be to include context in this analysis. In future, we will be very interest to analyze the changes in the relations that depends on the context.

It is important to project that a domain conceptualization is useful to build ontologies, which represent (globally) the context of a certain domain, while the vocabulary of concepts and its relations describe the semantics (locally).

This descriptor can be applied to improve the results and performance in spatial analysis process; for instance, in automatic generalization process.

Acknowledgements

The authors of this paper wish to thank the IPN, CIC, SIP and CONACYT for their support.

References

1. Cohn, A., Randell, D. and Cui, Z.: Taxonomies of logically defined qualitative spatial relations, *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Kluwer, 1994.
2. Cohn, A., Bennett, B., Gooday, J. and Gotts, N.: Qualitative spatial representation and reasoning with the region connection calculus, *GeoInformatica*, 1:275-316, 1997.
3. Cui, Z., Cohn, A.G., and Randell, D.A.: Qualitative and Topological Relationships in Spatial Databases, *Third Symposium on Large Spatial Databases*, Lecture Notes in Computer Science Nº 692, pages 296-315, Singapore, June 23-25, 1993.
4. Egenhofer, M. and Franzosa, R.: Point-Set Topological Spatial Relations, *International Journal for Geographical Information Systems*, 5(2): 161-174, 1991.
5. Egenhofer, M.: A Model for Detailed Binary Topological Relations, National Center for Geographical and Analysis and Department of Surveying Engineering, Department of Computer Science, University of Main, Orono, ME 04469-5711, U.S.A., 1993.
6. Egenhofer, M. and Herring, J.: Categorizing topological spatial relationships between point, line and area objects, *The 9-intersections: formalism and its use for natural language spatial predicates*, Technical report 94-1, National and Analysis, Santa Barbara, 1994.
7. Instituto Nacional de Estadística Geografía e Informática, “Diccionario de Datos Topográficos (vectorial), escala 1:250 000”, 1995, (in Spanish).
8. Instituto Nacional de Estadística Geografía e Informática, “Diccionario de Datos Topográficos (vectorial), escala 1:50 000”, 1996, (in Spanish).
9. Instituto Nacional de Estadística Geografía e Informática, “Diccionario de Datos Topográficos (vectorial), escala 1:1000 000”, 1997, (in Spanish).
10. Martínez, M.: Descriptor Topológico para Mapas Topográficos, Master in Science Thesis, Mexico, 2006, (in Spanish), (in Spanish).
11. Molenaar, M.: An Introduction to the Theory of Spatial Object Modelling for GIS, Department of Geo-Informatics, International Institute for Aerospace Survey and Earth Science, Enschede, The Netherlands.
12. Mustière and Moulin, B.: What is spatial context cartographic generalisation?, *Symposium on Geospatial Theory, Processing and Applications*, Symposium sur la théorie, les traitements et les applications des données Géospatiales, Ottawa, 2002.

13. Randell, D. and Cohn, A.: Modelling topological and metrical properties of physical processes, Proceedings First International Conference on the Principles of Knowledge Representation and Reasoning, Morgan Kaufmann, Los Altos, pages 55-56, 1989.
14. Randell, D., Cui, Z. and Cohn, A.: A spatial logic based on regions and connection, Proceedings Third International Conference on Knowledge Representation and Reasoning, Morgan Kaufmann, San Mateo, pages 165-176, 1992.
15. Stell, J.: Part and Complement: Fundamental Concepts in Spatial Relations, Annals of Mathematics and Artificial Intelligence 0: 1-17, 2004.

Some Remarks on Topological Abstraction in Multi Representation Databases¹

Andreas Thomsen and Martin Breunig

Institute for Geoinformatics and Remote Sensing
University of Osnabrück, Germany
{andreas.thomsen,martin.breunig}@uos.de

Abstract. Topology is playing a central role in GIS and data integration. However, the standardisation of topological data models, especially for data represented at different levels of detail, has not yet far progressed. The multi-representation of geo-objects poses new challenges, resulting in the development of Multi-Representation Databases. MRDB manage objects with changing scale and level of detail. As spatial models change their content over time, the users require various representations of geo-objects. A general model based on oriented hierarchical d-Generalized Maps is introduced to represent topology in a MRDB. The model can be used as a data integration platform for 2D, 3D, and 4D topology. The realisation of the approach results in a topological toolbox comprising elementary and complex tools for single and multiple representations. An application example is presented, which uses 2D cartographic datasets from Hannover University. Finally, an outlook to ongoing research is given.

Keywords: Topology, abstraction, geoinformation, data integration, LOD, MRDB.

1 Introduction

Topological data models represent a central component of GIS. However, to our knowledge, the database representation of topology in different spatial dimensions, time, and different levels of detail (*LOD*) has not been investigated in detail. However, a general topological data model can be easily embedded into general models such as ontologies (Guarino, 1998;

¹ This work is funded by the German Research Foundation (DFG) in the project “MAT” within the DFG joint project “Abstraction of Geoinformation”, grant no. BR 2128/6-1.

Berners-Lee et al., 2001) or knowledge representations, e.g. frames and semantic networks used in the AI community (Winston 1992).

As digital spatial models are state of the art in science and engineering, and GIS applications have become commonplace, new questions arise concerning the acquisition, management, and representation of spatial information. Geo-data are represented at LOD related to different scales; subsurface 3D models represent deposits at different LOD from exploration to exploitation. Ideally, a spatial model should yield many different representations varying with time, whereas redundancy should be reduced by automatically deriving many “views” at different LOD, starting from a small set of basic representations, in order to reduce costs and to avoid inconsistencies.

Whereas the application of database technology for GIS is well established, multi-representation poses new challenges, resulting in the development of *Multi Representation Databases (MRDB)*, that manage discretely and continuously changing LOD. Although generalisation operations affect the topology of a spatial model, research about the representation and management of topology in MRDB is still at its beginning.

From a database point of view, DBMS software should not rely on a priori knowledge about the internal structure of the data. In a classical RDBMS, such information must be explicitly represented. As this requirement is too restrictive for spatial data, specific spatial access structures have been introduced, whereas object-relational extensions handle objects with internal structure, and support functionality beyond the possibilities of the relational model. Whereas spatial indices are independent of the internal structure of the managed objects, topological access structures tend to be problem-specific. In 3 dimensions, a multitude of different representations range from simplicial complexes to boundary representations. In this paper, we investigate how a more general concept, namely oriented hierarchical G-Maps, can be used to handle the topology of a digital spatial model at different levels of detail in a multi-representation DBMS, providing a more generic, less application-dependent approach. The method is general enough to support 2- and 3-dimensional models, as well as 2D-manifolds in 3D space.

2 A general Approach to the Abstraction of Topology

Cellular complexes, and in particular cellular partitions of d-dimensional manifolds (d-CPM) are able to represent the topology of an extensive class of spatial objects (Mallet, 2002). Based in algebraic topology, they pro-

vide a more general, less rigid framework than simplicial complexes. By stepwise aggregation of cells, hierarchies of d-CPM model a succession of generalisations (Köthe, 2000). The topology of d-CPM can be represented by d-dimensional Cell-Tuple Structures (Brisson, 1993) respectively d-dimensional Generalized Maps (d-G-Maps) (Lienhardt, 1991). These possess the combinatorial structure of abstract simplicial complexes, each d-cell-tuple comprising $d+1$ cells of different dimension, and related to its neighbours by involution operations. Lévy (1999) has shown that 3D-G-Maps have comparable space and time requirements as the well-known DCEL and radial edge structures, but a much wider range of application, and generate a more concise code. Lévy introduces hierarchical G-Maps (HG-Maps) for the representation of nested structures. Fradin et al. (2002) use G-Maps to model architectural complexes in a hierarchy of multi-partitions. G-Maps have been used to represent the topology of land-use changes (Raza and Kainz, 1999), and are applied in the geoscientific 3D-modelling software GOCAD (Mallet, 1992, 2002).

In this paper, we investigate to what extent oriented G-Maps can be used in 2- and 3-dimensional MRDBs, to support topological operations and queries on all LOD such as

- systematic traversal by ordered circuits (“orbits”)
- boundaries of complex objects
- neighbourhood relationships
- identification of connected components
- topological consistency checks
- passage from one level of detail to another

The data structures shall manage the topology of complex spatial objects in 2 and 3 dimensions. They can be realised as a network of cell-tuples that is made persistent by *object-oriented DBMS (OODBMS)* software on the one hand, and as relations in an *object-relational DBMS (ORDBMS)* on the other hand. We are currently implementing both types of representation in order to examine their respective advantages and disadvantages.

2.1 The Topological Model

2.1.1 Motivation

A well-known discrete approximation of a curved surface is *triangulation* combined with appropriate interpolation, e.g. in a 2.5D digital terrain model that represents the topographical information of a map. The con-

cept of such a surface can be extended to the discrete representation of *2D-manifolds* in 3D space by triangulated surfaces - a concept that can also be applied to situations such as overhangs, which cannot be handled by the simple 2.5D approach.

It seems thus reasonable to choose an approach to the representation of topology that not only supports the generalisation of 2.5D triangulated surfaces, but will permit the extension towards true 3D models. The chosen representation of topology should be general, formal and based on a clear and concise mathematical concept, allowing the development of simple algorithms to support topological queries. Such models are cellular d-complexes represented by cell-tuple structures (Brisson, 1993) or by the closely related Generalized Maps (G-Maps, Lienhardt 1991). Comparing the two approaches one might say that Lienhardt's G-Maps are a more theoretical concept having its foundations in algebraic topology, while Brisson's approach to cell-tuple structures is closer to practical implementation. Lévy (1999) notes that G-Maps cover a somewhat wider range of special non-manifold configurations.

2.1.2 Generalized-Maps and Cell-Tuples

As defined by Lienhardt (1991), a *d-dimensional Generalized Map (d-G-Map)* consists of a finite set of objects called "*darts*" and a set of one-to-one mappings α_i , $i = 0, \dots, d$ linking pairs of darts, that are *involutions*, i.e. that verify $\alpha_i(\alpha_i(x)) = x$, and that further verify the condition

for all i , $0 \leq i < i+2 \leq j \leq d$, $\alpha_i \alpha_j$ is an involution, i. e.
 $\alpha_i(\alpha_j(\alpha_i(\alpha_j(x)))) = x$.

Furthermore, the G-Maps are *embedded* in space by a mapping that associates to each dart a unique combination of a node, an edge, a face, and in 3D a solid in space.

In Brisson's (1993) terminology, *Cell-tuple Structures* consist of a set of *cell-tuples* (*node, edge, face[, solid]*) attached to the corresponding spatial objects. The cell-tuples are pairwise linked by "*switches*" defined by the exchange of exactly one component, which correspond to Lienhardt's involutions:

$$\alpha_0: (\mathbf{n}, \mathbf{e}, \mathbf{f}, \mathbf{s}) \rightarrow (\mathbf{n}', \mathbf{e}, \mathbf{f}, \mathbf{s}), \quad \alpha_1: (\mathbf{n}, \mathbf{e}, \mathbf{f}, \mathbf{s}) \rightarrow (\mathbf{n}, \mathbf{e}', \mathbf{f}, \mathbf{s}), \quad \alpha_2: (\mathbf{n}, \mathbf{e}, \mathbf{f}, \mathbf{s}) \rightarrow (\mathbf{n}, \mathbf{e}, \mathbf{f}', \mathbf{s}), \quad \alpha_3: (\mathbf{n}, \mathbf{e}, \mathbf{f}, \mathbf{s}) \rightarrow (\mathbf{n}, \mathbf{e}, \mathbf{f}, \mathbf{s}')$$

In the following, we use Lienhardt's terminology with one exception: we do not distinguish between the abstract *darts* and their practical implementation as *cell-tuples*, and in consequence systematically use the term *cell-tuple* instead of *dart*.

Cellular complexes can be considered as a generalisation of *simplicial complexes*, but lack the algebraic properties of the latter. The involution operations of the G-Maps provide cellular complexes with the combinatorial structure of an *abstract simplicial complex*, where the cells and cell-tuples play the role of *abstract nodes* and *abstract simplexes*, whereas the involution operators define the *neighbourhood relationships* between the abstract simplexes. Note that the abstract nodes n, e, f, s of a 3-G-Map belong to 4 classes distinguished by different dimensions, whereas all nodes of a simplicial complex belong to the same finite set of vertices in space.

A d-G-Map can be represented as a graph with cell-tuples as nodes, and edges defined by the involution operations (fig. 1). Moreover, it can be stored as a relation in tabular form (fig. 2).

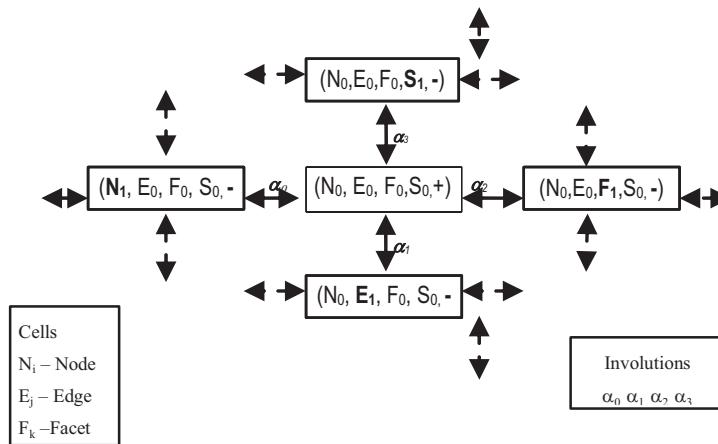


Fig. 1. Representation of an oriented 3-G-Map as a graph with symmetries determined by the combinatorial character of the involutions

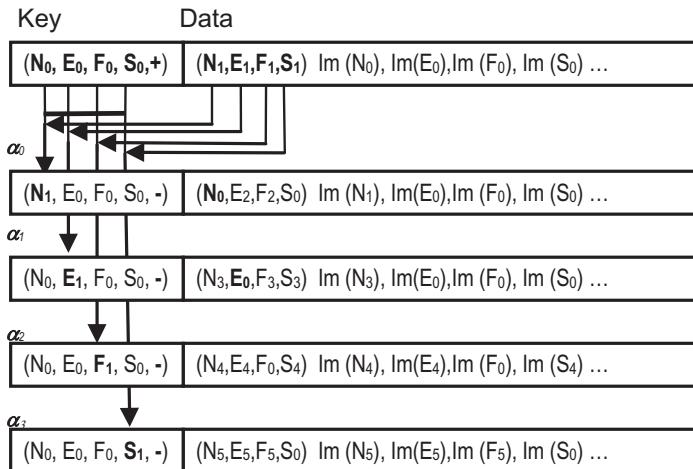


Fig. 2. Representation of a 3-G-Map as a relation with additional symmetry constraints determined by the combinatorial character of the involutions

2.1.3 Basic Objects and Relationships

We require the cellular complexes to be orientable, and the corresponding G-Maps to be oriented. This implies that there are two classes of cell-tuples of the same cardinality, but carrying different *polarity*. Different from Lienhardt (1991), we exclude the possibility that an involution attaches a cell-tuple to itself ($f(x) = x$), e.g. at the boundary of the cellular complex. Thus we make sure that involution operations always link pairs of cell-tuples of opposite sign. Instead, we introduce a special non-standard cell, the outside, or “*universe*”, which in general needs not be simply connected and may comprise holes and islands, and provide the *universe* also with cell-tuples and involutions on the boundary. This approach increases the number of objects to be handled, but simplifies some operations and algorithms.

2.2 The Topological Toolbox

The proposed topological extension of a MRDB provides the client with a set of basic tools for the management, navigation and retrieval of topological information. These tools can be combined into short programs that ful-

fil more complex tasks. At a later stage, we intend to support the combination of these tools, together with spatial and thematic queries, into user defined scripts. The *topological toolbox* outlined here is based upon the model of oriented 2D- or 3D-manifolds represented by oriented 2D/3D-G-Maps. Within the toolbox, different classes of functions and operations can be distinguished:

Basic tools:

1. Methods that instantiate connected components or entire G-Maps
2. Orbits and other iterators that support the navigation on a G-Map.
3. Counting functions that deliver information on the topological structure
4. Elementary transformation operations on the cell level
 - 4.1. Euler operations that conserve the connectivity properties
 - 4.2. Non-Euler operations
5. Methods that handle the spatial embedding of the G-Map
 - 5.1. Representation of celltupels and cells by geometrical objects in the MRDB
 - 5.2. Methods that handle “colourings” of the cell complex, i.e. classifications of cells

Complex tools for single representation:

6. Consistency checks
7. Topological queries
8. Methods to handle combinations of topological, spatial and other queries
9. Exchange of data with the exterior
 - 9.1. Construction of consistent GMaps from imported spatial data.
10. Methods that handle topological exceptions.

Complex tools for multiple representation:

11. Hierarchical structures based on links between Levels Of Detail
 - 11.1. Building of a Hierarchical G-Map.
 - 11.2. Navigation on the hierarchical structure.
12. Multiple classifications (Fradin et al., 2002)
 - 12.1. Management of multiple classifications
 - 12.2. Navigation on multiple classifications
13. Progressive transformations, especially agglomeration.
 - 13.1. Creation and management of “delta” operations
 - 13.2. Variation of LOD by progressive application of delta operations

With growing experience, the toolbox may of course be extended. In particular, at a later stage of development, the toolbox may be integrated into a scripting language accessible to client applications. In the following we discuss some of the above-mentioned tools.

2.2.1 Instantiation

The creation of an empty G-Map is straightforward, but not very useful. In order to be able to apply transformation operations, we must start with the creation of a first d-cell. An initial cell of dimension 2 may either be a triangle, or a somewhat “smaller” minimal cell consisting of two nodes, two edges, the inner face and surrounded by the outer *2D-universe*. A minimal 3-cell is composed of a solid bounded by two minimal 2-cells, and is surrounded by the *3D-universe*.

2.2.2 Navigation on the G-Map

Orbits. By definition, the repeated application of the same involution α_i , means cycling between the two states $\alpha_i(x)$ and $x = \alpha_i(\alpha_i(x))$. If we combine two or more different involutions $\alpha_i, \alpha_j, \dots$, the picture changes: $\alpha_i(\alpha_j(x))$ in general cannot be expected to coincide neither with x nor with $\alpha_j(\alpha_i(x))$. However, Lienhardt’s definition of a d-G-Map (see above) implies that

$\beta(x) = \alpha_0(\alpha_2(x))$ is an involution for dimension $d = 2$ and $d = 3$, as are
 $\gamma(x) = \alpha_0(\alpha_3(x))$ and $\delta(x) = \alpha_1(\alpha_3(x))$ for $d = 3$.

An *orbit* in a d-G-Map is defined as the maximal subset of d-cell-tuples that can be reached from a start d-cell-tuple cts by any combination of a given subset of the involutions $\alpha_0 \dots \alpha_d$. Let us note such an orbit $orbit^d(cts, \alpha_i, \dots, \alpha_j)$ or shorter $orbit^d_{i\dots j}(cts)$.

Singletons $\{cts\}$ and pairs $\{cts, \alpha_i(cts)\}$ may be considered as the result of *trivial orbits*.

In 2D, non-trivial orbits are:

- $orbit^2_{01}(cts)$ enumerates the cell-tuples associated with a face ,
- $orbit^2_{02}(cts)$ enumerates the cell-tuples associated with an edge,
- $orbit^2_{12}(cts)$ enumerates the cell-tuples associated with a node,
- $orbit^2_{012}(cts)$ enumerates all cell-tuples of the connected component containing cts .

In 3D, the following classes of non-trivial orbits exist:

-
- $\text{orbit}^3_{01}(\text{cts})$, $\text{orbit}^3_{02}(\text{cts})$, $\text{orbit}^3_{03}(\text{cts})$, $\text{orbit}^3_{12}(\text{cts})$, $\text{orbit}^3_{13}(\text{cts})$, $\text{orbit}^3_{23}(\text{cts})$
 - $\text{orbit}^3_{012}(\text{cts})$, $\text{orbit}^3_{013}(\text{cts})$, $\text{orbit}^3_{023}(\text{cts})$, $\text{orbit}^3_{123}(\text{cts})$ enumerate the cell-tuples associated with a solid, face, edge, node n, respectively.
 - $\text{orbit}^3_{123}(\text{cts})$ enumerates all cell-tuples of the connected component containing cts.

By marking each cell that has once been encountered, and successively skipping marked cells, we can use orbits to enumerate all cells of a given type encountered on the way. Most of these orbits can be implemented using nested simple programming loops, except for the following ones: $\text{orbit}^2_{012}()$, $\text{orbit}^3_{012}()$, $\text{orbit}^3_{123}()$, $\text{orbit}^3_{0123}()$. These can be implemented recursively e.g. as depth-first search (Lévy, 1999). This approach has a shortcoming: whereas non-recursive implementations yield sequences of cell-tuples that form a continuous path in the graph representation of the G-Map, recursively implemented orbits yield sequences that may comprise discontinuities when the procedure is forced to backtrack. We are therefore currently investigating to what extent these orbits can be implemented as continuous closed loops.

Loops. For the basic *split* operation on 3-cells (solids) described below, we need to define a closed loop of cell-tuples on the inner surface of the solid, i.e. a sequence of the form

$$ct_0 - \alpha_{i0} - \dots - \alpha_{ik-1} - ct_k - \alpha_{ik} - \dots - \alpha_{iN-1} - ct_0,$$

passing from each cell-tuple ct_k to the next by an involution α_{ik} . This sequence may be user-defined rather than generated by an orbit. Once it has been defined, e.g. as a linked list, it is easy to use an iterator to produce the sequence of cell-tuples in similar fashion as by an orbit. On the other hand, an orbit that does not make any “jumps” produces a continuous closed path of cell-tuples that can be stored in form of a loop. It can be represented in condensed form, either as a start cell-tuple followed by a sequence of pairwise adjacent cell-tuples, or followed by a sequence of involution selectors $i = 0, 1, 2 [3]$, that at each step determine the next involution to follow. Obviously it takes at most 4 bits to encode such a selector. A possible external representation is a start cell-tuple $ct_0(n_0, e_0, f_0, s_0)$ followed by a sequence of cells c_i , where each transition α_{ki} from cell-tuple $ct_i(n_i, e_i, f_i, s_i)$ to the next is defined by exchanging c_i against the corresponding cell in ct_i .

Orbits and loop iterators are used to traverse connected subsets of a d-G-Map. They are instantiated with a start cell-tuple, and in the case of loop iterators with the sequence of involution selectors. Methods are `.hasNext()`, `.getNext()`, `.hasPrevious()`, `.getPrevious()`, `.size()`, and `.restart()`.

2.2.3 Counting Functions

A number of consistency checks can be implemented by counting cells or cell-tuples and comparing the result to an expected value. In an oriented G-Map, the number of cell-tuples obviously must be pair. As we provide the outside of the boundary of the cellular complex with cell-tuples, every edge of a 2-G-Map is accompanied by four cell-tuples - therefore their total number must be a multiple of four. If we start an orbit $orbit^2_{012}$ with any cell-tuple cts , it will enumerate all cell-tuples of the connected component containing cts . If their number is different from the total number of cell-tuples in the G-Map, there must be more than one component. The most prominent counting function however calculates the *Euler-Poincaré-characteristic*

$$\chi = \sum (-1)^i N_i,$$

where N_i denotes the number of cells of dimension i . For a cellular 2D-manifold in \mathbf{R}^3 , χ amounts to $V-E+F$ where V, E, F are the total numbers of vertices, edges and faces and verifies the Euler-Poincaré formula

$$\chi = V - E + F = 2(S - G) + (L - F),$$

where S is the number of inner “shells”, G is the *genus* of the surface, and L is the number of *bounding loops*. Operations on the meshing that do not affect χ are considered as conserving the topological properties of the surface, and are called *Euler operations* (cf. Mäntylä 1988).

2.2.4 Elementary Operations on Cells

Lévy (1999) describes the attachment and detachment of a complete d-cell to a cell complex and the induced `sew()` and `unsew()` operations in the G-Map, i.e. the insertion or deletion of a set of cell-tuples in the G-Map and of new involution connections. We distinguish Euler operations that transform the cell structure, but preserve the topological properties represented by the Euler characteristic, from non-Euler operations that affect the connectivity.

Euler Operations on Cells

- “Split a d-cell in two by the insertion of a separating d-1-cell”, and the inverse operation
- “Merge two d-cells by the elimination of a separating d-1-cell” (fig. 3, 4). We do not consider geometry issues, as we make no assumption about the geometrical representation of the G-Map. In 3D, we obtain the following 6 operations:
- “Split an edge inserting a node, a face inserting an edge, a solid inserting a face.”
- “Merge two edges removing a node, two faces removing an edge, two solids removing a face.”

Note that for the merge operations, a number of topological conditions must be verified to avoid the creation of inconsistent configurations.

By passage to the *dual* G-Map, which is achieved by inverting the roles of dimension and co-dimension, we obtain a second set of 6 dual operations:

- “split a node inserting an edge, an edge inserting a face, a face inserting a solid.”
- “merge two nodes removing an edge, two edges removing a face, two faces removing a solid.” Again these operations are applicable only in specific configurations.

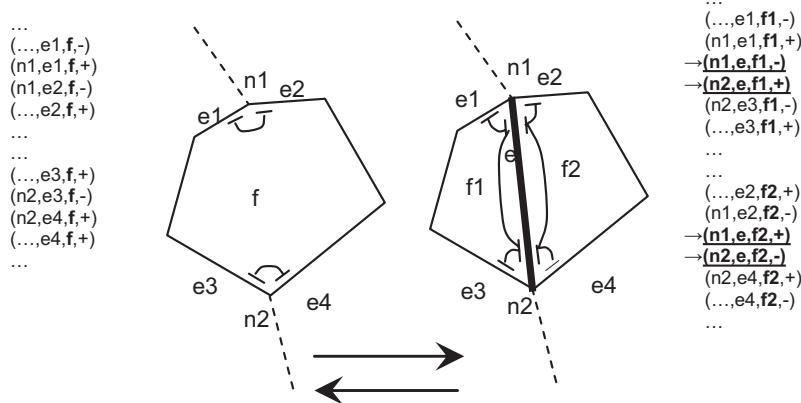


Fig. 3. Elementary topological Euler 2d-operation: a face f is divided in two faces f_1, f_2 by the insertion of a dividing edge e . Sewing generates newly inserted cell-tuples. The inverse operation requires additional validity checks

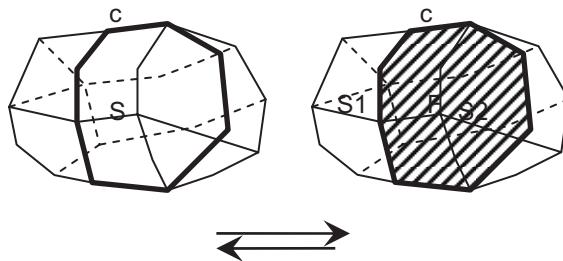


Fig. 4. Elementary topological Euler 3d-operation: a solid s is divided in two solids s_1, s_2 by the insertion of a dividing face f . The inverse operation requires additional validity checks

Non-Euler Operations on Cells

These operations change the connectivity. Let us mention here only the following:

- “Create a disconnected simple d-cell.”
- “Destroy an isolated d-cell.”
- “Merge two connected components inserting a bridge cell.”
- “Split a connected component removing a bridge cell.”

3 Persistence

The graph representation of a d-G-Map can be made persistent using object-oriented DBMS, while the tabular representation can be stored in an ORDBMS. We intend to implement both types of representation in order to examine their respective advantages and disadvantages.

3.1 Object-Oriented DBMS

OODBMS software is used to make the graph representation of a d-G-Map (**fig. 1**) persistent and map it onto external storage in a more or less transparent way. In an earlier project, we have developed an object-oriented 3D-geometry DBMS named GeoDB3D (Breunig et al., 2004) based on commercial OODBMS software, and we plan to extend GeoDB3D by a topological access structure based on 3D-G-Maps.

3.2 Object-Relational DBMS

We implemented 2D- and 3D-G-Maps using the ORDBMS PostgreSQL (PostgreSQL.org, 2006), with the intention to later combine the topological software with the open source GIS *PostGIS* (PostGIS.org, 2006). This combination will serve as the platform for the planned 2D-application example outlined below. Cell-tuples lend themselves to a natural relational representation. Involutions can be modelled e.g. using foreign keys. In a cell-tuple (*node_id*, *edge_id*, *face*, [*solid_id*]) the combination of cell identifiers, augmented by a positive or negative sign,

$$(\textit{node_id}, \textit{edge_id}, \textit{face_is} [, \textit{solid_id}], \textit{sign})$$

is used as a unique *cell-tuple key*, while the identifiers of the cells to be exchanged by the involutions are stored as data (**fig 2**). The data access by cell-tuple keys is enhanced by the use of a unique sorted index or by a hash index.

The involutions are implemented in two steps:

1. From a given cell-tuple entry, create a new cell-tuple key by exchanging exactly one *cell_id*.
2. Use the new cell-tuple key to retrieve the corresponding complete entry from the database.

By iterating these two operations, entire orbits can be implemented. This method may turn out to be very slow, and therefore we investigate, to what extent the topological operations can be accelerated by RDBMS functionality like sorting, indexing, clustering and caching. Also a hybrid approach may be considered, namely using the RDBMS to retrieve ordered subsets of cell-tuples, and performing the topological operations in memory.

4 Multi-Representation

Agglomeration, simplification, suppression, displacement and typification are well-known generalisation transformations. Agglomeration and suppression directly affect the topology of a map. Simplification may affect the interior structure of an object, whereas displacement may be employed in order to maintain topological consistency under a geometrical generalisation operation - e.g. if smoothing a river bend would leave a building on the wrong side. In a first step, we concentrate on the agglomeration of

contiguous cells by the application of sequences of Euler transformations, being aware that this approach covers only a selection of generalisation operations. In a later second step, we will try to model the agglomeration of disjoint cells using transformations of classifications/colourings of cellular complexes. Whereas the choice of the generalisation method is taken by the geoscientist, supported by specialised software (cf. Haunert 2004, Haunert & Sester 2005), we focus on the representation of the given transformations and of the resulting relationships between LOD in the MRDB. Relationships between cells at different levels can be defined by explicit links, or by indicating the sequence of elementary operations that transform a cellular complex at scale A into a cellular complex at scale B . It is the task of the database software, to keep track of the incurred changes, and if possible to support *transitions* with *commit* and *rollback* operations. Moreover, by storing incremental transformations rather than entire versions, the amount of storage space may be reduced at the cost of some extra calculation time.

4.1 Hierarchies of Maps

For the representation of multi-scale topology, Lévy (1999) proposes Hierarchical G-Maps (HG-Maps): The Agglomeration of neighbouring cells results in a classification of cells on the more detailed level A , each class being associated with one cell on the less detailed level B . It can be represented by an n:1-mapping from one level A to level B . As cells are merged, and interior boundaries disappear, the number of cell-tuples is reduced. The cell-tuples on level B can be associated with a selection of cell-tuples on the lower level A , or be identified with a subset of the latter. If the geometry of the remaining cell boundaries is not changed after the agglomeration step, higher level cell-tuples may delegate their geometrical embedding (co-ordinates, lengths, angles etc.) to their counterparts on the lower level (**fig. 5**), so that a higher-level edge is geometrically represented by a sequence of lower-level arcs and vertices. Otherwise, links with a new higher-level geometrical embedding must be established.

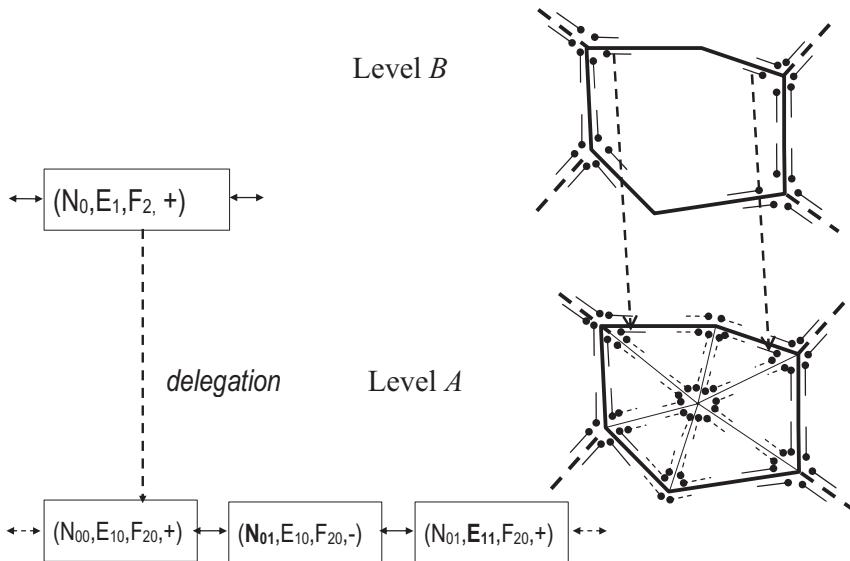


Fig. 5. Generalisation by aggregation in a hierarchical 2-G-Map. Cell-tupels (darts) are symbolised by small pins

4.2 Progressive Variation of LOD

Due to the necessity of keeping all levels of detail consistent with each other, any changes in an MRDB are first introduced at the greatest scale, and then propagated upwards using appropriate generalisation methods (Haunert & Sester, 2005). Carrying this “dynamical” approach a step further, we investigate the applicability of progressive meshes. The progressive triangulation method (Hoppe, 1996) uses two localised elementary operations, namely the “edge collapse” and its inverse, the “vertex split”, to coarsen or to refine a triangle network incrementally in both directions, by successively applying a sequence of stored “delta” operations. This method is well suited for *progressive transmission*, as it can reduce the amount of data exchanged between a geo-database server and a local client (cf. Shumilov et al., 2002).

Though generalized maps are abstract simplicial complexes, Hoppe’s method cannot be adapted: Though a d-cell-tuple is an abstract d-simplex, its $d+1$ components belong each to a different class defined by dimension, and therefore cannot be merged, like in an “edge collapse” operation on a

triangle network. An analogous argument holds for the inverse "vertex split" operation. Instead, we investigate the possibility to use combinations of the Euler elementary split and merge operations on cells to model the transformation of topology induced by generalisation. Different from Hoppe's method, the progressive mesh transformation is controlled by the external generalisation method, and not by a given optimisation criterion. Note that the merge operations are applicable only in certain configurations and hence require supervision. We expect progressive mesh transformation to allow for dynamical navigation between levels of detail in both directions, at the expense of some calculation and storage overhead. Moreover, under the above-mentioned condition, the implementation of "rolling back" topology updates by applying inverse operations in reverse order will be straightforward, so that transactions can be implemented.

5 Application Example

Our project partners at Hanover Institute of Cartography (IKG) are investigating methods that generalise land use maps by an automatic aggregation of parcels using thematic and/or geometric criteria (Haunert, 2004, Haunert & Sester, 2005). The resulting hierarchies of maps at different LOD are stored in a MRDB (Anders and Bobrich, 2004). The n:1 relationships between polygonal faces at greater and lesser scale are represented in tabular form (**Fig. 6**).

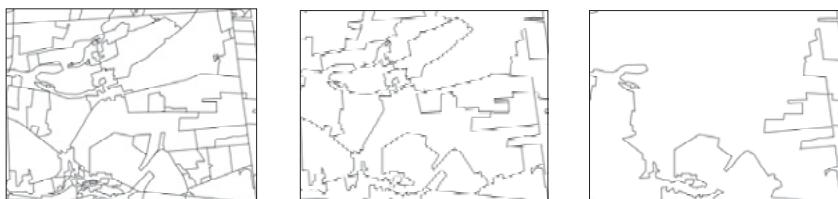


Fig. 6. Application example by courtesy of J. Haunert, IKG Hannover University: a section of ca. 2 % of a digital map on land-use at three different scales. The generalisation was controlled only by thematic criteria.

From a set of separate maps at different scales imported into PostGIS/PostgreSQL, we derive corresponding G-Maps. Topological consistency is checked and the Euler characteristic and some basic statistics are established. The n:1 relationship between maps at different LOD induces a classification of the cells of greater scale. Using the elementary merge

operations described above, groups of cells of the same class are agglomerated either until a 1:1 correspondence is established, or until inconsistent configurations are detected. Thereafter, unnecessary nodes on the boundaries of the agglomerated cells are eliminated while edges are merged. If no premature stop has been encountered, the 1:1 relationship between faces and agglomerated cells is used to determine the relationships between edges, nodes, and in consequence cell-tuples. The resulting hierarchical G-Map represents the interrelations between the topologies at different LOD.

Conclusion and Outlook

In this paper we have discussed topological abstraction applied to Multi-Resolution Databases. A general approach to the handling of topology in MRDBs based on oriented hierarchical d-G-Maps has been proposed. The approach can be used as a data integration platform for 2D, 3D, and 4D topological data, starting the standardisation of topological data models at different dimensions and LOD. A topological toolbox containing basic and complex tools for single and multiple representations has been proposed. In our future work particular interest will be given to the transformation between external representations and HG-Maps. As application examples we will use 2D datasets from IKG Hannover (Haunert and Sester, 2005), and own 3D datasets.

Acknowledgements

We thank our students Uwe Sander, Edgar Butwilowski and Bjoern Bro scheit for programming first implementations within the DFG joint re search project “Abstraction of Geoinformation”.

References

- Anders K.-H. and Bobrich J. (2004): “MRDB Approach for Automatic Incremental Update”. ICA Workshop on Generalisation and Multiple Representation, Leicester.
- Berners-Lee T., Hendler J., Lassila O. (2001): „The Semantic Web. Scientific American, 05/2001.
- Breunig M., Bär, W. and Thomsen A. (2004): “Usage of Spatial Data Stores for Geo Services.” 7th AGILE Conference on Geographic Information Science, Heraklion, Greece, 687-696.

- Brisson E. (1993): "Representing Geometric Structures in d Dimensions: Topology and Order." *Discrete & Computational Geometry* 9, 387-426.
- Fradin D., Meneveaux D. and Lienhardt P. (2002): "Partition de l'espace et hiérarchie de cartes généralisées." AFIG 2002, Lyon, décembre 2002, 12p.
- Guarino N. (1998): Formal Ontology and Information Systems. Proceedings FOIS'98, IOS Press, Amsterdam, 06/1998.
- Haunert J.-H. (2004): "Geometriotypwechsel in einer Multi-Resolution-Datenbank". Mitteilungen des Bundesamtes für Kartographie und Geodäsie, Aga-Tagung 2004, 9p.
- Haunert J.-H. and Sester M. (2005): "Propagating updates between linked datasets of different scales" XXII Int. Cartographic Conference, A Coruna, Spain July 11-16.
- Hoppe H. (1996): "Progressive meshes." ACM SIGGRAPH 1996, 99-108
- Köthe U. (2000): "Generische Programmierung für die Bildverarbeitung". Dissertation, FB Informatik, Universität Hamburg, 274 p.
- Lévy B. (1999): "Topologie Algorithmique - Combinatoire et Plongement" PhD Thesis, INPL Nancy 202p.
- Lienhardt P. (1991): "Topological models for boundary representation: a comparison with n-dimensional generalized maps." *Computer Aided Design* 23(1), 59-82.
- Mallet J.L. (2002): "Geomodelling." Oxford University Press, 599 p.
- Mallet J.L. (1992): "GOCAD: A computer aided design programme for geological applications". In: Turner, A.K. (Ed.): *Three-Dimensional Modelling with Geoscientific Information Systems*, NATO ASI 354, Kluwer Academic Publishers, Dordrecht, 123-142.
- Mäntylä M. (1988): "An Introduction to Solid Modelling". Computer Science Press, 401 p.
- PostGIS.org (2006): <http://postgis.refractions.net/documentation>
- PostgreSQL.org (2006): <http://www.postgresql.org/docs>
- Raza A. and W. Kainz (1999): "An Object-Oriented Approach for Modelling Urban Land-Use Changes." ACM-GIS 1999: 20-25.
- Shumilov S., Thomsen A., Cremers A.B. and Koos B. (2002): "Management and visualisation of large, complex and time-dependent 3D objects in distributed GIS", ACM-GIS 2002, 113-118.
- Winston P.H. (1992): Artificial Intelligence, 3rd edition, Addison Wesley, 1992.

Intelligent Simulation of Hydrophysical Fields by Immunocomputing

Alexander O. Tarakanov¹, Ludmilla A. Sokolova¹ and Sergey V. Kvachev²

¹ St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences
14-line 39, St. Petersburg, 199178, Russia

[\(tar,lucy\)@iias.spb.su](mailto:(tar,lucy)@iias.spb.su)

² Transas, Maliy pr. 54-4, St. Petersburg, 199178, Russia
sergey.kvachev@transas.com

Abstract. This paper proposes a novel approach of simulation of hydrophysical fields (HPF), using immunocomputing (IC). The essence of the approach is the recognition of the value of scalar field in any point at any time by using a formal immune network (FIN) which is trained by the known values in some points of the field at some instants. Numerical experiments include simulations of sea surface temperature (SST) by utilizing average data obtained from the space monitoring of the Caspian Sea. The obtained results suggest that the IC simulator is promising for on-line modeling and visualization of real-world HPF.

1 Introduction

The ability to simulate HPF is important for many applications, including research, monitoring and three-dimensional (3D) computer modelling of the ocean floor, surface and volume [12]. However, existing methods of computation are too complicated and too slow to satisfy the practical applications of real-world end-users. Usually, the computation of HPF is undertaken using a sophisticated hydrodynamic model that replicates the factors that define a hydrophysical environment.

Theoretically, this task can be solved by two different ways based on 1) wave or 2) ray propagations. In the first case, HAF is determined by partial differential equations of mathematical physics; in the second case, the task becomes analogous to ray tracing in 3D computer graphics.

A lot of methods of conventional computing have been proposed to cope with this task [2, 3]. However, this makes difficulties for finding a convenient method of 3D simulation of HAF on-line with accuracy sufficient for real-world applications at the same time.

This paper proposes a novel approach to simulation of HPF in the seas and oceans, using IC which was being developed and applied in our previ-

ous work as an approach of computational intelligence [6, 9]. Within this approach, the task of the simulation of HPF can be formulated using an analogy of the computation of the ecological atlas [9] or index [4, 10].

The essence of the approach is the recognition of the value of scalar field in any point at any time by using FIN which is trained by the known values in some points of the field at some instants. The developed IC simulator also includes the embedded data generator (DG), which performs linear interpolation by time and spatial interpolation by splines. This DG provides the IC simulator with realistic training and test data as well as fast computing and visualization of HPF with any resolution. Numerical experiments include SST simulations of the Caspian Sea during 2006 by using average data obtained from the space monitoring [1].

2 Formalization of the Task

Let G be set of points of 3D grid of HPF: $g_{ijk} \in G$, where coordinates (latitude, longitude, depth) are determined by indices i, j, k . Let $t \in N$ be discrete moments of time (instants). Let scalar function $f(N_0, G_0)$ be known for some points of this field $G_0 \subseteq G$ at some instants $N_0 \subseteq N$. For example, these data can be taken from monitoring, atlas, data base, modeling, etc. The task is to compute the function $f(N, G)$ for any point of HAF at any instant.

When known values of the field $G_0(t \in N_0) \subset G$ do not form the boundary conditions $G_0 = G$, this real-world task hardly yield to any conventional method of mathematical physics including pseudodifferential equations [3]. However, the task can be formulated as a task of pattern recognition.

Let us form a set of parameters (column vector) for any point: $X = [x_1 \dots x_n]'$ which includes also the coordinates of the point and the time. Note that upper stroke $[]'$ is the symbol of matrix transposing so that X is column vector whereas X' is string vector. Let the values of function f_1, \dots, f_m are known for some training vectors X_1, \dots, X_m . Then the task is to recognize the value $f(X)$ for any vector X .

3 IC approach to Pattern Recognition

According to [9], pattern recognition by IC simulates *molecular recognition* by computation of the *binding energy* between any protein (*antigen*) and the proteins of immune system (*antibodies*). In mathematical terms, any n -dimensional input vector X ("antigen") is projected to q -dimensional space of FIN and recognized by class of the nearest point of FIN. Coordinate axes of FIN's space are determined by the right singular vectors ("antibodies") of singular value decomposition (SVD) of the training matrix $A = [X_1 \dots X_m]$, where X_1, \dots, X_m are n -dimensional training vectors.

According to [6], pattern recognition algorithm of IC is as follows (in a pseudocode):

```

Training
{
    1st stage training // map data to FIN ("antigen processing")
    {
        Get training patterns;
        Form training matrix;
        Compute SVD of the training matrix;
        Store q singular values // "binding energies"
        Store q right singular vectors; // "antibody-probes"
        Store left singular vectors; // cells of FIN
    }
    2nd stage training // compress data by FIN's "maturation"
    { // compute consecutively for all cells of FIN:
        Apoptosis; // kill unnecessary cells
        Immunization; // correct mistakes of Apoptosis
    }
}
Recognition
{
    Get pattern; // "antigen"
    Map the pattern to FIN;
    Find nearest cell of FIN;
    Assign class of the nearest cell to the pattern;
}
```

4 IC approach to Simulation of HPF

The above IC algorithm of pattern recognition with a discrete output (class) has been modified in our previous paper [8] to provide real-valued (continuous) output. Consider the mathematical description of the modified algorithm to provide the recognition of the value $f(X)$ of HAF for any input vector $X = [x_1 \dots x_n]'$.

1. Form training matrix $A = [X_1 \dots X_m]'$ of dimension $m \times n$, where m is total number of points in the training subfields $G_0 (t \in N_0)$ over all training instants N_0 .
2. Compute first q singular values s_1, \dots, s_q and corresponding left and right singular vectors L_1, \dots, L_q and R_1, \dots, R_q by SVD of training matrix, where $q \leq r$ and r is rank of the matrix:

$$A = s_1 L_1 R_1' + \dots + s_q L_q R_q' + \dots + s_r L_r R_r'.$$

3. For any n -dimensional vector X , compute its mapping $Y(X) = [y_1 \dots y_q]'$ into q -dimensional space of FIN:

$$y_1 = \frac{1}{s_1} X' R_1, \dots, y_q = \frac{1}{s_q} X' R_q.$$

4. Among the training points of FIN $Y(X_1), \dots, Y(X_m)$, determine p nearest to $Y(X)$ points Y_1, \dots, Y_p and their Tchebishev distances (according to [6]):

$$d_1 = \|Y_1 - Y\|, \dots, d_p = \|Y_p - Y\|.$$

5. Interpolate $f(X)$ by the following sum:

$$f = \sum_{i=1}^p a_i f_i,$$

where $f_i = f(Y_i)$ are training values, which correspond to the nearest points of FIN, whereas the coefficients a_i are determined by the distances:

$$a_i = \frac{1}{1 + d_i \sum_{j \neq i}^p \frac{1}{d_j}}.$$

It can be shown that

$$\sum_{i=1}^p a_i = 1.$$

It can be also shown that $f = f_i$ if $d_i = 0$ for any i (then $d_j \neq 0$ for any $j \neq i$). Thus, IC should not produce inaccurate data for the training set.

5 Numerical Experiments

Data for this example have been obtained from the space monitoring of the Caspian Sea [1]. These data represent twelve arrays of the monthly average temperature of the surface in centigrade degree ($T^\circ C$) which is given for all months of 2006: $month = \{1, \dots, 12\}$.

Consider 2D grid $G = \{g_{ij}\}$, $i = 1, \dots, 22$, $j = 1, \dots, 14$, which is determined by the northern latitude $lat = \{36.5, \dots, 47.5\}$ N and the eastern longitude $lon = \{47.0, \dots, 54.0\}$ E and covers the surface of the Caspian Sea with the resolution of 0.5° ($= 30'$).

Let the function $f = T(t, G)$ be daily SST, where $t = [month, day]$, $day = \{1, \dots, days(month)\}$, $days(1) = 31$, $days(2) = 28, \dots$, $days(12) = 31$. Let the task is to compute this function $T(t, G)$ for any day of the year t and for any point of the grid G .

Consider the monthly average SST as an SST for the 15th day of every month. Table 1 shows an example of such data in December: $month = 12$, $day = 15$. Note that the points where $T = -5.0$ either do not belong to the

Caspian Sea or the corresponding SST may be unavailable (due to the cloud clusters or other causes).

Table 1. Average SST of the Caspian Sea in December 2006 obtained by space monitoring

-5.0	-5.0	-5.0	-5.0	-5.0	-5.0	-5.0	3.4	3.6	-5.0	-5.0	-5.0	-5.0	-5.0
-5.0	-5.0	-5.0	-5.0	-5.0	3.7	2.1	1.0	2.0	2.5	2.2	2.1	3.2	-5.0
-5.0	-5.0	-5.0	-5.0	4.8	2.0	1.2	1.0	2.0	2.5	2.3	3.0	3.6	-5.0
-5.0	7.0	5.8	3.7	2.6	2.9	1.8	2.0	2.9	5.1	0.7	2.4	-5.0	-5.0
-5.0	4.3	3.8	5.8	5.7	4.5	-5.0	-5.0	-5.0	-5.0	-5.0	-5.0	-5.0	-5.0
5.8	5.3	8.0	7.4	8.0	8.1	7.6	-5.0	-5.0	-5.0	-5.0	-5.0	-5.0	-5.0
-5.0	6.5	7.0	9.8	-5.0	10.0	10.0	10.4	-5.0	-5.0	-5.0	-5.0	-5.0	-5.0
-5.0	7.9	10.0	10.6	10.3	-5.0	-5.0	11.4	10.2	-5.0	-5.0	-5.0	-5.0	-5.0
-5.0	7.6	9.4	10.4	11.3	10.9	-5.0	10.7	11.8	10.2	-5.0	-5.0	-5.0	-5.0
-5.0	-5.0	9.0	10.9	10.9	10.8	10.9	10.8	10.9	11.4	12.1	9.9	-5.0	-5.0
-5.0	-5.0	-5.0	8.0	10.7	10.9	11.0	11.2	11.3	11.3	12.5	-5.0	-5.0	-5.0
-5.0	-5.0	-5.0	-5.0	8.6	10.5	10.7	10.4	11.0	12.0	12.5	12.6	-5.0	-5.0
-5.0	-5.0	-5.0	-5.0	-5.0	10.2	11.4	10.9	10.7	12.1	13.2	14.2	-5.0	-5.0
-5.0	-5.0	-5.0	-5.0	-5.0	-5.0	11.1	11.2	12.3	14.1	13.9	-5.0	-5.0	-5.0
-5.0	-5.0	-5.0	-5.0	-5.0	11.7	11.4	11.4	12.5	13.7	15.2	13.9	-5.0	-5.0
-5.0	-5.0	-5.0	-5.0	-5.0	12.5	12.4	13.1	13.0	14.1	14.9	14.6	13.4	-5.0
-5.0	-5.0	-5.0	-5.0	12.5	13.6	14.3	14.6	14.2	15.6	15.4	15.1	10.9	-5.0
-5.0	-5.0	-5.0	-5.0	14.2	14.5	14.9	15.1	15.0	15.5	16.1	15.6	13.6	11.5
-5.0	-5.0	-5.0	-5.0	14.3	15.2	15.4	14.9	15.2	15.5	15.7	15.9	14.1	12.4
-5.0	-5.0	-5.0	-5.0	-5.0	15.3	15.1	15.6	15.4	15.3	15.2	15.8	16.1	13.6
-5.0	-5.0	-5.0	-5.0	-5.0	-5.0	-5.0	15.9	15.9	15.7	15.4	15.8	16.5	16.0
-5.0	-5.0	-5.0	-5.0	-5.0	-5.0	-5.0	-5.0	-5.0	-5.0	-5.0	-5.0	-5.0	-5.0

Let us define vector $X = [x_1 \dots x_5]'$ by the following parameters: $x_1 = month$, $x_2 = day$, $x_3 = lat$, $x_4 = lon$, $x_5 = T_{DG}$, where T_{DG} is the SST computed by DG (using the linear interpolation of the temperature by time). Let training vectors X_1, \dots, X_m , where $m = 22 \times 14 \times 12 = 3696$, be determined by the known values of SST: $T_1 = f(X_1), \dots, T_m = f(X_m)$, so that $x_2 = 15$ for any training vector. Using these training vectors, the task of the IC simulator is to compute $22 \times 14 \times 365 = 112420$ values of $T = f(X)$ for any day of the year and for any point of the grid.

The results of the simulation of T_{IC} have been estimated relatively to the linear interpolated T_{DG} by using the following standard deviation (mean square error):

$$e = \sqrt{\frac{1}{k} \sum_{i=1}^k (T_{IC} - T_{DG})_i^2},$$

where $k(year) = 22 \times 14 \times 365$ for the full year or $k(day) = 22 \times 14$ for any day.

Depending on the parameters of FIN (q, p), Tab. 2 shows the yearly deviation, whereas Tab. 3 shows the maximal daily deviation. According to

Tab. 2, and Tab. 3, an optimal performance of the IC simulator is obtained for 2D FIN ($q = 2$) with 5 nearest points for the interpolation ($p = 5$). Figure 1 shows the graphic of daily deviation of T_{IC} for such optimal FIN during the year of simulation. Figure 2 shows the screenshots of the DG and IC simulator on December 2nd when daily deviation is maximal (according to Tab. 3). The spline grid on both screenshots is equal to 0.1° ($= 6'$) by latitude and longitude.

Table 2. Yearly deviation ($^\circ\text{C}$)

Number of nearest points	Dimension of FIN	1	2	3	4	5
1		93.26	0.56	0.50	0.56	141.84
2		77.26	0.48	0.45	0.54	135.55
3		71.59	0.46	0.44	0.54	128.75
4		68.71	0.44	0.44	0.55	131.59
5		66.92	0.43	0.44	0.56	131.04
7		64.76	0.41	0.45	0.58	116.45
10		63.03	0.40	0.46	0.6	110.6
20		60.99	0.39	0.49	0.64	105.39

Table 3. Maximal daily deviation ($^\circ\text{C}$)

Number of nearest points	Dimension of FIN	1	2	3	4	5
1		14.78	1.36	1.51	1.79	22.36
2		13.94	1.29	1.48	1.72	22.21
3		13.64	1.25	1.48	1.69	21.93
4		13.48	1.24	1.48	1.67	22.04
5		13.35	1.23	1.48	1.67	22.04
7		13.29	1.22	1.49	1.66	21.27
10		13.25	1.21	1.48	1.69	20.84
20		13.21	1.19	1.46	1.65	20.49

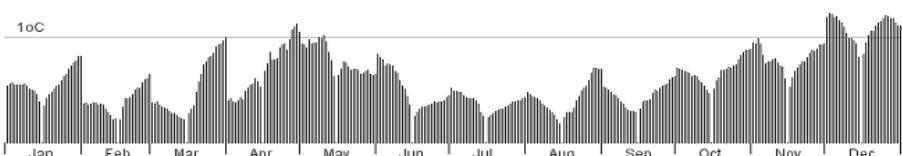


Fig. 1. Daily deviation of SST for the optimal FIN

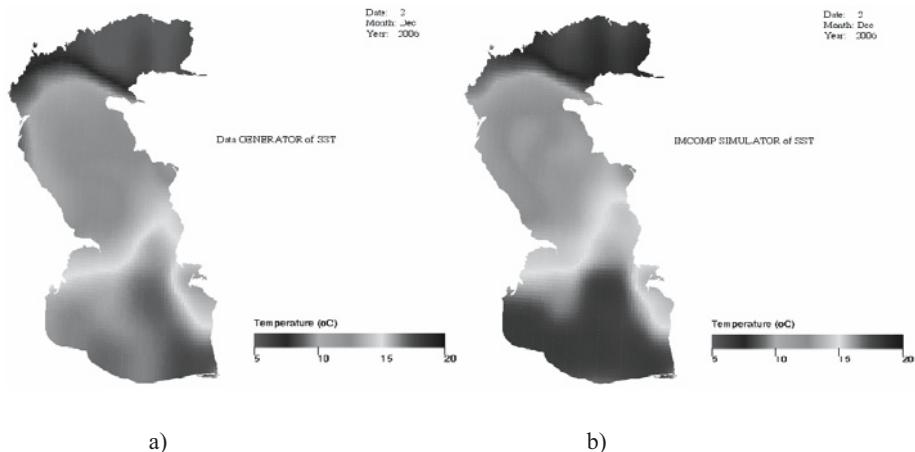


Fig. 2. SST of the Caspian Sea on December 2nd computed and visualized by the
a) data generator and b) IC simulator

6 Discussion

Tables 2 and 3 as well as Fig. 2 demonstrate that the IC simulator provides quite sufficient accuracy of recognition of SST (yearly deviation $\approx 0.4^{\circ}\text{C}$ and maximal daily deviation $\approx 1.2^{\circ}\text{C}$). Meanwhile, the training data in the above experiments make less than 3.3% (3696/112420) of total data to be simulated and the simulator reconstructs the training SST exactly (see Fig.1 where daily deviation $e = 0$ for any $day = 15$).

It is worth noting that the time of IC simulation of the above SST (22×14) for any day is less than 0.5 sec (Intel 2.0 GHz), which includes also interpolation by splines (with 6' resolution by latitude and longitude) and visualization of the results (Fig. 2a or 2b). Apparently, the advanced methods of mathematical physics (finite elements method, pseudodifferential equations, etc. [2, 3]) could not provide similar speed of computation.

IC simulation of monthly and daily dynamics of SST (Fig. 3 and Fig. 4) demonstrates that a forecast of long term behavior of real-world HPF seems questionable for any conventional model (e.g., draw attention to the diversity of configurations of SST in Fig. 3, Fig. 4, and Fig. 2b). In addition, it is rather problematic to select the parameters of such model so that

it would correspond to the real-world training data (irregular, inaccurate, incomplete, etc.).



Fig. 3. Simulation of the monthly dynamics of SST

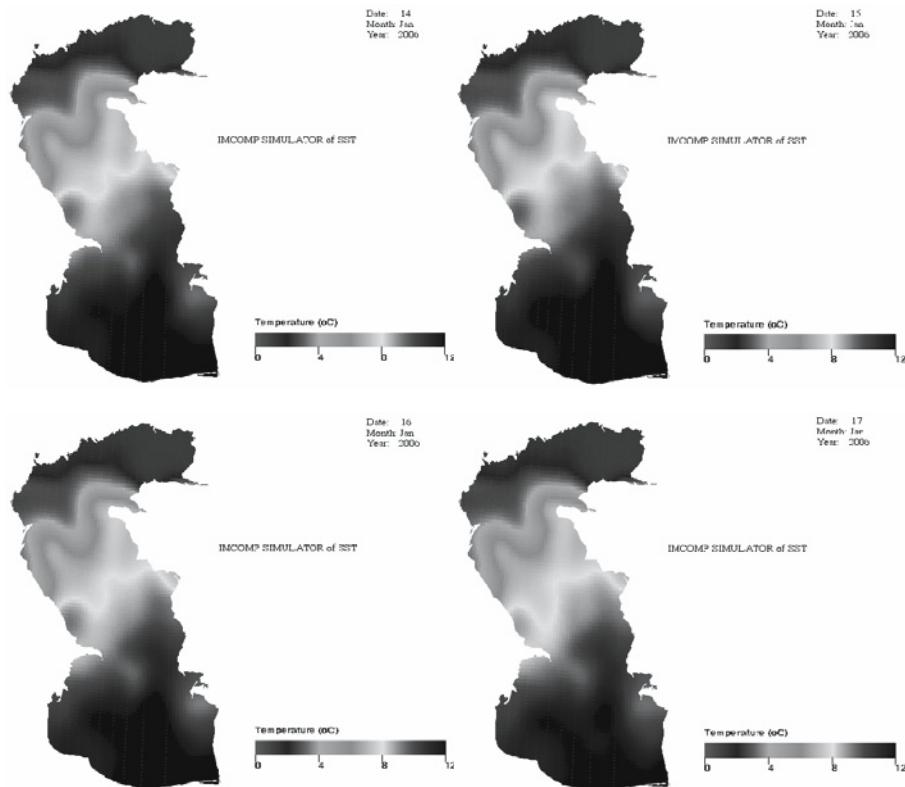


Fig. 4. Simulation of the daily dynamics of SST

Recent results [7, 8] also show the clear advantage (in training time and accuracy) of the use of IC over both neurocomputing and the more conventional least square method of interpolation. These results together with the advantages in 3D modeling [5] make IC rather promising for on-line simulation and visualization of real-world HPF including the fast simulation of hydroacoustic fields [8].

References

1. NASA: Ocean Color Time-Series Online Visualization and Analysis [<http://reason.gsfc.nasa.gov/Giovanni>]
2. Brekhovskikh, L. and Lysanov, Y. Fundamentals of Ocean Acoustic (3rd ed.). Berlin: Springer-Verlag, Berlin (2003)

3. Saranen, J. and Vainikko, G. Periodic Integral & Pseudodifferential Equations with Numerical Approximation. Springer-Verlag, Berlin (2006)
4. Sokolova, L.A. Index design by immunocomputing. Lecture Notes in Computer Science, Vol. 2787. Springer-Verlag, Berlin (2003) 120-127
5. Tarakanov, A. and Adamatzky, A. Virtual clothing in hybrid cellular automata. *Kybernetes (Int. J. of Systems & Cybernetics)* 31(7/8) (2002) 394-405
6. Tarakanov, A.O., Goncharova, L.B., and Tarakanov, O.A. A cytokine formal immune network. Lecture Notes in Artificial Intelligence, Vol. 3630. Springer-Verlag, Berlin (2005) 510-519
7. Tarakanov, A. and Prokaev, A. Identification of cellular automata by immunocomputing. *Journal of Cellular Automata* (accepted for publication)
8. Tarakanov, A., Prokaev, A., and Varnavskikh, E. Immunocomputing of hydroacoustic fields. *International Journal of Unconventional Computing* (accepted for publication)
9. Tarakanov, A.O., Skormin, V.A., and Sokolova, S.P. Immunocomputing: Principles and Applications. Springer-Verlag, New York (2003)
10. Varnavskikh, E., Kravchenko Y., and Sokolova L. Possibilities of application of immunocomputing in practical tasks of military pedagogic and psychology. *Maritime Digest* 8 (2005) 41-43 (in Russian)
11. Varnavskikh, E.A., Prokaev, A.N., and Tarakanov, A.O. Problem statement of the estimation of indicators of physical fields by immunocomputing with an example of control of acoustic field. *Science-Technical Digest "Frontier-2005"* (ed. Zотов, Ю.М.). Graphics, Moscow (2006) 378-382 (in Russian)
12. Wille, P. Sound Images of the Ocean in Research and Monitoring. Springer-Verlag, Berlin (2005)

Visual Modeling of Spatial Processes

R.P.Sorokin

St Petersburg Institute for Informatics and Automation, Russian Academy of Sciences
39, 14 Line, VO, St. Petersburg, 199178, Russia
sorokin@oogis.ru

Keywords: Visual modeling, spatial process, artificial intelligence, ontology, geographic information system (GIS), inference engine, scenario, rule, open source.

1 Introduction

Visual computer simulation unlike statistic simulation concentrates on visual presentation of process or phenomenon being simulated. The main result of such a simulation is rather ‘a picture, possibly changing in time’, than a number, and a human observer is a consumer of this result. Such a simulation permits to clearly imagine the real process and concentrate one’s entire mental faculties on solving various problems relevant to the simulated process. The visual computer simulation is mostly developed in aircraft simulators, however, the problem solved by their means – pilotage training and aircraft control - is extremely narrow, at that, the simulation is only limited by realistic idea of what can be seen by a pilot from the cockpit and on the monitoring devices.

GIS allows to clearly visualize geodata, i.e. physical and other values related to geographic coordinates. Electronic maps serve as an example. So far an abundance of geodata has been accumulated, and most of them are visually available via Internet (Google Map, Live Search, etc.). Those applications allow visualizing any part of terrain. The issue of visual *dynamic* representation of different processes going on the earth surface became demands of the times. A spatial process concept needs to be defined in order to arrive at a technical solution of this problem by corresponding software development.

2 Spatial Processes

Various spatial processes are taking place in the spatial environment simulated by GIS. The process is known as a successive change of states. Accordingly, *the spatial process* - is a successive change of the space states. The spatial processes may include both the natural ones going on in the space, and various processes of human activities. Such processes can be sufficiently complex and include a variety of different objects. The complexity of processes can be a hierarchical one, when more global processes result from joint proceeding and interference of separate, specific processes, which, in turn, decompose into elementary processes. In terms of the object-oriented simulation, the processes are considered a totality of participating objects. The objects participating in spatial processes can be divided into the point and extensive ones.

The *point* objects shall include the objects whose own dimensions can be disregarded at simulation of a given process. Such objects of spatial environment are primarily characterized by coordinates of the point of their location at a given instant of time. In different scenarios the examples of such objects can include the transportation means, residential localities, buildings, facilities, etc.

As *extensive* shall be considered the objects whose own dimensions can not be disregarded at simulation of this or that other scenario. Such objects are simulated by means of spatial figures of both regular and arbitrary shape. The atmospheric fronts, oil spills caused by accidents, contaminated areas, detection areas, etc. can be mentioned as examples. It shall be noted that the same physical objects in different scenarios or even, more specifically and with reference to GIS, in different map scales, can be regarded both as the point and extensive ones, for instance, the settlements.

The simulation of spatial processes also distinguishes the *moving* and *stationary* objects. The point moving objects are additionally characterized by motion vector, i.e. by direction and velocity. The extensive regular objects can also be sometimes characterized by a moderate number of parameters. For instance, a transient circular area as well as a point object can be characterized by a displacement vector. The ‘radiation’ can be described as an extensive circular area with a determined velocity of extension. However, an arbitrary figure is described, as a rule, by a spline consisting of multitude of points, and each point of the spline can have its own displacement vector. This results in arbitrary change of form and location of the spatial object.

The external manifestations of spatial processes consist in various *events* in different points of space and various *changes* in the space.

Among those the following can be emphasized:

- Emergence and disappearance of objects in different points of space
- Displacement of the point objects along different paths
- Change of shape and size of extensive objects according to different laws
- Display of various events by means of conventional signs (by color, sound, signs, etc.) and written text (voice).

3 Sets of Regulations and Scenarios

For the purposes of visual computer simulation the spatial processes can be described by two different approaches. One approach is based on the concept of scenario, the other one – on the concept of a set of regulations. In a sense these two approaches are the opposite ones. The scenario-based approach is used when the general algorithm of the process, i.e. the consecutive actions it consists of, is available. The approach based on a set of regulations is used when the general algorithm of the process is not available except for the certain factors it is conditioned to that can be formulated as regulations for the ongoing process.

The approach based on regulations is a conventional one and widely used in expert systems and described in publications in detail.

The concept ‘scenario’ is a generalized concept of algorithm. This generalization is aimed at a possible representation of the totality of parallel interrelated processes as a scenario, since the complex spatial processes develop exactly like this.

A *scenario* can be formally defined as a succession of phases and solutions.

A *phase* is the totality of elementary actions realized in succession or **in parallel**. This is the main distinction of scenario from algorithm, as the actions are realized in a strict succession in the latter.

The *decision* is a point where the process flow may alter its direction extensively due to certain conditions formed by the moment. Hence, formally, the decision is defined as a totality of *branches* (directions of the scenario continuation). The realization of procedure of decision making, i.e. the selection of direction for the further realization of scenario can vary between an automatic (programmed) and ‘manual’ (which includes human involvement).

The *actions* are the building blocks for scenarios. They represent specific elements of activity of the scenarios’ participants which may be real-

ized in different ways. The actions result in certain events or spatial changes.

The second important distinction of scenario from algorithm is the fact that while the actions have a duration in time, the duration of actions in an algorithm is usually disregarded (sometimes a number of steps is taken into account). The difference in actions' duration along with a possibility of their parallel realization in scenarios results in the necessary synchronization of the scenario branches realized in parallel.

The scenarios and their phases are visually presented as schemes. The scheme of a scenario is a two-dimensional oriented graph whose nodes are phases and points of decision, and whose arcs are the transition lines from one phase (decision) to another. The scheme of a phase is also a two-dimensional oriented graph whose nodes are the actions and whose arcs are the transition lines from one action to another.

Both types of graphs can be cyclic ones.

An example of scheme of a simplified scenario of rescue operation after a collision of two oil tankers in the sea, and one of its phases is shown in Fig. 1.

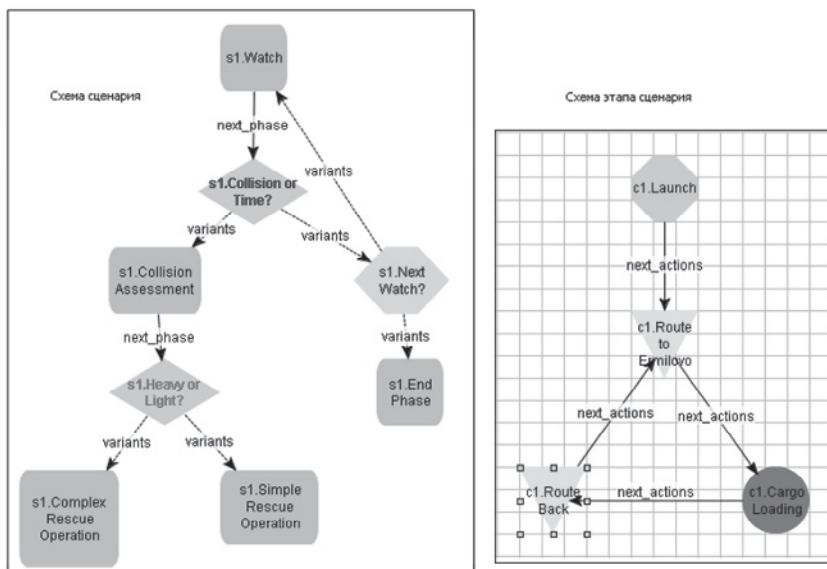


Fig. 1. Example of scenario and phase schemes

4 Implementation through RETE Algorithm

Notwithstanding the difference between above two approaches to the formal description of complex spatial processes, a united instrument based on the RETE network-type inference engine, is expedient to be used for their computer implementation.

The RETE algorithm [1] has been developed and successfully used during decades for computer implementation of inference engine in the systems of representation and processing of the rule-based information, particularly, in expert systems. This algorithm that proved its efficiency is a *network* one in a sense that all regulations are transferred to the network before the algorithm starts working. The input information for the algorithm consists of the facts describing the states of the outer world and its changes. During the work the facts “travel” through the network from the net input nodes to the output ones and are stored in the nodes at intervals between their “travels”. Due to such structure and organization the efficiency of RETE algorithm is independent of the regulations number. Its performance is accounted for the fact that only new facts are subject to testing.

The efficient performance of RETE algorithm requires a sufficiently slow alteration of the multitude of fact. The current implementations deal successfully enough with hundred of thousands of facts when tens of thousands of regulations are imposed on.

According to the research, the computer implementation of various cycle-based parallel processes as per all simultaneously realized actions can be adequately substituted by RETE network consisting of regulations describing the implementation of each type of action. At that, the samples of actions for specific objects act as facts. For instance, for the action «Arrive X to point Y» the network has the Arrive Rule regulation. Accordingly, for the parallel realization of two actions it will be sufficient to enter two facts into the network: ‘Arrive A to point B’ and ‘Arrive C to point D’.

Thus, the uniform approach based on regulations and RETE algorithm will stipulate for the following general scheme of realization of spatial processes visual modeling (Fig. 2)

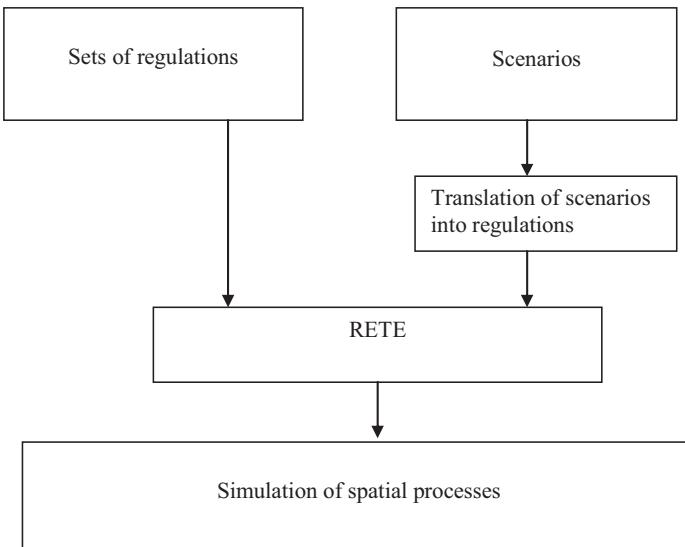


Fig. 2. General scheme of implementation of visual modeling of spatial processes

5 Scenarios Ontology

To represent the information being the contents of both scenarios as well as of regulations' sets, an advanced approach based on the object-oriented *ontologies* is used. A universal scenarios' ontology common for all applications, has been developed for this purpose. This approach implies a possible specification of the universal ontology with reference to various specific applications by defining the classes of specific actions to form more specific applied scenarios from.

The most general concept of the ontology of scenarios is the concept 'activity' represented by an abstract **Activity** class. The specific subclasses of this class include the following:

- **Scenario** – scenario,
- **Phase** – phase of scenario,
- **Decision** – decision,
- **Action** – action.

The main attributes of Activity class and its subclasses include:

- **status** – the state of given type of activity,
- **run** – ‘run’, the specific realization of given scenario,
- **context** – the context or ‘environment’ of this realization of scenario.
- **period** – the limited period of time for realization of action.

The attribute ‘status’ may take on different values depending on the state of a given specific type of activity. For instance:

- **START** – beginning (initialization) of given type of activity,
- **DOING** – the given type of activity is being realized,
- **REPEAT** – the repeated action,
- **DONE** – this activity has been realized, etc.

The scenarios may be *common* (universal), i.e. appropriate for different cases. The *variables* defining the objects and values contained in the scenario are used for presentation of such scenarios. When the simulation of such scenarios is started, the variables shall be given specific values. The aggregate of pairs of the ‘variable– value’ type is called *context*. If one common scenario is started at the same time with two different sets of variables values, two concurrently realized implementations of the scenario take place. The *run* concept serves for their identification and distinction. The *run* identifier is used, for instance, at interruption of specific implementation of scenario, based on it the inference engine determines the phases and actions to be interrupted and to be kept going.

The ontology of scenarios contains the following classes to support the universal scenarios: **Context** (context), **Variable** (variable), **Pair** (‘variable– value’ pair), **Object** (any object). Besides, there are classes in the ontology for visual presentation of scenarios, as follows: **ScenarioScheme** (scenario scheme), **PhaseScheme** (phase scheme).

The ontology of regulations sets depends on the implementation of the used RETE inference engine. In any case it contains the concepts of regulation and set of regulations.

6 Implementation of Prototype

For implementing a prototype of considered architecture of visual modeling of spatial processes the ready program packages with open source free available in the Internet, have been selected as follows:

- **Protégé** – editor of ontologies [2],
- **JBossRules** – RETE type inference engine [3],

- **OpenMap** – library of GIS [4].
- **Groovy** – interpreter of scenario language for Java virtual machine [5].

Protégé has a special built-in technology to connect additional modules (plug-ins). The other above-listed packages are connected to Protégé via this technology means. Besides, Protégé has a special GraphWidget component which enables visual design of scenario schemes and its phases. The ontology of scenarios, ontology of regulations' set supporting the JBossRules inference engine, and ontology of geoinformational systems supporting the GIS OpenMap library, have been developed in Protégé.

A special **RuNA** plug-in to directly support a simulation has been developed with the functions as follows:

- flow of *simulated time* in real and arbitrary time scale and generation of events with time reference;
- displacement of transient objects according to the current time scale;
- control of simulation process including start, stop, alteration of time scale and map, map navigation, creation, mapping and removal of objects, control of their displacement parameters, etc.

JBossRules has a built-in technology for *domain subject languages* (DSL); via it a simple language to describe regulations for a specific data domain and a ‘vocabulary’ for translating expressions in this language into expressions in the Java language can be easily developed. This technology is used in the system of visual modeling, in particular, for description of specific scenarios’ actions. The additional advantage of such approach consists in the following: if we want to switch to another ‘actuator system’, for instance, to another GIS library or another simulating system, we have only to change the vocabulary in the ‘right part’, i.e. as regards translation of expressions in the language of data subject in the ‘left part’ into the Java language, substituting the API calls of corresponding new components of the actuator system. As a result, all developed scenarios will start to work unchanged in the new actuator system.

The Groovy language is used in the proposed prototype for the special **Calculus** type of action. This is a ‘low-level just in case’ which allows to introduce any text in the Groovy language and realize by its means any processing of information in the parts of scenario without typical actions available. Besides, the ‘scripts’ realizing various auxiliary tasks during the work of the system, are written in the Groovy language.

This prototype of environment of the visual modeling of spatial processes has been published together with the sources on the Internet site <http://www.oogis.ru> under **JBRSTab** name.

Conclusion

The lengthy testing of the considered prototype of the visual modeling environment has proved the adequacy of the approach to the simulation of multiple parallel spatial processes based on RETE algorithm. The implementation of complex models consisting of numerous nested scenarios incorporating hundreds of phases and thousands of actions, has not revealed any delays, failures or computer overloads even in the multi-accelerated simulation mode. However, it should be noted that the point objects have mainly participated in the most of these scenarios while the extensive objects, especially those of irregular form, were only tested fragmentary. The further study of this approach to the spatial simulation is planned to be aimed, primarily, at improvement of details of simulating the extensive objects of complex shape changing in time and the interaction of various objects and shapes during simulation process. Besides, a significant updating of general architectural design of the system is planned in regard to appearance of a new version of the JBossRules system which is considerably improving its capacity in developing systems based on regulations.

References

1. Charles Forgy, "Rete: A Fast Algorithm for the Many Pattern/Many Object Pattern Match Problem", Artificial Intelligence, 19, pp 17-37, 1982
2. Holger Knublauch, An AI tool for the real world. Knowledge modeling with Protégé, JavaWorld.com, 06/20/03
3. James Owen, Open source rule management, InfoWorld.com, November 02, 2006
4. Owen Densmore, Java Geography for the Smart Mob, Part 1: OpenMap, O'Reilly OnJava.com March 6, 2003
5. Dierk Koenig with Andrew Glover, Paul King, Guillaume Laforge and Jon Skeet. Groovy in Action . Manning, 2006

Universal Data Model for Intelligent GIS

Sergey N. Potapychев and Andrey V. Pan'kin

St Petersburg Institute for Informatics and Automation, Russian Academy of Sciences
39, 14 Line, VO, St. Petersburg, 199178, Russia
{potapychev, pankin}@oogis.ru

Abstract. The advanced geographic information technologies despite their rapid development not always can provide for a solution of wide scope of problems like development of rather complex systems incorporating a large number of components and dealing with a great variety of information sources. The above mentioned problems can be solved by using in intelligent GIS (IGIS) a proposed universal data model and a method consisting in reduction of all information flows to a common universal data model along with arranging for informational exchange via reducing all flows thereto.

Keywords: Object-oriented approach, intelligent geographical information system, data model, metaclass, relation, XML.

1 Introduction

The efficiency of use of geographic informational technologies in various fields of human activities is determined by a fact that 85% of information a person deals with during a life time is territorially conjunct. Geographic informational systems (GIS) are actively introduced in various areas of management, industry, transportation, ecology, health care, etc.

GIS are actively entering scope of many activities, including every day life and business, thus, providing for better-founded decision making. However, GIS are rather means and techniques improving the decision-making procedure effectiveness than a decision-making tool. GIS provides for queries' answers, and spatial data analysis functions, visual representation of analysis results in easy-to-use form. At the same time, such systems are expected to process heterogeneous information. This problem is caused by a need for concurrent processing of large body of information received from heterogeneous sources. As a rule, each information type is described by a certain data model, or, a so-called data format. So, a prob-

lem of providing for interpretation of data from heterogeneous sources, i.e. data harmonization arises. The following two tasks should be resolved to receive the above problem solution:

- To develop a common universal data format, providing for processing the data from any source;
- To transform data from whatever source to the above universal data format.

One of the ways for solution of this problem is The use of the common universal data model developed in OOGIS Laboratory of SPIIRAS and the method of data reduction to this model set an approach to the given problem solving.

2 Universal Data Model IGIS

This model development is based on the object-oriented paradigm. All information regarding the interaction of external components with IGIS is reduced to a form representing a list of data entities (Entities) and the hierarchy of their properties (Keys). This model introduces a new essence of relations (Relations) as an extension of the object-oriented approach.

The following elements – classes are assumed in the developed model (see Fig. 1):

1. List of entities – Entities.
2. Separate entity – Entity.
3. List of entities' properties – Keys.
4. Separate property – Key.
5. List of relations involving the entity – Relations.
6. Separate relation – Relation pointing to the relation's object and determining the object's role in this relation.
7. Union – Union. Consolidates all above entities in one model.

The key element of the model is the Entity class aimed at representing data on any entity of the data domain regardless of their nature (it can be a ship, a ship control system, etc.).

A data message may contain data about several entities, not about one entity, of the data domain. The Entities class is intended to implement a feasibility of several entities' data representation. This allows forming the list of data entities (Entity).

Each informational entity may have a totality of properties determining an entity state. The Keys class is intended to deal with the list of properties for each entity (Entity).

The Key class is intended to deal with each property of a data entity (Entity). The examples of such properties may include ship's location, her course and speed. The properties can be both simple ones to hold the value of one parameter (course, speed of ship), and complex ones to represent the complicated complex hierarchy-type characteristics of the entity (ship's location including latitude and longitude). To represent simple properties a certain property's value is brought to conformity with each named property. In case of presentation of a complex property its value would consist in its list of properties (Keys). The properties in the list can also be simple and complex.

The Relation class is intended to represent relations binding the entities. It shows what relation the entity is involved in and its role therein. The 'Relation' category has no analogies in the object-oriented approach. It is introduced because currently available object-oriented models do not allow general realizing the properties of relations binding different entities. In a result, these models have no explicit universal mechanism for binding the entities. The object-oriented approach does allow a good description of separate entities, however, a union of entities resulting in emergence of new properties (a typical example of a system) is hard to describe. The implicit techniques have to be used, that lead to a complication of the model's realization and, consequently, to impossibility of complex systems development. To eliminate the above disadvantage the new essence – the relation (Relation) was introduced. Its introduction enabled to describe the properties of interrelated entities including the system forming ones. That is, apart from properties (Keys), the entity (Entity) acquired a structure of relation (Relation).

The relations can be illustrated by the following example: The ship is moored to a berth (see Fig. 2). At that, three entities exist: 'Sylvia Line ferry', 'Berth no. 1', 'Mooring to a berth' and two relations binding these entities. The first relation binds up the entity 'Sylvia Line ferry' with entity 'Mooring to a berth', specifying the entity's role in this relation. The plays a role of 'Ship at a berth'. The second relation binds up the entity 'Berth no. 1' with entity 'Mooring to a berth', the role played is 'The berth the ship is moored to'.

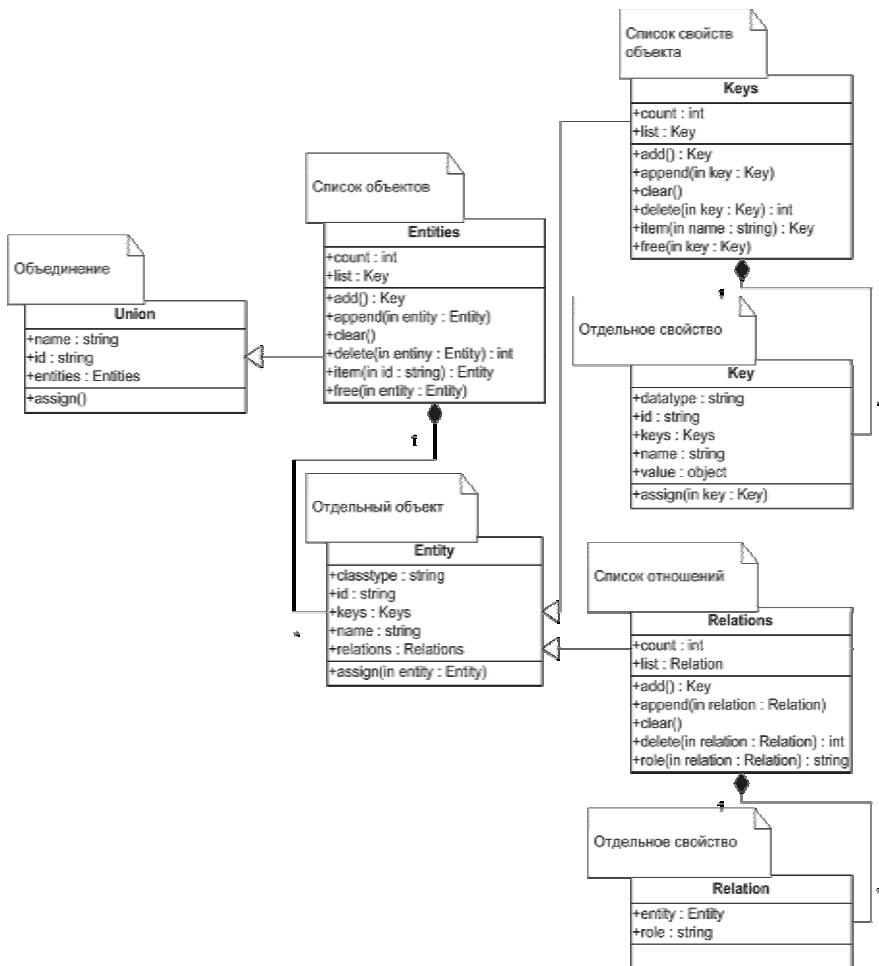


Fig. 1. Diagram of classes of the universal model for data representation

Another example is a search of entity by the ship. The yacht ‘Black Cuttlefish’ is searching for the wrecked yacht ‘Disaster’ in the specified area ‘near Moshchny Island’. This example considers four entities: ‘Search operation’, ‘Yacht “Black Cuttlefish”’, ‘Yacht “Disaster”’, ‘Area near Moshchny Island’ and three relations. The first relation binds up the entity ‘Yacht “Black Cuttlefish”’ with the entity ‘Search operation’, the role assigned is ‘Search forces’. The second relation binds up the entity ‘Yacht “Disaster”’ with the entity ‘Search operation’, the role played is ‘Search Object’. The third relation binds up the entity ‘Area near Moshchny Island’ with the entity ‘Search Operation’, the role assigned is ‘Search Area’.

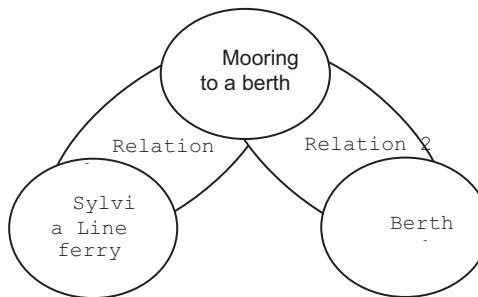


Fig. 2. Example of relation realization

Using the relations' mechanism allows to provide for an information on any relations binding the entities up.

The last class of the model — union class (Union) is intended for presentation of one complete message and is a container enclosing a list of data entities with all their properties and relations.

Thus, the universal data model in IGIS allows to:

1. use data from external information sources, including the automatic ones
2. prepare and enter the basic data to perform necessary calculations and simulation
3. provide for exchange of working documents
4. provide data for the situation's electronic mapping.

The model supports a dynamic development of necessary data entities with the characteristics and properties required by the concrete application. The use of the model allows realizing the information exchange between components of intelligent GIS and external systems.

3 Method of Data Reduction to Universal Model

As specified above, the second task is the reduction of data acquired from any source to the universal data format. For this purpose the method consisting in development of a logic-and-semantic dependence of data acquired from external components is used. The main point of this method consists in developing the logic-and-semantic dependence of data for each component of the kind (see Fig.3).

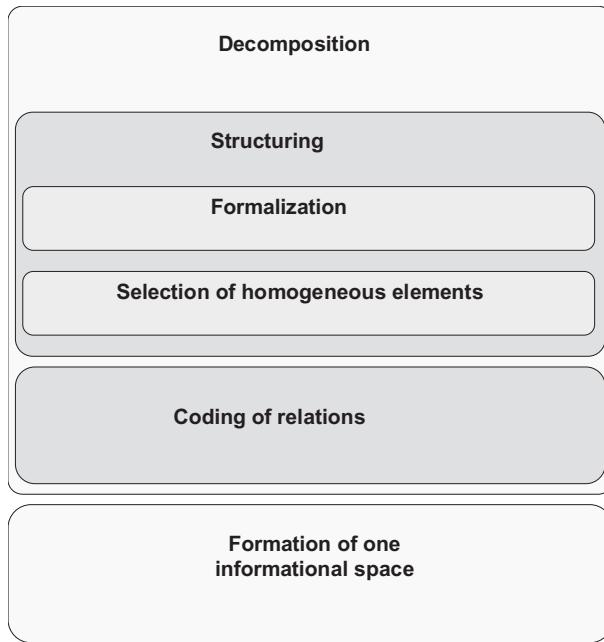


Fig. 3. Method of data reduction to the universal data model in IGIS

The first step of data reduction to the universal data model representation in IGIS consists in decomposition of data acquired from the date component. The decomposition process includes three steps: two stages of its structuring and the stage of encoding the relations (semantic relations) between the elements.

At the first stage the data acquired from the component are formalized, and the basic scheme of the data structure description is selected. The acquired structural elements do not disclose the inner contents of elements at this stage.

The second structuring stage objective is the separation of maximally possible number of addressable homogeneous elements using the necessary number of structuring levels and combined schemes of the elements of description of relations between components.

The result of the first step of data reduction to the universal data model representation in IGIS consists is the logical model of data acquired from the data component provided with specification of all structuring schemes in conformity with these schemes used therein. All homogeneous and heterogeneous components, as well as relations in the form of semantic relations between components, are addressed.

As elements of the universal data model, the components of logical model are divided into two categories:

1. data on the entity (entities);
2. data on the entity (entities) properties including class of entity, name of entity, relation the entity is involved in.

Reduction of data from the component containing the data on the entity implies the creation of new entity of the universal data model, while the reduction of data from the field containing the data on property of the entity implies altering the feature of existing entity of the universal data model.

At the second step of data reduction to the universal data model representation in IGIS the entire totality of data received from all components is considered. The main objective of the second step consists in forming a comprehensive information space of IGIS. This allows considering the information received from a component in the context of information from other components, thus, permitting to assign the reliability priorities, determine the order of interaction with components, etc.

The developed method of data reduction allows transforming the data of existing components to the developed universal data model in IGIS. IGIS based on the universal data model permits its integration with any other external components, and to use of information from this subsystem by other components.

4 Schemes of XML-Document for Universal Data Model

The answer to the emerging question about the best possible material date medium in the format of universal data model is the document of XML format, as the most universal currently due to the following reasons:

1. The XML format allows to describe the structure of any document using the DTD rules (document type definition) or XSD (XML scheme definition), thus, arranging for an adaptation of the XML language constructions specific for the data domain. MathML (mathematic markup language), SVG (scalable vector graphics), GML (geographic markup language), DAML+OIL and many others may serve as examples of such applications;
2. XML is easily understood by a human understanding, as well as for easily amenable to code walkthrough;
3. XML allows validating data stored in the documents, verifying hierarchical relations inside, and establishing a uniform standard for

- the structure of documents whose contents can be heterogeneous data. This means that it can be used at development of complex information systems, for which the issue of information exchange between different applications, active in the same system, is highly important;
4. The documents in XML format can be a unique way of data storage, both incorporating means for information deciphering and representation;
 5. The processing software for documents in XML format are not complicated, and currently various software products intended for a work with documents in XML format are available and freely distributed;
 6. The XML format is adapted for exchange of information in computer networks.

The following scheme of document in XML format (see Fig.4), developed according to the notation of W 3 C consortium, is suggested to be used for the universal data model.

```
<?xml version="1.0" encoding="UTF-8"?>

<!--W3C sceme of universal model in IGIS-->

<xss:schema
xmlns:xss="http://www.w3.org/2001/XMLSchema"
elementFormDefault="qualified">

    <!--root element-->

    <xss:element name="union">

        <xss:complexType>

            <xss:sequence>

                <xss:element ref="entities"/>

            </xss:sequence>

        <!--file name-->

        <xss:attribute name="name"
type="xss:string" use="required"/>

    </xss:element>
```

```
<!--File ID-->

    <xss:attribute name="id"
type="xss:string" use="required"/>

    <!--File description-->

        <xss:attribute name="description"
type="xss:string"/>

    </xss:complexType>

</xss:element>

<!--objects search-->

<xss:element name="entities">

    <xss:complexType>

        <xss:sequence>

            <xss:element ref="entity" maxOc-
curs="unbounded"/>

        </xss:sequence>

    </xss:complexType>

</xss:element>

<!--separate object-->

<xss:element name="entity">

    <xss:complexType>

        <xss:sequence>

            <xss:element ref="keys" minOc-
curs="0"/>

        </xss:sequence>

    </xss:complexType>

</xss:element>
```

```
<!--the object unique ID-->

<xs:attribute name="id"
type="xs:string" use="required"/>

<!--Object name-->

<xs:attribute name="name"
type="xs:string" use="required"/>

<!--Object description-->

<xs:attribute name="description"
type="xs:string"/>

<!--Object type-->

<xs:attribute name="classtype"
type="xs:string"/>

<!--Object owner-->

<xs:attribute name="ownerid"
type="xs:string"/>

</xs:complexType>

</xs:element>

<!--list of the object/entity specifications-->

<xs:element name="keys">

<xs:complexType>

<xs:sequence>

<xs:element ref="key" maxOccurs="unbounded"/>

</xs:sequence>

</xs:complexType>
```

```

    </xs:element>

    <!--specification of entity-->

    <xs:element name="key">
        <xs:complexType>
            <xs:sequence>
                <xs:element ref="keys" minOccurs="0"/>
            </xs:sequence>
            <!--name of specification-->
            <xs:attribute name="name" type="xs:string" use="required"/>
            <!--value of specification-->
            <xs:attribute name="value" type="xs:string" use="required"/>
            <!--type of specification-->
            <xs:attribute name="datatype" type="xs:string"/>
            <!--specification description-->
            <xs:attribute name="description" type="xs:string"/>
        </xs:complexType>
    </xs:element>
</xs:schema>

```

Fig. 4. XML-scheme of universal data model in IGIS

Let us consider XML elements and their attributes in detail.

‘Union’ is the root element of a document in the XML format. It is a container for all entities with information transferred in XML-document. The same element contains the name of the document, its description.

The element ‘union’ contains the element ‘entities’. The element ‘entities’ represents a list of all entities transferred in XML-document. Each entity is transferred in the element ‘entity’. This element contains the name, description, type of entity and specification of its owner. The element ‘entity’ contains its properties in the element ‘keys’. The element ‘keys’ represents a list of properties of the entity. Each property of the entity is stored in the element ‘key’. The element ‘key’ contains the property name, its value, description and type of property. The element ‘key’ may be a simple one, that is, contain one value, or a group one, that is, that property is a group of other properties. In case of the group property, it in turn contains the list of properties in the group. The properties may contain the values of simple types, such as line, number, logical expression, or the entity may be the value of the property. In that case the entity containing that property will be the owner of that entity. All described above can be presented as a table (see Table 1).

Table 1. Description of the elements of XML-scheme

nos	Element	Description
1.	Union	<p>Root element of XML-document intended for identification of document, is a container for all entities and has the following attributes:</p> <ul style="list-style-type: none"> - id – unique identifier; - name – name of union; - description – description of union (optional attribute); <p>Contains the element – ‘entities’</p>
2.	Entities	Element representing a container of entities (entity)
3.	Entity	<p>Element intended for storage of information on the entity, has the following attributes:</p> <ul style="list-style-type: none"> - id – unique identifier; - name – name of entity; - description – description of entity (optional attribute); - classtype – class of entity (optional attribute); - ownerid – unique identifier of the entity owner of this. (For implementation of the aggregation relation) (optional attribute). <p>Contains the element ‘keys’</p>
4.	Keys	Element representing a container of properties of the entity (key)

nos	Element	Description
5.	Key	<p>Element intended for storage of information on the concrete property of the entity, has the following attributes:</p> <ul style="list-style-type: none"> - id – unique identifier; - name – name of property; - description – description of property (optional attribute); - datatype – type of property (optional attribute). <p>May contain the element ‘keys’ for implementation of hierarchical properties.</p>

The use of documents in the XML format based on the above scheme, allows to work with data corresponding the universal data model IGIS.

5 Conclusion

The use of universal data model in IGIS will provide for a decreased laboriousness of the dataware automation due to the uniform approach to the harmonization of data from heterogeneous sources, such as:

- external, including the automatic information sources;
- special and calculation tasks;
- workflow system;
- geographical information interface.

Besides, it will allow increasing the integration level both between IGIS and components, and between separate components, providing for the possibility of quick data ware adaptation to changing requirements and increase of its functionality without considerable alterations of existing elements.

The suggested universal data model has been practically implemented and used in various developments of the laboratory.

References

1. A.V.Pankin, S.N.Potapychev. The Object - Oriented Approach to Creation of Geoinformational Systems//Industry, no. 3 (33), 2003. – PP. 108-109
2. A.V.Pankin, S.N.Potapychev. Informational System as Principal Support for Decision Taking // Innovations, no. 8 (65), 2003. – PP. 61 - 64

3. A.V.Pankin, S.V.Saitov, Ya.A.Ivakin. Directions and Methods of Development of the Functional System of the Navy Electronic Workflow //Sea Collection, no. 8, 2004. – PP. 31-33
4. A.V.Pankin. Integration of the Functional System of Electronic Workflow into Geoinformational Systems IX St Petersburg International Conference ‘Regional Informatics – 2004’, St Petersburg, 22 - 24 June 2004.
5. Pankin, V. Popovich, S. Potapichev, R. Sorokin Intelligent GIS for monitoring systems development. CORP 2005, February 22 - 25 2005, University of Technology Vienna, Austria.
6. Pankin. Integration of Heterogeneous Information Flows Circulating in the Control Loop. IF & GIS 2005, September 25 - 27 2005, St. Petersburg House of Scientists, St. Peterburg.
7. Pankin, V. Popovich, Y. Ivakin. Data for GIS. CORP2006, February 13 th – 16 th 2006. Congress Center Vienna, Austria.

3D Graphics Applied to Maritime Safety

Christopher Gold and Rafal Goralski

University of Glamorgan, School of Computing
CF37 1DL Pontypridd, Wales UK
cmgold@glam.ac.uk, rigorals@glam.ac.uk
Web: www.voronoi.com

Abstract. An important aspect of maritime safety involves the visualisation of the situation close to any particular ship, so that appropriate action may be taken. While many features may be seen directly from the bridge, bad lighting and weather may obscure them, and much of the necessary information is only available on charts or in pilot books. It seems clear that an integration of all these within a single simple view would improve maritime safety. We have developed a visualization system, based on a laptop PC, that gives a 3D-games type navigational view. We believe that the resulting 3D model provides new facilities for ship navigation, and may be introduced very economically for many types of shipping. The combination of basic GIS concepts with 3D modelling, user interaction and real-time data could contribute significantly to future maritime safety.

Keywords: Marine GIS, Maritime Safety, Kinetic Data Structures, Voronoi Diagram, 3D Visualization, AIS, ECDIS, ENC, Marine Charts

1 Introduction

An important aspect of maritime safety involves the visualisation of the situation close to any particular ship, so that appropriate action may be taken. While many features may be seen directly from the bridge in good conditions, bad lighting and weather may obscure them, and much of the necessary information concerns maritime control, and is only available on charts or in pilot books.

It seems clear that an integration of all these components within a single simple view would improve maritime safety – although it is not obvious how best this should be done. While a heads-up display – a “Mobile Augmented Virtual Reality” – might be ideal, it is not yet practical for most shipping. However, a sufficiently-good integrated computer display,

if it could mimic the real view from the bridge, could be developed at moderate cost and should be sufficiently straightforward to match with the real world while adding those extra features necessary for navigation.

Several components are required for such a system. Firstly, an appropriate 3D graphics view of the surrounding land- and sea-scape is needed. The main tools for this are readily available: modern 3D graphics cards, languages and algorithms, developed for the games or engineering industries, are readily available, and a small laptop PC may now be sufficiently powerful. Existing games engines, however, do not usually provide the level of topological connectivity needed for navigation and collision detection, nor the desired link to a cartographic database. We therefore needed to write our own, using a basic scene-graph structure.

Secondly, we needed to populate this engine with the terrain, landscape, buildings and ships appropriate to the geographic location. This can be a significant expense, especially in the case of harbour simulators, where precise building forms and textures are needed. In our less-demanding open-sea context the biggest concern was augmenting the terrain model of the coastline, as standard charts do not provide enough depth of topography to simulate the coastal silhouette, which is a necessary part of registering the computer image with the real-world view. So far we have not attempted to model coastal buildings in general.

Thirdly, bathymetry was modelled directly from samples of survey soundings, giving a triangulated terrain model of the sea floor. This is not normally visible on-board ship, and it is not directly shown on the computer view – partly because it has no legal validity for navigation, and partly because it often conflicts with the generalized chart information. It is of particular value because the terrain model may be updated in real time, taking account of the latest data. Its primary use is for the real-time interpolation of the depth of water below each of the ships in the display – especially the “active” ship with the current view from the bridge.

However, as a fourth component we needed to be able to insert specific navigational objects into our model – examples are buoys and lighthouses. These (like ship models) are developed separately with modelling software and placed at the appropriate locations on the land- or sea-scape. The OpenGL graphics language permits the modelling of cameras (views) and lights in the same way as visible objects. This, in conjunction with the scene graph model of transformation hierarchies, permitted the insertion of lights within lighthouses, cameras on the bridge, and even the simulation of lighthouse beams. An additional aspect was the integration of a simple database for instances of these objects, so that a simple mouse click could request the properties of the selected object. This was appropriate for ship, lighthouse and chart information, among others.

Fifthly, chart information was obtained from ECDIS data, and new techniques were needed in order to view this 2D information in a 3D display. Rights-of-way, anchorages, restricted zones etc. are drawn on the sea surface, and this appears to be comprehensible and effective. More problematical is the display of chart contours – while these can be drawn on the surface they obviously refer to the sea-floor terrain model, but as this model is derived directly from soundings the chart contours will rarely match the model precisely, and are taken as indicative of navigation warnings. We therefore leave them drawn on the sea surface – but with the option of draping “curtains” from them to the sea floor, reminiscent of “Northern Lights.” Especially for chart “safety contours” these are highly effective signals.

Sixthly, it is quite feasible to provide updates to the model, via radio, similar to the “Notice to Mariners.” The most obvious cases involve the warning of wrecks, obstacles, etc. At present the shipboard AIS data is collected and used to give the location and characteristics of vessels within range of the system. A preliminary interface between the Marine GIS and the AIS has been developed [7], in collaboration with the Ecole navale, Brest, France. We are also attempting to include their Tractable Role-based Agent for Navigation Systems – TRANS – intelligent navigation system that helps involved parties to negotiate the rules of way automatically in real-time [8], to propose automated collision avoidance procedures.

Stages One to Five above were implemented in prototype form in the work of Gold et al. (2004), although extensive revisions are being performed. In Stage Six the AIS integration is currently being tested. Other data inputs are being evaluated. We believe that there is a need for a small on-board 3D visualization system, using standard chart data, to assist many classes of sailors with their navigation.

2 Marine Cartography – From 2D to 3D?

Traditional marine charts have a long-standing and effective symbolism, developed over a long period of time for demanding applications, where misinterpretation can be serious. However, now that less traditional forms of display, such as 3D views and dynamic simulation are starting to be developed, some of the cartographic methods need to be re-examined. Recent collaboration with the Ecole navale, Brest, illustrated the point: a six-screen marine simulator was erected in front of a mock-up of a ship’s bridge. Observations were taken by trainee seamen, using traditional

sighting equipment to view marine markers on the screens, the features were identified on the paper chart and their directions were plotted to give the ship's location. As required the ship's course was changed, the simulator adjusted, and the simulation proceeded. It is clear that in the future these two components – screen and paper chart – will need to be integrated. However, chart symbolism may not transfer effectively to a 3D view.

Perhaps the ultimate example of a graphics-free description of a simulated Real World is the Pilot Book, prepared according to international hydrographical standards to define navigation procedures for manoeuvring in major ports [6]. It is entirely text-based, and includes descriptions of shipping channels, anchorages, obstacles, buoys, lighthouses and other aids to navigation. While a local pilot would be familiar with much of it, a foreign navigator would have to study it carefully before arrival. In many places the approach would vary depending upon the state of the tides and currents. It was suggested that a 3D visualization would be an advantage in planning the harbour entry. Ford in [2] demonstrated the idea of 3D navigational charts and concluded that 3D visualization of chart data had the potential to be an information decision support tool for reducing vessel navigational risks. Our intention was to adapt IHO S-57 Standard Electronic Navigation Charts [4, 5] for 3D visualization. The study area was Hong Kong's East Lamma Channel.

Traditional land-based GIS is two-dimensional and static. The new game technology is 3D and dynamic. We have attempted in [3] to develop a "games engine" for 3D dynamic GIS, and to evaluate the process with several applications – primarily a "Marine GIS", where objects and viewpoints move, and where a realistic simulation of the real-world view should be an advantage in navigation and training. While computer game developers have spent considerable time in developing imaginary worlds, and the ways in which players may interact in a "natural" fashion, graphics software and hardware developers have provided most of the necessary tools. However, there has been limited examination of the potential of these technologies for "Real World" spatial data handling.

For "GIS", we presume that our data represents some sampling of the real world. This real world data maps directly to the simulated world (objects and coordinates) of the computer representation that we wish to work with. The only difference between this activity and some kinds of computer game playing is the meaning inherent in our data, and the set of operations that we perform on it. Our perception or mental ability to manipulate and imagine the simulated world that we work with is usually limited by our concept of permissible interactions. Historically, we started with paper and map overlays, and this led to a particular set of techniques that

were implemented on a computer. Game development is capable of freeing us from many of these constraints. The intention of this work was to bring these concepts, and the improved algorithms and hardware that they have inspired, into the context of geographic space, where the objects and locations correspond in some way to the physical world. We then need to adapt our cartographic symbolism to this model.

2.1 The 3D Graphics Engine

Stage One, the development of our graphics system, is based on “Geo-Scene”, which is the Graphic Object Tree, or scene graph. This manages the spatial (coordinate) relationships between graphic objects. These graphic objects may be drawable (such as houses, boats and triangulated surfaces) or non-drawable (cameras and lights). The basis of the tree is that objects can be arranged in a hierarchy, with geometric transformations expressing the position and orientation of an object with respect to its parent object – for example the position of a wheel on a car, a light in a lighthouse, or a camera on a boat. This was described in Gold et al., (2004). Our recent work was to take the virtual world manipulation tools already developed, and add those features that are specific to marine navigation.

2.2 Terrain, Buildings, Objects

In Stage Two we had to develop a landscape model close to the coast, in order to permit reasonable coastline silhouettes [1], and in Stage Three we added the bathymetry based on samples of the survey soundings. These steps involved significant manual work, as the contours supplied with the charts were insufficient and the level of detail of the terrain changed at the shoreline to the sparser depth soundings. Shoreline points were calculated from the intersection of the triangulated terrain model with the sea surface, which may be changed at any time to simulate tides.

These shoreline points were incorporated within a kinetic Voronoi diagram layer, expressing the neighbourhood relations on the sea surface, and this was used for collision detection by adding the real-time ship locations. Maintenance of the adjacency relations during ship movement provided a good mechanism for detecting local collisions and planning avoidance strategies. The kinetic Voronoi diagram is illustrated on Fig. 1.

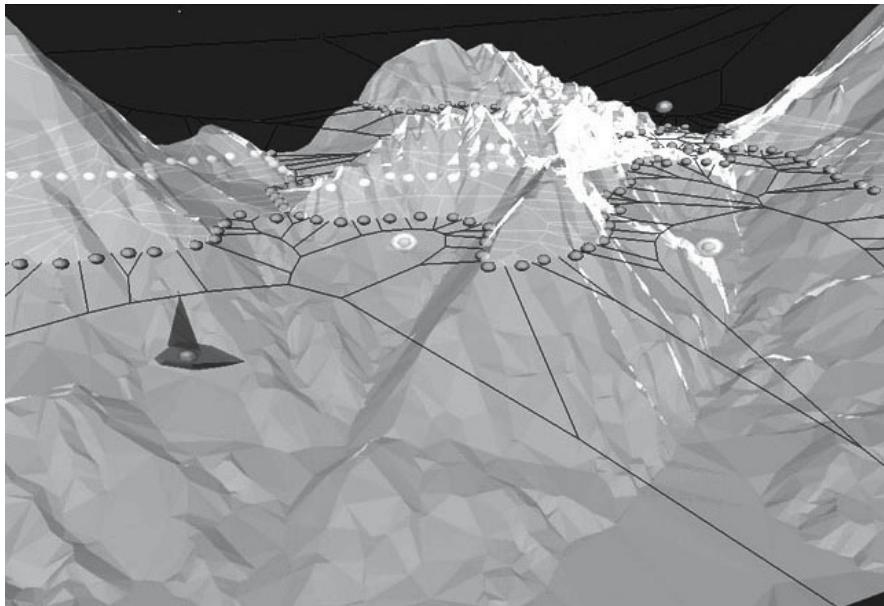


Fig. 1. The kinetic Voronoi diagram for collision detection

3D models, of ships, buoys, lighthouses etc. were created externally and instances were introduced at the desired locations. These locations could be maintained in real time based on incoming information from the AIS.

The graphics system allowed interaction by the user in a variety of ways, including the selection and querying of individual objects such as ships and the modification of the view direction and ship's course.

2.3 IHO Data Display

Marine features identified in the IHO S57 standards were incorporated. Fig. 2 shows the selection of particular S57 data items for display. These included navigational buoys, navigational lights, soundings, depth contours, anchorage areas, pilot boarding stations, radio calling-in points, mooring buoys, traffic separation scheme lanes, traffic scheme boundaries, traffic separation lines, precautionary areas, fairways, restricted areas, wrecks, and underwater rocks. Ship models, sea area labels, range rings and bearing lines, as well as oil spill trajectory simulation results were also added. Selecting a ship permits the viewpoint to be changed to that ship, permitting an evaluation of the point of view of the oncoming traffic. Other selected objects may be queried for their attributes – e.g. lighthouse and buoy locations and flashing timings, oncoming ship descriptions, etc.

3 Marine GIS – Overview

The Marine GIS as a navigation aid is meant for use on board of marine vessels using a PC-laptop integrated with on-board equipment (GPS, AIS, ARPA) through the standard NMEA interface. Different features of the Marine GIS developed during the stages described in the previous section are shown on Figs. 2÷10.

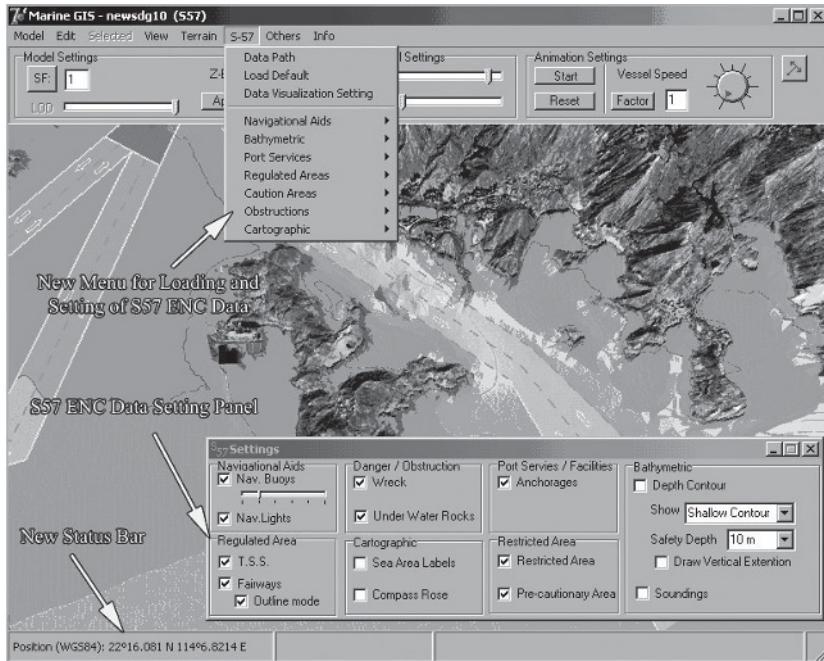


Fig. 2. S-57 data menu and S-57 settings dialogue

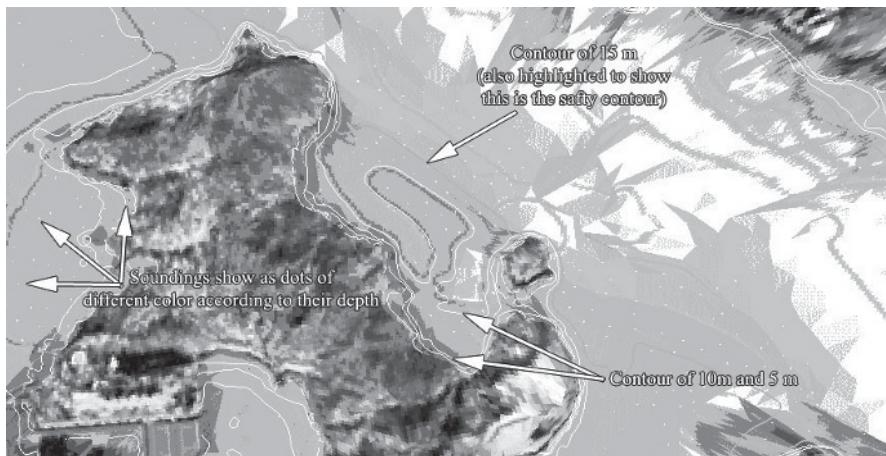


Fig. 3. Visualization of soundings and depth contour

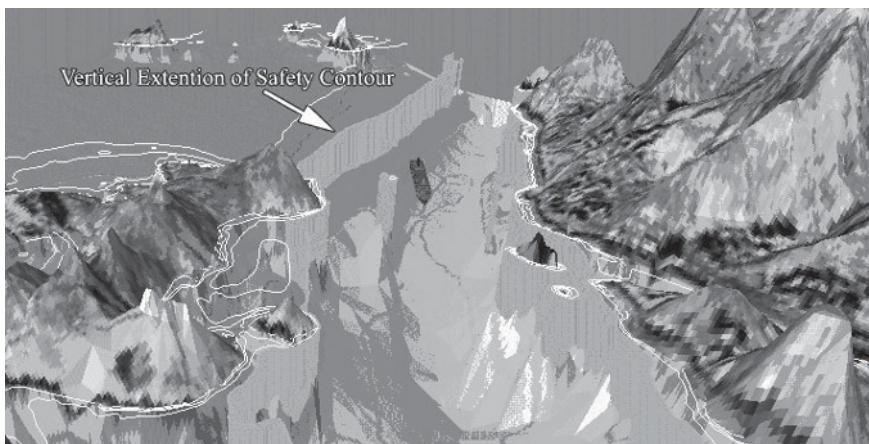


Fig. 4. Visualization of safety contour with vertical extension

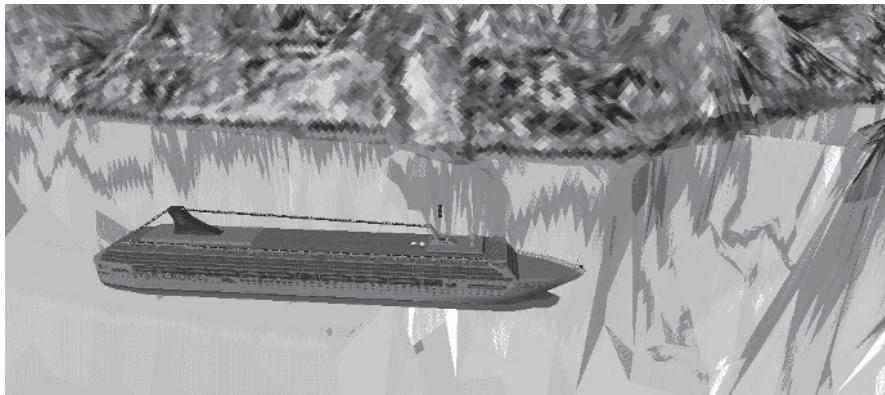


Fig 5. Visualization of boat using 3DS model

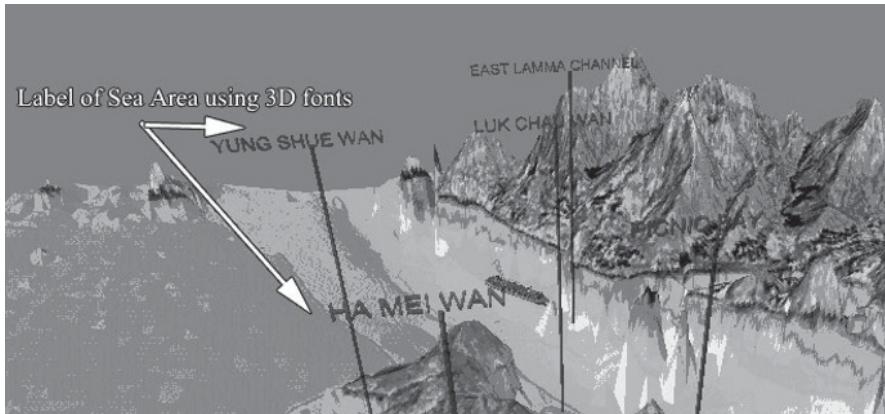


Fig. 6. Visualization of sea area label using 3D fonts

Safety contours may be displayed along the fairways, and a 3D curtain display emphasizes the safe channel. Fog and night settings may be specified, to indicate the visibility of various lights and buoys under those conditions. Safety contours and control markers may appear illuminated if desired, to aid in the navigation. The result is a functional 3D chart capable of giving a realistic view of the navigation hazards and regulations.

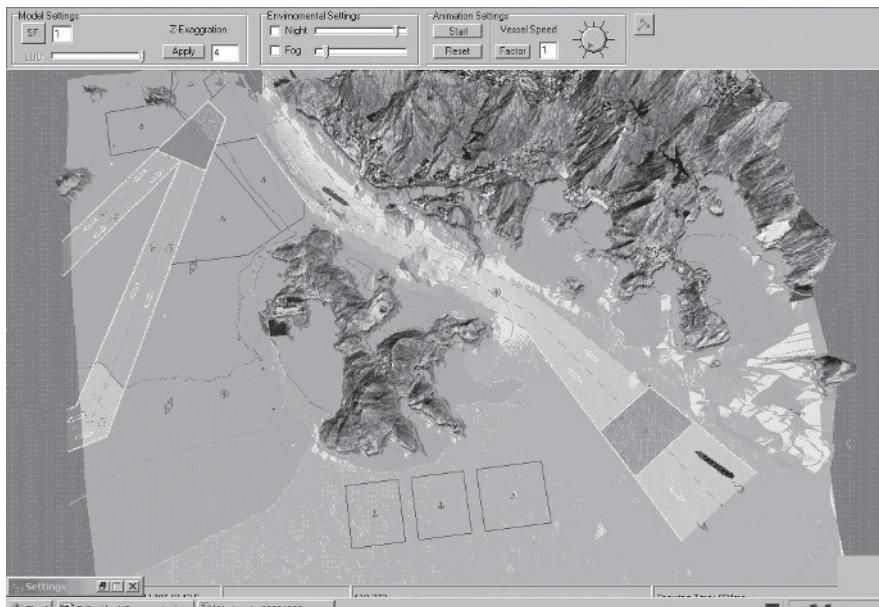


Fig. 7. Scene showing the terrain, traffic separation scheme, anchorage area, etc.

While the preliminary results appear to be satisfactory, several questions remain. Bathymetric contours, for example, are drawn on the sea surface, although they refer to the sea floor. Safety contours, drawn as curtains, are visually effective, but they correspond to the chart contours, not to the terrain model being used (otherwise each one would meet the sea floor at the advertised depth). Vertical exaggeration of the terrain is user specified, but may be unreasonable, and makes the draping of imagery difficult to achieve with good visual effect. Vertical 3D place names could be improved to look less like supermarket labels.

To tackle these problems in the next stage terrain model will be generated directly from the ENC marine chart data. The vertical exaggeration will be adjusted automatically with limited sensible range of manual adjustment.

Currently there are no buildings on the terrain model. The source of data and the techniques appropriate for incorporation of these and other landmarks on the sea shore for heads-up display are currently being attempted. Traffic separation lanes, including night-time illumination, appear at the moment to be effective, as does the ability to perform point-and-click querying of marine features.

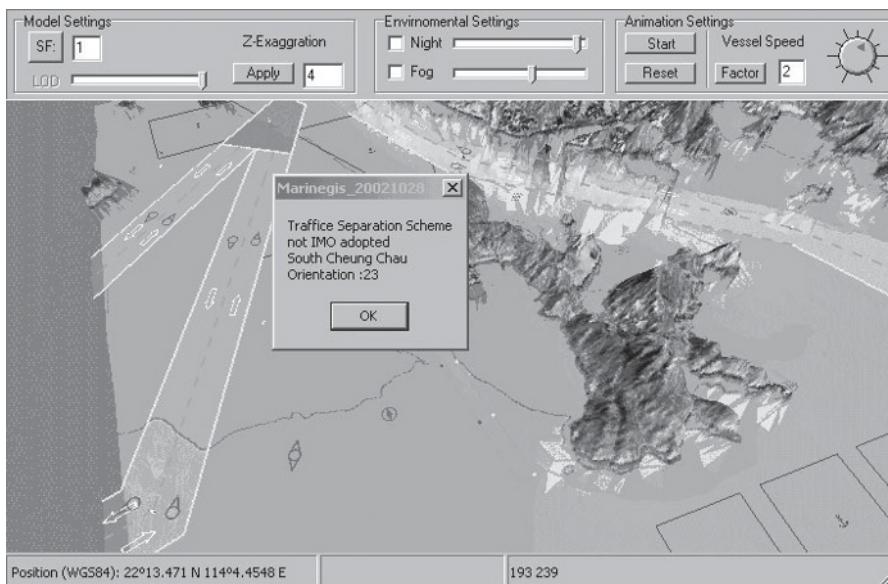


Fig. 8. Scene showing the query result of a traffic separation scheme. The attributes of the selected object are displayed

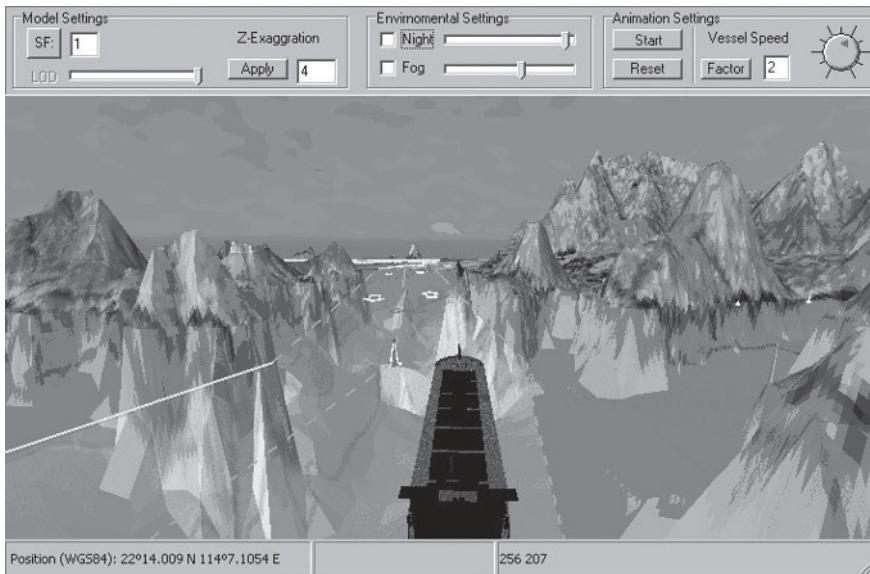


Fig. 9. Scene in navigational mode. When animation mode is activated, the viewpoint will follow the movement of the ship model

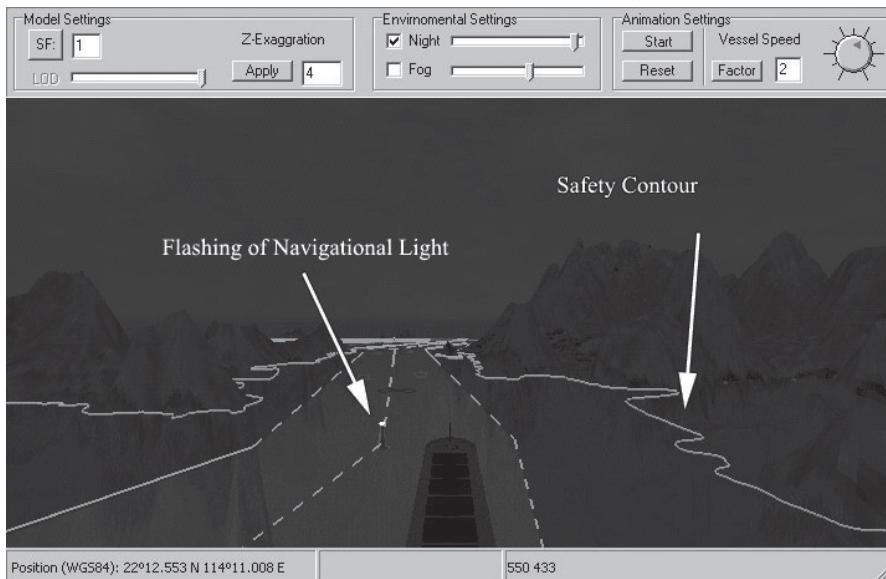


Fig. 10. Night scene with navigational lights, safety contour and traffic separation scheme

4 Maritime Safety

Marine safety is becoming more and more important with the constant growth of the sea shipping together with the growth of other boats segments. A gradually growing number of cargo ships that carry hazardous loads make possible collisions and groundings of sea vessels pose serious treat to the environment, as well as the life and health of the people and animals inhabiting coastal zones. The disaster and damage caused in the event of major sea collision can be very difficult and costly to deal with. Apart from the environment this has a huge impact on the world economy where the cost of shipping, clearly related to the level of safety, is a big factor.

Knowing about the importance of the problem the world sea authorities introduced many standards and new technologies like ECDIS and AIS [9]. However we still see a lot of points of the marine safety management process where improvements can be introduced. These include bathymetry data collection and processing, nautical charts production, providing navigation aid software on board the boats, vessel traffic planning, analysis and monitoring, as well as improving the level of training accessible for mariners and pleasure boat owners.

To address these needs we proposed a new type of special GIS software, based on kinetic 2- or 3-dimensional data structures, a sophisticated 3D visualization engine, and intelligent navigation rules libraries. In collaboration with the Ecole navale, Brest, an interface was developed between the shipboard AIS, which integrates GPS positioning together with the reception of information from other ships concerning their position, course and other properties. This involves careful multi-threading to manage the real-time inputs and processing. The results are easily used within the system, as ship locations are readily updated in the graphics model and the kinetic collision-detection tools employed.

4.1 AIS Integration

Integration of AIS transponder into the Marine GIS was performed in two steps. First an external specialized library for reading AIS messages was designed and implemented using UML modeling tools and the best object-oriented programming practices in an elastic way that allows for future expanding to other on-board NMEA devices. It was tested with a purpose-written two-dimensional graphical user interface. Then the library was incorporated into the Marine GIS 3D interface.

The software created allows for real-time tracking and recording of the AIS data, as well as for its later playback for test and simulation purposes. Several safety features related to the AIS specificity were implemented and tested. The integration with the NMEA multiplexer allowed for incorporation of the GPS data of the observer's own position. Figs. 11 and 12 present different views of AIS targets moving in the Bristol Channel recorded during field tests in Newport. Special menu options and the display of AIS target data may be seen.

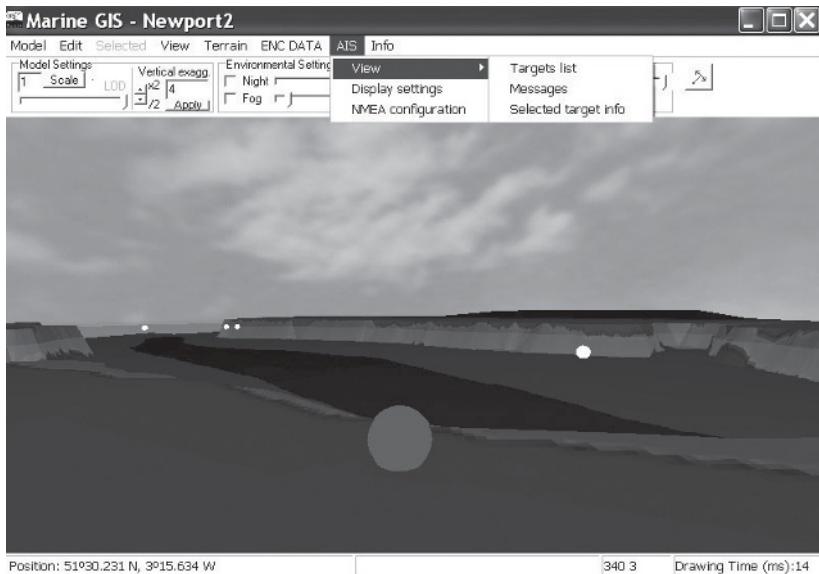


Fig. 11. AIS targets on Bristol Channel (distant white spheres) seen from the observer's point of view (the gray sphere in the forefront). The targets sizes are exaggerated for the purpose of this figure as the distance is about 10 NM

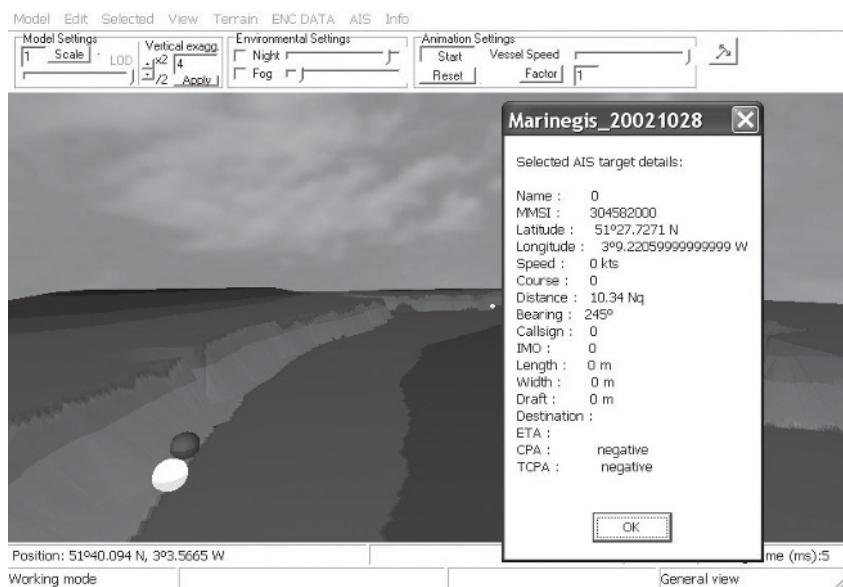


Fig. 12. AIS targets shown from a different viewpoint. All data transmitted over AIS for the selected target (dark grey sphere) are displayed

5 Conclusions and Acknowledgements

We believe that the resulting 3D model provides new facilities for visualization for ship navigation, in a form that may be provided very economically for many types of shipping. With sufficient data the screen view may easily be matched with the bridge view, allowing the easy visualization of the ship's location with respect to a variety of chart and safety data, as well as other ships. The combination of basic GIS concepts with 3D modelling, user interaction and real-time data could contribute significantly to future maritime safety.

The authors would like to thank the Ecole navale, Brest for their contribution, encouragement and support, and the EU Marie-Curie Chair in GIS of the senior author for financial assistance.

References

1. Dakowicz, M. and Gold, C. M. Extracting Meaningful Slopes from Terrain Contours. *International Journal of Computational Geometry and Applications*, 2003. Vol. 13, pp. 339-357.
2. Ford, S.F. The first three-dimensional nautical chart. In: *Under Sea with GIS*. ESRI Press: Wright, D., 2002. pp 117 -138.
3. Gold, C., Chau, M., Dzieszko, M. and Goralski, R. 3D geographic visualization: the Marine GIS. In: *Developments in Spatial Data Handling*. Springer, Berlin: Fisher, P., 2004. pp. 17-28.
4. International Hydrographic Bureau. IHO transfer standard for digital hydrographic data edition 3.0, Special publication No. 57. 2000.
5. International Hydrographic Bureau. Regulations of the IHO for internal (INT) charts and chart specification of the IHO. 2001.
6. The United Kingdom Hydrographic Office. Admiralty sailing directions – China Sea Pilot volume I, fifth edition. United Kingdom National Hydrographer, Taunton, 2001.
7. STROH, B., SCHULDT, S. Integration of an AIS transponder into the PC-based operational marine simulator 'Marine GIS' using NMEA interface/communication protocol. University of Glamorgan, 2006
8. Fournier, S., Brocarel, D., Devoge, T. and Claramunt, C. TRANS: A Tractable Role-based Agent Prototype for Concurrent Navigation Systems. European Workshop on Multi-Agent System (EUMAS), Oxford, 2003.
9. SOLAS: International Convention of the Safety of Life at Sea, 1974 - Consolidated Edition. International Maritime Organization Publishing, 2004. 92-801-4183-X.

Northern Shield

US-Russia Maritime Energy Security Cooperation

Mr. Gabe Collins

United States Naval War College

The views expressed in this paper are the author's personal analysis and opinions and in no way reflect official US government policies and positions.

1 Introduction

Russia is a major maritime energy player, producing more than 9.5 million barrels per day of oil and exporting approximately 3 million barrels/day of crude by sea. Additionally, Russia aims to become a significant liquefied natural gas (LNG) exporter. Russian energy production is becoming more maritime in character, with projects on Sakhalin Island, Shtokman, and other areas. Finally, Russia is not simply a supplier of seaborne oil and gas. It has significant oil transportation interests, and after the 2006 merger of Sovcomflot and Novoship, now boasts the world's fourth largest tanker fleet[1].

US-Russian energy cooperation at sea would be both logical and mutually beneficial. Russia is a top exporter of oil and gas and cooperation could help move US-Russia relations onto a more positive track. The US and other oil and gas importers would benefit from the stabilizing effect that expanded Russian oil and gas supplies could have on the global market. Russia would also gain from secure energy production, as the oil sector alone currently generates roughly 15 percent of Russian GDP [2]. In exploring possible avenues for maritime energy security cooperation between Russia and the US, this paper aims to:

- Illustrate Russia's large and growing maritime role in the global energy system.
- Identify threats to Russia's maritime energy exports and possible joint responses.
- Demonstrate how the US and Russian Navies, US Coast Guard, Russian Maritime Border Guards, and private sector might foster maritime energy security cooperation between the United States and Russia.

2 Russia's Growing Maritime Energy Interests

Russia already exports roughly three million barrels of oil per day by ship. Tankers take on oil at Primorsk on the Baltic Sea, Tuapse, Novorossiysk, and Odessa on the Black Sea, and at De Kastri and Prigorodnoe on Sakhalin Island in the Pacific. There are export facilities near Murmansk (based on the supertanker Belokamenka, which is used as a floating storage facility for loading other tankers) as well as at Varandei on the Arctic coast.

The Russian government hopes to see oil production reach approximately 10.5 million bbl/d by 2015 [3]. Unless internal oil demands outstrip production growth, Russia's export volumes will likely increase. Maritime outlets are vital because Russia's overland pipelines are already running near full capacity.

3 Forces Driving Russia's Seaborne Energy Trade

3.1 Economics

Maritime oil shipping is desirable from both the political and economic standpoints. It is simpler and often cheaper to build tankers and access new markets by sea than to build multibillion dollar pipelines across sovereign states who may not share Russian interests. Current Russian oil and gas export pipelines to Europe were built during the Cold War when the Soviet Union did not mind the cost of multi-thousand kilometer pipelines. Today, building new pipelines is even more expensive. In addition, it costs far more to transport a barrel of oil by pipeline than by sea. Third countries often enact steep tariffs on oil crossing their territory, cutting exporters' revenue.

The construction and transport costs of major new oil and gas pipelines will have to be considered against maritime routes, particularly for shipping oil from Russia's northern fields to Western Europe and North America. **Fig. 1** shows the cost advantages of maritime oil shipping over pipeline transport.

Distance 4500 KM (~3000 miles)	Cost per Barrel
Through Russian pipeline network*	\$3.56
By Suezmax tanker**	\$0.38
Sources: Platts Oilgram, Yukos Presentation	*Transneft tariff currently \$0.58/ton/100KM ** Assumes Suezmax tanker hauling 1 million bbl of oil @ \$50K/day@25 km/hr

Fig. 1. Tanker vs. Pipeline Transport Costs

3.2 Transit Country Risk

Russia's existing export pipelines to Europe were laid by the Soviet Union across Soviet satellite states that had no choice in the matter. Today, Russia is embroiled in increasingly frequent disputes with Ukraine, Belarus, and other neighbors across which its oil and gas pipelines must pass. As Russia comes to grips with transit countries' leverage over pipelines, maritime oil shipment will become more attractive. During a January 2007 meeting with German Chancellor Angela Merkel, President Putin noted that Russia will actively work to bypass transit countries such as Poland, Ukraine, and Belarus [4].

The January 2007 spat between Russia and Belarus over gas pricing, and the retaliatory imposition of a tariff on Russian oil exports by Belarus, led to a temporary cutoff of exports through the Russian Druzhba pipeline and triggered wide condemnation of Russian actions. Transneft is actively investigating the possibility of increasing the Baltic Pipeline System's capacity by 160 percent (to 120 million tons of crude/year), so that it can avoid having energy exports held hostage by a transit country in the future [5]. If this plan succeeds, tanker traffic in the Baltic could rise significantly.

Russia has also had problems with the Bosphorus, which suffers from congestion and strict transit rules formulated by Turkey. These factors have thus far prevented Russia from substantially expanding its Black Sea oil shipment volumes. However, Transneft, Rosneft, and GazpromNeft have established a consortium with Greece and Bulgaria to build a 700,000 bbl/day pipeline from Burgas, Bulgaria to Alexandropolis, Greece that would by-pass the Bosphorus and allow increased Russian and Caspian oil via Black Sea ports [6].

3.3 More Market Choices

Maritime energy shipping gives Russian exporters more flexibility in choosing markets. Pipelines tie consumers and producers together and erode both sides' ability to deal with sudden changes in energy prices and demand. A pipeline delivers oil from one fixed point to another. A tanker at sea can be rerouted and the cargo resold, giving great flexibility to both importers and exporters of oil. Certain traditional energy markets in Eastern and Central Europe can only be accessed by pipeline, but major new demand centers for Russian energy in Asia and the US are accessible only by sea.

3.4 Growing Products Exports

New tanker export growth will not come only from crude oil. Russian oil products exports stand to grow substantially because Russian firms currently leave money on the table by not having large refinery operations.¹ Experts estimate that state oil producer Rosneft currently foregoes at least \$2 billion/year in income because it can only refine a fraction of its oil production [7]. To remedy the situation, Russian state and privately owned firms are expanding domestic refining capacity and looking to acquire refining assets in Asia and Europe. Rosneft plans to boost its refining capacity to 150 million tons of crude per year [8].

The refineries that Rosneft and other companies hope to acquire or build are located in Greece, Great Britain, the Netherlands, China, and the Russian Far East. If even half of these deals materialize, Russian firms could gain 500,000 or more barrels/day of refining capacity [9]. Traditional Russian refineries lie far inland. Virtually all of the new ventures are seaside and well situated for tanker exports of products and tanker imports of crude oil feedstock.

The highest levels in the Kremlin appear to support increased maritime oil and products trade. In October 2006, Russian Prime Minister Mikhail Fradkov noted the lack of pipelines to key ports and encouraged the development of oil product export infrastructure so that Russia can boost its seaborne product exports [10]. **Figure 2** compares current Russian seaborne energy exports to expected export levels in 2015.

¹ The sum of the refined components made from a barrel of oil typically exceeds that of the crude oil's original value. Thus, companies with significant refining operations can capture more rent from each barrel, particularly when products are in high demand. Expanded refining operations would also create jobs within Russia.

	2006	2015	Growth
Crude Oil	~3 million bbl/day	More than 4 million bbl/day ²	~35%
LNG	0	~20 MTA	N/A

Note: As Russian oil producers expand their refining capacities and look to acquire overseas refineries, Russian oil product exports could also expand substantially in coming years

Fig. 2. Russian Seaborne Energy Export Growth

3.5 LNG Production Growth

While Russia is already a large maritime oil supplier, any discussion of seaborne energy transport remains incomplete without mentioning LNG. Russia is traditionally a pipeline gas exporter because her primary fields lie far inland in Western Siberia. Although the bulk of Russia's gas exports will continue flowing through pipelines, the country is nevertheless poised to become a significant LNG exporter. The global LNG market—driven by increased demand for clean fuel—has grown substantially over the past decade. Many of Russia's newest gas fields are near the sea and well situated for LNG exports. **Figure 3** shows the location, estimated start up dates, and projected production for Russian LNG projects.

Project	Estimated Start-up Date	Projected capacity (Million tons/year)
Sakhalin-2	2008	9.6 MTA current, 16 MTA under consideration
Baltic LNG	~2010	7 MTA
Yamal (Arctic)	2010 or later	7.3 MTA
Shtokman (Arctic)	2015	Initial 5 MTA, possible expansion to 15-20 MTA
Total Russian LNG Production Circa 2015= ~20 MTA		
Sources: Gazprom reports, World Gas Intelligence		

Fig. 3. Russian LNG Projects

² Assumes current seaborne crude shipment levels plus 25 million tons/year of Sakhalin oil production plus a 25 MTA expansion of the Baltic Pipeline System due to problems with transit countries plus 250,000 bbl/day of Arctic shelf oil production.

Gazprom is slated to be Russia's largest LNG exporter and is already positioning itself to have a global LNG presence. Company managers frequently state their desire to become a "serious player" in the North American LNG market. The firm has followed this up by locating its global LNG trading and marketing office in Houston, and by looking to acquire natural gas distribution infrastructure in North America [11]. Although Gazprom won't have its own LNG production until at least 2008, it has traded swap cargoes purchased from British Gas and other producers in order to establish a market foothold and gain trading and marketing experience.

3.6 "Seaside" Location of New Russian Oil/Gas Production

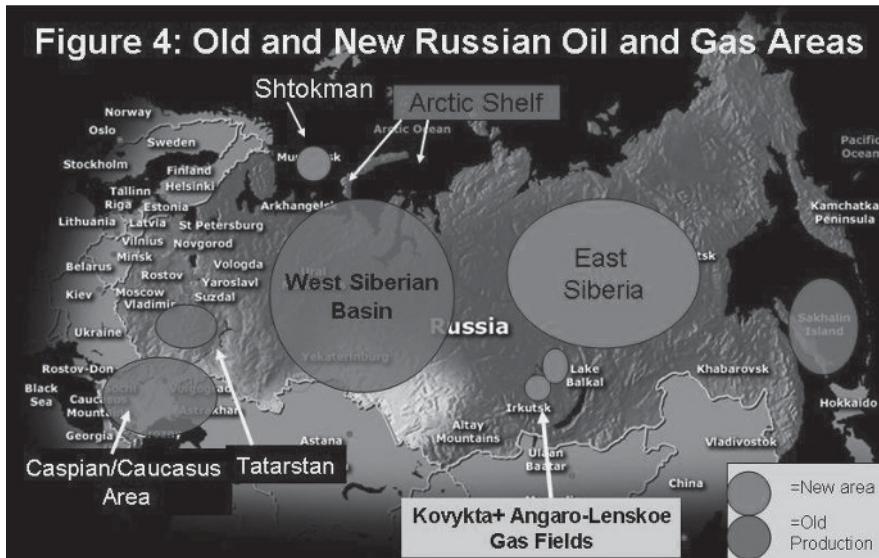
New Russian oil and gas developments are increasingly shifting from the Western Siberian "core" to areas that are either offshore or lie close to the coast, making maritime transport the most cost effective way to bring production to market. New areas already in production include Sakhalin Island, whose oil has already been shipped to India and other markets, as well as the Timan-Pechora area, where LUKOIL's Varandei terminal may ship 240,000 barrels/day of crude by the end of 2007 [12].

Several projects seem poised to join the maritime energy ranks in the near future. The Sakhalin-II LNG facility will be operational by 2008. Eastern Siberian developments will also send oil to a terminal near the Pacific port of Nakhodka, although it remains unclear whether this oil will come by ship or rail, and, exactly how many barrels per day the terminal will send out. Meanwhile, the ultimate destination of gas from the giant Shtokman field remains unclear. Over the past 5 years, Shtokman has been successively earmarked for pipeline sales to Europe, re-tagged as an LNG supplier for the US market, and re-prioritized for pipeline gas sales to Europe [13]. Whatever the outcome, Shtokman will likely supply LNG (although this may be 10 years or more into the future) and will be a true offshore production project.

Lastly, developments in the Arctic and Caspian regions will be inherently maritime in character.³ The Russian Ministry of Natural Resources predicts that by 2010, the Russian continental shelf (mainly in the Arctic) could be producing 200,000 barrels/day of oil, and up to 1.9 million barrels per day by 2020 [14]. The Ministry also estimates that by 2010, the shelf could produce enough natural gas to make nearly 20 million tons/year of

³ Russia has recently expressed a desire to become a major player in the Caspian shuttle tanker trade and also has production interests in the Caspian's Russian sector.

LNG, and that by 2020, there could be enough to produce more than 100 million tons of LNG [15]. Not all gas fields will be large enough to justify construction of liquefaction facilities, but the predicted magnitude of production suggests just how maritime-dependent future Russian oil and gas production may become. **Figure 4** shows “old” and “new” Russian oil and gas production areas.



The Russian oil and gas sector’s shift toward a more maritime orientation will create a significant need for protection of production facilities, export terminals, and shipping. The next section of this article will discuss key maritime threats to the energy industry, as well as ways to tackle them.

4 Threats to Maritime Energy Security

Seaborne energy transport is a critical link in the global oil and gas supply chains. More than 1/3 of the oil traded internationally crosses the oceans in the holds of nearly 500 Very Large Crude Carriers (VLCCs) and thousands of smaller tankers. Three quarters of this oil passes through 16 key global chokepoints that could be blocked by armed conflict, terrorist attacks, or piracy. Secure maritime oil transport is critical to oil consumers, as well as to producers, who would suffer significant losses from disrupted seaborne oil trade. The increasingly tight global oil market, rise of unconventional threats such as terrorism, and major consumers’ and producers’

renewed emphasis on overall energy security make energy security's maritime dimension an urgent concern.

Meanwhile, the global LNG market has grown at nearly 8 percent annually over the past decade and more than 200 tankers now transport 157 million tons of LNG per year [16]. LNG and crude oil shipments pass through many of the same chokepoints, and face common threats. **Fig. 5** shows the crude oil and LNG supply chains and their respective vulnerabilities.

Figure 5: The Oil and LNG Supply Chains



5 Responses

As shown in Figure 5, Russian seaborne energy shipments face several key threats. While there is no magic solution to this diverse range of threats, the Russian and US Navies, Russian and US Coast Guards, and private sector firms can share best practices and work to promote mutually beneficial US-Russia maritime energy security cooperation.

5.1 US Coast Guard-Russian Maritime Border Guard Cooperation

Navy-navy cooperation is a critical component of maritime energy security. However, port and critical facilities protection is typically a Coast Guard responsibility. Thus, we will first discuss the possibility of maritime energy security cooperation between the US Coast Guard and Russian Maritime Border Guards.

Russia faces substantial security threats in the energy arena. Moscow grapples with an insurgency in Chechnya and general instability in the North Caucasus not far from its major Black Sea oil loading ports. It is conceivable that a non-state group wishing to harm Russia might consider attacking key energy export infrastructure. Chechen terrorists have already demonstrated the capacity to strike deep into Russia. Notable attacks include the Beslan massacre in 2004, the 2002 Dubrovka theater hostage taking (“Nord-Ost”), and the 1996 raid on Budyonnovsk. Such an enemy might consider attacking major energy export facilities.

Russian oil and gas infrastructure is spread over a wider geographical area and would be tougher to secure than facilities in countries whose main fields are centralized. In addition, because Russia’s core oil production base presently lies far from tanker loading terminals, the country must take a comprehensive approach to oil and gas supply security. Russia must protect tankers at sea, secure loading terminals, and safeguard the production facilities and long pipelines that bring the oil to the seashore. Russian companies are moving to enhance onshore energy infrastructure security, as Gazprom and Transneft are both working to establish armed security detachments that would be responsible for protecting production, processing, storage, and transport infrastructure [17]. Yet there remains a need to secure the sea zones around key terminals, producing fields, and export routes in the near shore waters most easily reachable by terrorist actors.

Protecting key seaside energy infrastructure is an area where the US might share knowledge and experience and collaborate with Russia. The United States Coast Guard (USCG) has substantial experience protecting LNG and tanker loading facilities. In the wake of the 9-11 attacks, the USCG stepped up its port security operations and learned how to foster public/private partnerships, helping it secure some of the world’s busiest energy ports, most notably the Port of Houston [18]. This experience can be shared with the Russian security services as they work to formulate the most effective facilities protection plans. USCG personnel would also have the opportunity to learn from their Russian counterparts, who bring their own unique operational experiences to the table.

The Coast Guard's energy security experience spans the globe. USCG units have successfully protected Iraq's Persian Gulf oil export facilities for more than 3 years, developing a wealth of experience and combat tested techniques that could be usefully shared with Russian security forces. The USCG's maritime security response teams (MSTs) and maritime safety and security teams (MSSTs) are trained to operate in high threat areas and deploy anywhere in the United States within 12 hours [19]. If the USCG helped create Russian Maritime Border Guard and FSB teams with similar capabilities, this would enhance oil and gas export security.

Maritime energy security cooperation can build on the precedent set by the 6-year old North Pacific Coast Guard Forum, which brings Japan, Russia, Korea, China, Canada, and the US together to cooperate on regional maritime issues. Russia will hold the Forum's annually rotating presidency in 2007. This presents an ideal moment to begin discussions on bi and multi-lateral maritime energy security collaboration. Existing fisheries enforcement agreements between the US and the Russian Maritime Border Guards may provide a starting point for energy security cooperation in the North Pacific [20].

Coast Guard to Coast Guard cooperation would ideally include oil spill contingency planning and preparation of mitigation measures, since several new energy producing areas and shipping routes lie in fragile Arctic ecosystems and other sensitive zones such as the Baltic Sea and Sakhalin Island. Fisheries are a rich Russian natural resource and a joint oil spill response plan benefits both nations. Along with an oil spill response plan, both parties might also consider establishing joint iceberg and tanker tracking centers in order to further boost maritime safety on the Arctic, Baltic, and Black Sea energy shipping lanes.

5.2 Naval Cooperation

Cooperation between the U.S. Coast Guard and the Russian Maritime Border Guards will be very important. However, both are already heavily taxed by the operational and funding requirements of their current missions. In addition, the Russian and US Navies are better configured to handle portions of the maritime energy protection mission that might require operations beyond the 200 nautical mile exclusive economic zone.

The United States Navy is heavily occupied with the Persian Gulf, Indian Ocean, and Western Pacific. If the Russian Navy can independently secure oil shipments along key Arctic and North Atlantic sea lanes, this reduces the US global burden and both Russia and the US win. Such coop-

eration would involve bilateral action, as well as deeper cooperation between Russia and NATO on naval energy security matters.

The Russian Navy has significant capabilities for maritime energy security related missions. Moreover, the Kremlin appears to be re-emphasizing naval development. The new Project 20380 corvette marks the Russian Navy's first new naval shipbuilding project since the end of communism [21]. Protection of key maritime economic interests may be a possible new *raison d'être* for the Russian Navy. Russian Navy Commander Admiral Vladimir Masorin has stated that the new Project 20380 corvettes will help protect oil and gas transportation routes, particularly in the Black and Baltic Seas, where fleet assets currently lag behind the significant requirements that increased energy transit will create [22]. Russian naval forces can also help safeguard energy supply routes in the Arctic, North Pacific, and Caspian Sea areas.

It will be particularly critical to integrate US and Russian maritime energy security efforts in the Caspian Sea. The Caspian Sea and the Black Sea face the highest risk from terrorist attack of any major Russian energy transport routes. They also have the highest potential for becoming geopolitical battlegrounds between Russia and the US. Russia and the US have clearly demonstrated their intent on influencing events in the region. In 1997, the US began the CENTRAZBAT joint exercises with forces from Kazakhstan, Uzbekistan, and Kyrgyzstan [23]. It has since maintained military relationships with these countries, as well as Azerbaijan. In 2002, Russia held major live fire exercises following failed seabed demarcation talks between it and other Caspian littoral states. Moreover, the US has established the energy security focused Caspian Guard, which does not include Russia, while Russia has attempted to create a similar Caspian security bloc excluding the US [24].

US-Russian competition in the Caspian region is counterproductive and undermines both countries' maritime energy security plans. Fortunately, there are existing models for US-Russia cooperation in sensitive zones that might help pave the way for increased cooperation in the Caspian area. The US recently created an integrated defense strategy for the Black Sea that encourages regional states to join Turkey's "Black Sea Harmony" maritime security initiative, which includes Russia [25]. If Moscow and Washington can find a similar way forward in the Caspian Sea, both will gain.

The energy protection mission makes sense for several reasons. First, it is a tightly focused mission that is primarily directed against non state threats. As such, it will enjoy broad international support. Second, gaining proficiency in littoral energy security operations will enhance Russian interoperability with NATO and other international partners in a wide

range of other key maritime security missions such as drug interdiction and immigration enforcement. Third, focusing on energy protection would reflect the realization that most modern maritime security threats (i.e. piracy and terrorism) are littoral in nature and are best dealt with by dedicated multipurpose platforms such as the Project 20380 corvette.

Modern maritime security should emphasize making international forces complementary with respect to capabilities and operational expertise. Combining American global operational abilities with Russia's regional capacities and excellent new platforms fully embodies this idea. Today's Russian Navy is fully capable of executing energy protection missions within the Arctic, European, Black Sea, Caspian Sea, and North Pacific regions.

The Russian Navy already exercises with NATO forces in the Black and Mediterranean Seas and appears to be interested in a broader maritime security partnership. For both diplomatic and operational reasons, expanding current NATO-Russia activities might be the most effective way forward US-Russia naval energy security operations.

5.3 Private Sector Cooperation

Many private US firms have substantial energy security expertise and advanced technologies, making them logical implementers of critical energy infrastructure protection measures. Indeed, 95 percent of American pipeline operators have already implemented security plans and could exchange ideas and best practices with their Russian partners, many of whom also have significant facilities protection experience.⁴

US energy construction and engineering firms would be excellent partners for Russian firms looking to build new facilities and upgrade existing infrastructure. Moscow and Washington might facilitate this by establishing a US-Russia Energy Infrastructure Investment Working Group. The working group would facilitate collaboration on badly needed energy infrastructure construction and renovation projects including oil and gas pipelines, refineries, nuclear and conventional power plants, the electrical

⁴ It is also conceivable that US based private military firms (PMFs) could advise Russian internal security forces on facility protection techniques. American PMFs have trained internal security forces elsewhere in the world, notably in Saudi Arabia, where DynCorp and others have successfully trained portions of the Saudi National Guard, whose mission includes energy facility protection. It might also be politically more palatable for Russian forces to accept training from a PMF paid by the Russian government than to accept US aid.

grid, and LNG facilities. Projects would be collaborative and “turnkey” in nature, so as to assuage Russians’ fear of foreigners controlling key portions of their energy system.

6 Conclusion

Reversing the current negative trend in US-Russia relations requires a serious effort by both sides. As a first step, maritime energy security cooperation offers realistic and mutually beneficial ways to build the partnership. Consumer and producer interests naturally mesh when it comes to making sure exporters can securely deliver oil and gas to market.⁵ Both nations gain from useful exchanges that foster increased trust and understanding.

As Russia’s already significant maritime energy interests continue to grow, the importance of physical energy security will grow for both the Kremlin and energy consumers. While the biggest seaborne energy shipment growth lies in the future, it is critical to establish the necessary diplomatic and operational infrastructure to handle issues that may arise. A proactive approach helps avoid crisis driven decision making.

Vigorous US engagement on maritime energy security can be a new cornerstone of broader US-Russia cooperation. If the two countries can develop the trust to cooperate in an area as critical as securing energy supplies, the realm of possibilities for future engagement will grow substantially. Now is the time to begin this vital effort.

References

1. Frank, Jerry. “Russian Giant Tanker Tie Up Becomes Reality: Sovcomflot-Novoship Marriage Given Kremlin Blessing.” September 8, 2006. <http://web.lexis-nexis.com>
2. Khartukov, Eugene and Ellen Starostina. “Projects Focus on Pipeline, Terminal Expansions.” Oil and Gas Journal. March 27, 2006. (57-60)

⁵ Maritime energy security cooperation would reflect the reality that core US policy should emphasize bringing Russian oil and gas onto the market rather than focusing solely on shipping it to the US. The global oil market is fungible and Russian energy projects oriented toward Europe or Asia still bolster US energy security because every barrel of crude that goes to China or Japan is one less barrel that those countries must import from Angola or Saudi Arabia.

3. "Khristenko: By 2015, Oil Production Could Reach 509-542 Million Tons." Oil and Capital. October 10, 2006. <http://www.oilcapital.ru>. (in Russian)
4. "Putin: Russia's Energy Sector Will Reduce Its Reliance on Transit Countries." Oil and Capital. January 22, 2007. www.oilcapital.ru/news/2007/01/220952_104063.shtml. (in Russian)
5. "Transneft May Increase the Baltic Export Lines' Capacity in order to Bypass Belarus." Newsru.com. January 12, 2007. http://www.newsru.com/finance/12jan2007/truba_print.html (in Russian)
6. "Transneft, Rosneft, and Gazprom Neft Form a Consortium for the Burgas-Alexandropolis Project." February 1, 2007. www.newsru.com/finance/01feb2007/truba_print.html (in Russian)
7. "By Acquiring Refineries in Russia and Europe, Rosneft Aims for a 15-fold Expansion of its Oil Refining Capacity." Newsru.com Economics. October 24, 2006. http://www.newsru.com/finance/24oct2006/rosneftnpz_print.html.
8. Ibid
9. "Russia's Lukoil Interested in Shell's French Refineries." Prime-Tass Business News Agency. January 12, 2007. <http://web.lexis-nexis.com>
10. "Fradkov Calls for Development of Seaborne Oil Products Export." Oil and Capital. October 10, 2006. www.oilcapital.ru. (in Russian)
11. McElligott, Suzanne. "Russia's Gazprom to Expand Houston Office, Acquire North American Assets." Gasification News. Hart Energy Publishing LLP. December 15, 2006. <http://web.lexis-nexis.com>
12. "Growth Spurt: Lukoil Orders Tankers for Varandey Expansion." Nefte Compass. Energy Intelligence Group. September 15, 2005. <http://web.lexis-nexis.com>
13. "Endless Shtokman." Editorial. Vedomosti. October 11, 2006. <http://www.vedomosti.ru/newspaper/article.shtml?2006/10/11/113913>.
14. "Vedomosti: The Kremlin Assigns All of Russia's Shelf Fields to Gazprom and Rosneft." Newsru.com. January 22, 2007. http://www.newsru.com/finance/22jan2007/national_print.html. (in Russian)
15. Ibid.
16. "World LNG Trade Rises 12 Percent in 2006." LNG World Shipping. January 16, 2007. <http://www.lngworldshipping.com/content/news/compNews230.htm>
17. « Gazprom and Transneft Arm Themselves » Nezavisimaya Gazeta. March 2, 2007. from Oil and Capital. http://www.oilcapital.ru/print/press_round_up/2007/03/021038.shtml
18. D. Hauser. LCDR, US Navy. "Port Coordination in the Largest US Petrochemical Complex: A Public/Private Partnership." Proceedings. Spring 2006. (55-59)
19. Davenport, Aaron C. CDR, USCG. "Maritime Security and Safety Teams: A Force for Today." Proceedings. Spring 2006. (83-86)
20. Davis, John. CDR, USCG. "How International Enforcement Cooperation Deters Illegal Fishing in the North Pacific." January 2003. <http://usinfo.state.gov/journals/ites/0103/ijee/davis.htm>.

21. Mozgovoi, Alexander. « Steregushy Opens a New Page. » Military Parade. July/August 2006. (32-34).
22. « Severnaya Verf Launches Stoiky Corvette for Russian Navy. » October 11, 2006. RIA Novosti. <http://en.rian.ru/russia/20061110/55520921-print.html>
23. Butler, Kenley. “U.S. Military Cooperation with the Central Asian States.” Center for Nonproliferation Studies. September 17, 2001. <http://cns.miis.edu/research/wtc01/uscamil.htm>
24. Ismayilov, Rovshan. “Azerbaijan Ponders Russian Caspian Defense Initiative.” Eurasia Insight. February 1, 2006. http://www.eurasianet.org/departments/insight/articles/eav020106_pr.shtml
25. Kucera, Joshua. “US DoD Devises Black Sea Security Strategy.” Jane’s Navy International. February 13, 2007. <http://www8.janes.com>

Space-Extensive Targets Radar Search in the Presence of Noise

I. F. Shishkin, A. G. Sergushev

North-West State Technical Extramural University (SNTU), Saint Petersburg

Abstract. The report considers the theory of space-extensive targets search, being developed by the authors of the report. An approach, defined in the report, makes it possible to develop methods for assessing the efficiency of space-extensive targets search in the presence of noise and it is the logical development of the theory for conditions of search in the real environment. The report is based wholly on the original studies of the authors.

Keywords: Theory of search, search of space-extensive targets, efficiency of searching aids, search of traces on the sea surface, investigation of traces in water areas.

1 Introduction

One of the main features of search in the real environment consists in the fact that the search process takes place in the presence of noise. Technical aids operate in the presence of noise of natural as well as artificial origin, thus detection of an object is an accidental event. This leads to the fact that the range of technical detecting aids is characterized by random nature and, besides that, the event of the object presence within the technical aids range does not mean yet that it will be detected for sure (probability of the contact). B.Koopman, the initiator of the theory of search, was the first who noted in his papers this feature of search in the real environment, later this aspect was developed in the works of other investigators. However, initial axiomatics supposes the point nature of the target, thus all procedures giving possibility to take into account the random nature of detecting the targets due to the influence of noise on the search process were developed for point targets. Besides that, there exists another class of targets, the class of space-extensive targets, conformably to which this problem was considered by none of the investigators. This report considers search of space-extensive targets in the presence of noise.

2 General Principles

Let us assume that space-extensive targets are the targets with geometrical dimensions greatly exceeding resolution of the equipment used for their detection. These targets, mainly, include not the objects themselves, usually having small size, but their physical fields with substantial space extension. The space-extensive target is formed also by a group of objects with or even without their physical fields taken into account. It is obvious that, with other conditions being the same, search of space-extensive targets is more effective than that of point targets. This fact can be taken into account while assessing the efficiency of search with the help of the mathematical apparatus of the theory for space-extensive targets search, being developed by the authors. When developing the theory of space-extensive targets search, the authors did not take into account in the initial axiomatics the whole spectrum of natural and artificial noise, disturbing the search process. It was supposed that the search process took place under ideal conditions, i.e. without the influence of natural and artificial noise, disturbing the search process, and only the contribution of acquired effectiveness of search into gain of the search potential was taken into account. However, under real conditions search is made in the environment, in which search aids are influenced by the broad spectrum of diverse noise of natural as well as of artificial origin. These results both in decrease of search efficiency (mathematical expectation in the space searched during the time unit, due to decrease of potentially possible searched space) and in decrease of the probability of the contact (probability of correct detection with the optimum threshold). This report develops the views of the authors in the field of the theory of space-extensive targets, when they are searched in the real environment under conditions of noise of any type, disturbing the process of space-extensive target search and influencing the search aids.

3 Probability of Contact

A searched object location with the range of technical aids does not mean yet that it will be detected for sure. Technical aids operate in the presence of noise, thus detection of an object is an accidental event. The lower is some threshold level of the signal at the output of the technical aids (exceeding of this level is put into correspondence with the fact of an object detection), the higher is the probability of correct detection. Decreasing of the threshold level is followed by increasing of false alarm probability,

thus the threshold level possesses some optimum value that is to be set. The probability of correct detection at optimum threshold level is termed as the probability of contact.

Let us denote the density of distribution of signal probability at the output of the optimum filter of a searching aids receiver with the response to the desired input signal by $p_1(W)$ and with no response by $p_0(W)$. When the signal $W(t_0)$ exceeds the threshold value W_0 , it is considered that it is subject the law of probability distribution with the density $p_1(W)$. The probability that such a signal will have values in the interval $[W_0; \infty]$ is equal to:

$$P_0 = P\{W_0 \leq W \leq \infty\} = \int_{W_0}^{\infty} p_1(W) dW. \quad (1)$$

It is termed as the probability of correct detection. However, it is possible that at the instant t_0 the signal at the optimum filter output, having response to the desired output signal, takes the value, which is less than W_0 . Then a false conclusion about the absence of the desired signal at the output of the optimum filter will be taken. The probability of such false conclusion, termed as probability of error of the second type or probability of target omission, is equal to:

$$P_{II} = P\{-\infty \leq W \leq W_0\} = \int_{-\infty}^{W_0} p_1(W) dW. \quad (2)$$

When the signal at the output of the optimum filter has no response to the desired input signal, it is subject to the law of probability distribution with the density $p_0(W)$. When such a signal exceeds at the instant t_0 the threshold value W_0 , a false conclusion about the availability of the desired signal at the input of the optimum filter will be taken. The probability of such error of the first type or the probability of a false alarm is equal to:

$$P_I = P\{W_0 \leq W \leq \infty\} = \int_{W_0}^{\infty} p_0(W) dW. \quad (3)$$

Probability of correct conclusion about the absence of desired signal at the input of the optimum filter, taken when $W(t) < W_0$, is

$$P_{III} = P\{-\infty \leq W \leq W_0\} = \int_{-\infty}^{W_0} p_0(W) dW. \quad (4)$$

In all conditions

$$P_0 + P_{II} = P_I + P_{III} = 1. \quad (5)$$

Analysis of considered situations shows that in case of decreasing the threshold of detection W , the probability of correct detection P increases and the probability of error of the second type becomes less, but simultaneously, the probability of error of the first type P increases and the probability of correct conclusion about the absence of desired signal decreases. On the contrary, in case of increasing the threshold of detection W , the probability of correct conclusion about the absence of desired signal P increases, and the probability of error of the first type becomes less, but the probability of error of the second type P increases and the probability of correct detection P decreases. Under such contradictory conditions, the optimum value of the threshold of detection W should be chosen on the base of some criterion, e.g. Neyman-Pearson criterion, criterion of ideal observer, criterion of minimum average risk, criterion of weighted combination, Wald criterion. Threshold selected in this way, is termed as optimum one for the given type of the optimum detector of signals of the searching aids, and the probability of correct detection with such optimum threshold as probability of the contact. Introducing the probability of contact into the equation for probability of detection in the theory of search allows to take into account that the search process runs in the presence of noise and a searched object location within the range of the technical searching aids does not mean yet that this object will be detected. As a rule, it is presumed that the optimum threshold is chosen provided that the search process is influenced by noise of natural origin. In case of artificial noise acting upon the searching aids, the optimum threshold chosen in this way becomes no more optimum one. As this occurs, characteristics of the law of signal distribution at the output of the optimum filter of the searching aids (dispersion, mathematical expectation, RMS-deviation) are changed and the optimum filter can become non-optimum to discriminate

signals against a background of noise of artificial origin, acting upon the searching aids. Probability of correct detection and optimum threshold of detection for this case can be calculated in a similar manner.

4 Efficiency of Search

Search of space-extensive targets is characterized by the availability of acquired productivity of search which characterizes constant gain in efficiency of search of space-extensive targets as compared with search of point targets and in most of search situations does not depend on the range of the technical aids and takes place as a result of:

- using geometrical dimensions of a space-extensive target by the searcher;
- searching activity of the target itself, manifesting itself in the fact that it “sweeps” the space, actually increasing mathematical expectation of the area being searched in a unit of time.

Productivity of search:

$$\Pi = \Pi_C + \Pi_{II}, \quad (6)$$

where Π_C is own productivity of search, Π_{II} is acquired productivity of search.

Own productivity of search characterizes the activity of the target in searching the space and acquired productivity of search, owing to spatial target dimensions, characterizes constant gain in efficiency of search of space-extensive targets compared to the search of targets belonging the small-sized ones.

Own productivity of search is determined by the range of detection and speed of the searching aids carrier. Operating range of the searching aids becomes accidental and decreases when influenced by noise. In its turn, this leads to significant decrease of own productivity of search. However, the acquired productivity of search does not depend on operating range and its contribution to the search potential increases in noisy environment, which fundamentally changes the search situation in noisy environment, and allows to increase significantly the probability of target detection in the presence of noise owing to the use of property of target space-extension.

If the probability of detection of the target and its trace is considered as index of efficiency, taking into account that the probability of additional

search of the target by its trace is equal to 1, it is clear that the gain of probability of target detection owing to detection of its trace is not great in comparison with detection of the target itself. Similar small increase of the detection probability is observed both with the increase of trace life and with the speed increase of the target itself. Analysis of the situation shows that the dependence of the probability of detection of the target itself on the factor of spatial trace extent, determining the efficiency of the aids for detection of physical fields of a search object, is weak as long as the acquired productivity of search is less than own productivity of search. Search potential increases in this case due to such a weak factor as addition of area, being searched in a unit of time, when the space-extensive target sweeps the space, the probability of establishing the contact with the trace, with the target itself in specific conditions of the searching aids use and in conditions of the real environment. When the values of acquired search productivity are greater than those of own productivity of search, there is a stronger dependence of the search potential on the acquired productivity of search. The fulfillment of this condition depends on both searching parts, but for a fixed group of parameters, characterizing the search situation, and $D = \text{const}$, there are some critical values of technical characteristics of the aids for search of space-extensive targets, which make the use of these technical aids inexpedient. Values of these critical parameters depend to a great extent on the operating range of the searching aids. In conditions, when artificial noise exerts its influence and the detection range decreases, characteristics of the searching aids, using physical fields of the target, become supercritical and the efficiency of search of space-extensive targets increases sharply. Similar situation is observed, when the search process is influenced by a complex of natural noise like the change of hydrological conditions in the search area (e.g. positive or negative refraction) etc.

5 Conclusion

Approach, offered in the report allows to estimate the efficiency of space-extensive targets search in the presence of noise both of natural and artificial origin. It is the natural development of the apparatus used in the theory for search of space-extensive targets, being elaborated by the authors at present, as one direction in the activity of the Scientific School in the field of radiolocation of the sea surface at the North-West State Extramural Technical University. The problem is considered in this way for the first time. There are no scientific works of other authors on this subject.

The approach considered can be developed both in the direction of the general theory perfection, and when carrying out practical calculations of the efficiency of the space-extensive targets search in the presence of noise.

Empirical Bayes Trajectory Estimation Based on Bearings from Moving Observer

A. Makshanov, A. Prokaev

St Petersburg Institute for Informatics and Automation, Russian Academy of Sciences
39, 14 Line, VO, St. Petersburg, 199178, Russia
prokaev@oogis.ru

Abstract. The problem of estimating an object's initial distance and velocity components via noised bearings from a moving observer belongs to a class of linear models with stochastic regression matrix, at that, estimates has uncontrollable bias. Bayes combination of least squares and orthogonal regression aimed at the above bias reduction is considered in this paper.

Keywords: Trajectory estimation; bearings; stochastic regression matrix; orthogonal regression; empirical Bayes estimation.

1 Introduction

Estimation of the object initial distance D_0 and velocity components (V_x, V_y) based upon the noised bearings $P(t)$ and noised observer coordinates $\{x(t), y(t)\}$ conventionally uses a general system of balance equations. Under assumption of straightforward and uniform motion these equations take the following form:

$$\begin{aligned} D_0 \sin(P_0 - P_i) + V_x(t_i - t_0) \cos(P_i) - V_y(t_i - t_0) \sin(P_i) &= \\ = x(t_i) \cos(P_i) - y(t_i) \sin(P_i) \end{aligned} \tag{1}$$

The estimation procedure is based on applying least squares (LS) routine to sequentially augmenting sample of bearings [1, 2]. But the underlying model (1) belongs to a class of linear models with stochastic regression matrix and, moreover, the variance of regressors relevant to velocities increases with time t . This specific character of the problem assumes a possibility of estimate refining via stricter specialized estimation techniques.

Statistically the model (1) belongs to a class of incorrect problems: Fisher information matrix depends on parameters to be estimated, the maximal likelihood estimate does not exist.

Assume the observer's trajectory consists of two straight line segments (tacks), coordinates of self-motion being determined with uncorrelated additive error. The following layout of solution is proposed:

1. At the first tack there exists lacking for objective information on the distance D_0 , so Bayes procedure based upon some independent estimate of D_0 is to be used. At this step we receive biased estimate of velocity components (V_x, V_y) , at that, bias depends on accuracy of the distance determining.
2. Upon starting the second tack we can simultaneously estimate all three parameters (D_0, V_x, V_y) . Together with LS some variant of orthogonal projection (orthogonal regression, OR [4, 5]) is used, that provides an estimate with a reduced bias.
3. LS and OR estimates are strongly correlated, however, they are oriented to minimization of different loss functions, and they are to be used in complex: the distribution of current OR estimate provides with prior distribution for Bayes estimate of (D_0, V_x, V_y) . It leads to an estimate with ultimately reduced bias, thus, allowing minimizing the required observation time.
4. Under straightforward uniform motion of the object and of the observer at each tack the sequence of bearing measurements permits a polynomial smoothing in a sliding window, that in fact introduces into (1) supplementary information on smoothness of altering the current parameters.
5. All estimates in (1) are biased, and the biases' values are practically uncontrollable. To a great extent they are determined by accuracy of motion character at each tack. On the other hand, their covariances may be estimated by some specialized techniques based on the disturbance theory.

In general, the proposed approach reduces the observation time by 20% in comparison with other existing approaches and provides for more objective exactness estimates.

2 LS – Based Estimation

Transform (1) to linear regression model $AZ = B$, where i -th row of matrix A and i -th element of vector B have the following form

$$\begin{aligned} A_i &= [\sin(P_0 - P_i)(t_i - t_0)\cos(P_i) - (t_i - t_0)\cos(P_i)], \\ B_i &= x(t_i)\cos(P_i) - y(t_i)\sin(P_i), \\ Z &= [D_0 \quad V_x \quad V_y]^T, \quad i = 1, \dots, N \end{aligned} \quad (2)$$

Here a standard LS estimate has the form

$$\hat{Z} = (\hat{A}^T \hat{A})^{-1} \hat{A}^T B. \quad (3)$$

At the first tack there exists lacking for objective information on distance; the matrix $\hat{A}^T \hat{A}$ in theory has no inverse. Though as A is stochastic, such an inverse may exist. Here corresponding estimates make no sense; real estimation begins only at the second tack. Further discussion only concerns an estimation along the second tack.

Here the well known covariance estimate $\text{cov}(Z) = \sigma^2 (\hat{A}^T \hat{A})^{-1}$ is misleading, for it does not take into account a stochastic scattering of elements of A and corresponding bias. The only approach to acquiring information on properties of estimates is based on statistical modeling [3]: current sequences $\{P_i\}$ and $\{x(t_i)\}, \{y(t_i)\}$ are contaminated by additive white noises with estimated standard deviations. At each step an estimate (3) is to be calculated and afterwards covariance of the estimate is determined through the realizations' ensemble. Bias cannot be estimated that way, it is related to specifics of initial sequences.

In LS based procedures it is as a rule possible to refine an estimate by equalizing the regressors' variances, however, this does not but it does not happen in the considered model (2). The problem is that parameters in (2) are not equitable: regressor concerning D_0 is limited, and regressors related to V_x, V_y , increase with t . For that reason we may get sufficiently exact estimates of distance even for moderate N , but estimation of velocity components requires longer observation time. When equalizing, esti-

mate of distance deteriorates, and it does not lead to V_x, V_y estimates amending.

In absence of object maneuvering the differences $B_i - A_i \hat{Z}(i)$ for each component represent white noise, so the sum of squares for any such sequence of length k obeys chi-square distribution with k degrees of freedom. It provides for a simple identifier of object maneuvering.

3 Orthogonal Projections (Orthogonal Regression)

The method of orthogonal projections (orthogonal regression, OR) [4,5] is LS competitor in model (2). Construct the matrix V as follows: in symbols of (2) its i -th row is $V_i = [-B_i A_i]$. Let $U = [u_1 u_2 u_3 u_4]$ be the latent vector of $(V^T V)$, corresponding to its minimal latent value λ_1 . Then the OR-estimate of $Z = [D_0 \quad V_x \quad V_y]^T$ is $Z_{OR} = [u_2/u_1 \quad u_3/u_1 \quad u_4/u_1]^T$.

OR-estimate is specially intended for models like (2) with stochastic matrix A , but its theoretical properties are poorly known only for special cases, where errors in each column of V are additive with zero mean and known constant variance. Experiments with model (2) show, that covariance matrices of OR- and LS-estimates are rather close, while OR-estimate bias is 2-3 times less. Moreover, it is more stable in case of bad observability (inadequate choice of the second tack).

In available investigations on OR equalizing variances of the columns in V is regarded as highly important, but this is not so for model (2) not corresponding to any theoretical layout. The point is that both for LS and OR principal input to quadratic loss is caused by D_0 , and minimization by velocities only brings to negligible improvement.

OR estimates may also be used in the above proposed identifier of object maneuvering.

4 Robust Data Smoothing and Aggregating

Introduced approaches under real error levels in bearings and self positioning give satisfactory results within 7 miles range. At longer distances an observation time increase is inevitable, that contradicts the practice requirements. Moreover, over ten minutes increase of observation time may

lead to the estimates' divergence due to progressive growth of dispersions in last columns of A (2). The only way here is to introduce supplementary information such as smoothness of altering bearings and self positioning at every tack.

The proposed approach uses observations' aggregated values being averaged at some system of partly overlapping sliding windows instead of observations. When blunders in measurements are possible, robust procedures of linear or quadratic smoothing are inevitable.

Suppose data are to be aggregated in a sliding window of size $2l+1$ that is shifted at any step to value d . Let the window include bearings $P = \{P_1, \dots, P_{2l+1}\}$. Ascribe them to conditional values of time parameter $-l, -l+1, \dots, l-1, l$ and fit them by polynomial Q of degree $m < 2l$. Smoothed value of the bearing at central point \hat{P}_{l+1} is to be taken as the value of Q at point l . When there are no blunders the fitting is to be based on LS.

In more detail let $Q(t) = q_0 + q_1 t + \dots + q_m t^m$, coefficients form vector $q = [q_1, \dots, q_m]$. The problem is to find q to minimize the square error $\sum_{i=-l}^l [P_i - Q(i)]^2$.

Introduce Gram matrix $G = [g_{ij}]$, $g_{ij} = i^j$, $i = -l, \dots, l$, $j = 0, \dots, m$, then $q = (G^T G)^{-1} G^T P$, $Q(0) = gP = \hat{P}_{l+1}$, where G_l is the l -th row of $(G^T G)^{-1} G^T$. So the smoothed value \hat{P}_{l+1} is represented by linear combination of $P = \{P_1, \dots, P_{2l+1}\}$ with coefficients G_l .

When blunders in P are assumed the fitting is to be based on the least moduli (LM). It means that instead of (4) the sum of absolute values of discrepancies

$$\sum_{i=-l}^l |P_i - Q(i)| = \sum_{i=-l}^l \frac{|P_i - Q(i)|}{[P_i - Q(i)]^2} \cdot [P_i - Q(i)]^2 = \sum_{i=-l}^l \omega_i \cdot [P_i - Q(i)]^2 \quad (4)$$

has to be minimized. Let $\Omega = \text{dias}(\omega_i)$ be the weight matrix and $q^{(0)}$ be the arbitrary (for (5) is convex) initial value of q . Then $q^{(0)}$ defines the initial approximation $\Omega^{(0)}$ to Ω and the next approximation to q is found

by weighted LS: $q^{(1)} = (G^T \Omega^{(0)} G)^{-1} G^T \Omega^{(0)} P$. This process lasts until minimum of loss function (4) with given accuracy is achieved. This is the procedure of iteratively weighted LS [6, 7].

Experiments with model (2) show values $l = 8\text{--}10$, $d = 4\text{--}5$ to be nearly optimal; at distances over 15 miles l has to be increased.

5 Empirical Bayes Estimation

Applying Bayes approach supposes regarding Z as a random vector and using one of characteristics of posterior distribution (as a rule – middle value) as its point estimate.

Theorem [5]. Let prior distribution of Z be Gaussian with mean Z_0 and covariance matrix K . Then LS based Bayes estimate is

$$\hat{Z}_B = (A^T A + K^{-1})^{-1} \cdot (A^T B + K^{-1} Z_0). \quad (5)$$

If Z_0 does not coincide with true mean of Z , the estimate is biased:

$$M(\hat{Z}_B - Z) = (A^T A + K^{-1})^{-1} K^{-1} (Z_0 - Z).$$

For stochastic matrix A covariance $\text{cov}(\hat{Z}_B)$ may be estimated only on the base of statistical modeling. Here the sole theoretical result is that $\text{cov}(\hat{Z}_B) < \text{cov}(Z)$, though due to bias it is only plausible for sufficiently accurate information on Z_0 .

The theorem points the only way to get the estimate at the first tack: given an estimate of D_0 and rough estimates of V_x, V_y , using diagonal matrix K with high dispersions, (5) supplies sequential procedure to obtain more precise estimates of (V_x, V_y) their biases being determined by accuracy of D_0 . At the same time (5) shows how to combine different estimates of vector (D_0, V_x, V_y) (D_0, V_x, V_y).

Specific character of the regarded problem is that two competing algorithms, LS and OR, oriented to different properties of the model (2) are

used. Asymptotically OR leads to less bias, but at limited number of observations different combinations of random factors may give preference to any of these algorithms. The point is that they as a rule work in anti-phase: one estimate gives an overstated value and another an understated one, so they may be combined based on (5). At that OR supplies a priori point estimate, $K = \mu \ diag(A^T A)$, parameter μ decreasing as duration of the second tack augments, and final estimate of all three components of Z is defined by (5). Such a procedure eliminates biases and reduces the required observation time. They also may be used in the above proposed identifier of object maneuvering.

6 Comparison of Estimates on Simulated Data

Regard the further situational model.

The object velocity is 12 knots, course 180° , initial bearing is 45° . Initial distance takes values 7, 12, 20 miles. Duration of the first tack is 240 s, velocity 6 knots, course 45° . Duration of the second tack is 360 s, velocity 10 knots, course 15° . Bearings are measured every 2 s, std of measurements is $10'$, std of self-defining is 15 m. Estimation is based on data being aggregated in sliding windows 16 steps width and 4 steps shift. Three estimates are regarded: LS, OR and empirical Bayes (combination of LS and OR). Fig.1 shows their typical dynamics starting 30 s of the 2nd tack.

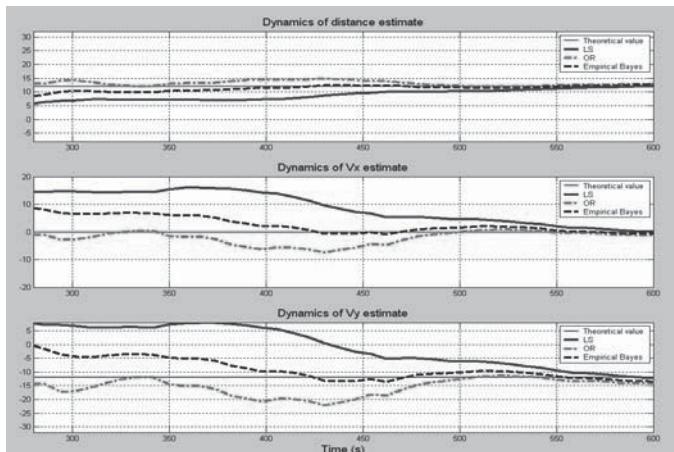


Fig.1. Initial distance 12 miles, sliding window width 32 c, shift 8 s

Tables 1 - 3 show dispersions and biases of estimates against “true” values of parameters based on 500 imitations. Principal result is that the dispersions of different estimates are approximately equal, at short distances the least bias has LS estimates, but at distances over 9 miles the combined estimates are preferable. When observation time exceeds 10 minutes OR gives better results, however, other algorithms are used here.

Similar results for 3 tacks are displayed in [8], in this case OR estimates are beyond competition.

Dispersions and biases evidently depend on distance and variance of bearings. Dependence on STD of self-defining (at values about 15 - 20 m) is much weaker. Fig.2, 3 show dependence of averaged biases and standard deviations with respect to STD of bearings at distance 12 miles. Fig. 4, 5 show similar dependences with respect to distance at bearings’ STD 10°. Advantages of combined estimates are evident.

Table 1. Statistical properties of estimates (averaging of 500 imitations). Distance 7 miles

	LS	OR	Empirical Bayes
STD of D_0 estimate (m)	0.16	0.16	0.16
Bias of D_0 estimate (m)	0.03	0.15	0.11
STD of V_x estimate (kn)	0.27	0.28	0.27
Bias of V_x estimate (kn)	0.08	-0.12	0.06
STD of V_y estimate (kn)	0.57	0.57	0.57
Bias of V_y estimate (kn)	-0.11	-0.53	-0.39

Table 2. Statistical properties of estimates (averaging of 500 imitations). Distance 12 miles

	LS	OR	Empirical Bayes
STD of D_0 estimate (m)	0.55	0.59	0.57
Bias of D_0 estimate (m)	-0.34	0.41	0.17
STD of V_x estimate (kn)	1.00	1.07	1.04
Bias of V_x estimate (kn)	0.83	-0.51	0.08
STD of V_y estimate (kn)	1.55	1.68	1.62
Bias of V_y estimate (kn)	1.06	-1.06	-0.3

Table 3. Statistical properties of estimates (averaging of 500 imitations). Distance 20 miles

	LS	OR	Empirical Bayes
STD of D_0 estimate (m)	1.37	1.90	1.67
Bias of D_0 estimate (m)	-2.56	1.30	0.06
STD of V_x estimate (kn)	2.68	3.55	3.16
Bias of V_x estimate (kn)	5.11	-2.07	0.24
STD of V_y estimate (kn)	3.47	4.69	4.15
Bias of V_y estimate (kn)	6.57	-2.99	0.08

7 Single Realization Based Covariance Estimation

In theoretical investigations true values of parameters are known and a possibility exists to receive N imitations getting every time some new estimate corresponding to next realization of random factors. These estimates averaging over the realizations' ensemble gives the mean bias and covariance matrix for the given estimation procedure and conditions. In such a way the results in Tab.1-3 and in Fig.2-5 are received.

In practice true values of parameters are unknown. After measurements are performed the sequences of bearings and self-positionings that define the estimate and proposed values for STD of errors are available. Perform N imitations using same sequences with different realizations of additive white noise. After averaging we arrive to Monte-Carlo estimate of covariance matrix [3] that coincides with similar matrix being received for the known "true" values of parameters that were used in initial model.

Example. Distance 12 miles. Averaging of 500 realizations with respect to true values of parameters gives for LS estimate with aggregated data the following covariance matrix

$$S_0 = \begin{bmatrix} 0.2775 & -0.5178 & -0.8025 \\ -0.5178 & 1.0310 & 1.5426 \\ -0.8025 & 1.5426 & 2.3531 \end{bmatrix}$$

and vector of biases $b_0 = [-0.1310 \ 0.5066 \ 0.5045]$. Monte-Carlo estimate based on single realization (without information on true values) is

$$S_1 = \begin{bmatrix} 0.2685 & -0.5194 & -0.7904 \\ -0.5194 & 1.0224 & 1.5411 \\ -0.7904 & 1.5411 & 2.3350 \end{bmatrix}$$

though this procedure does not supply any information on biases that are determined by a unique combination of random factors in the regarded single realization.

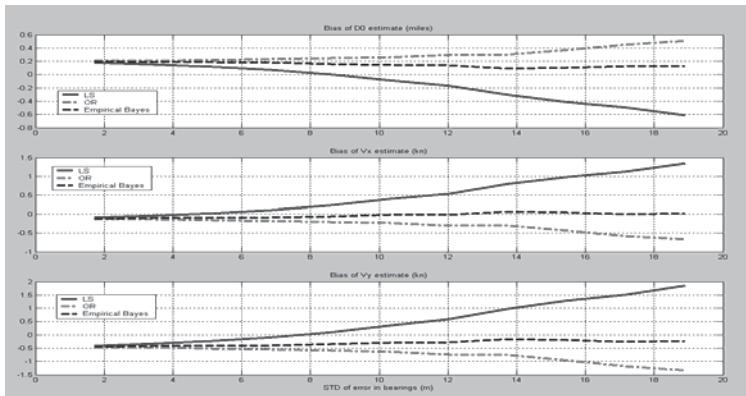


Fig. 2. Dependence of mean biases with respect to STD of errors in bearings. Distance 12 miles

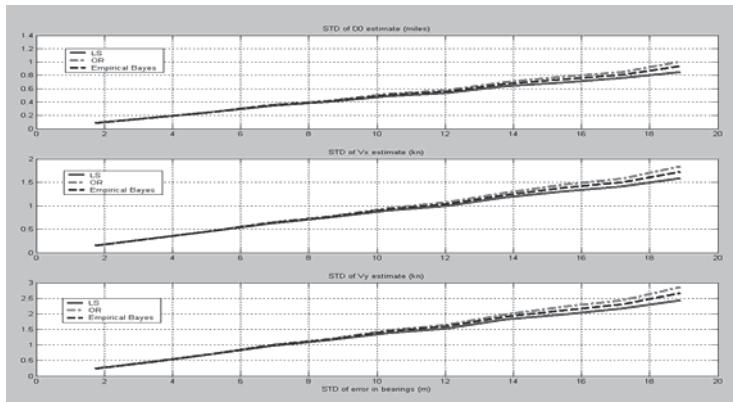


Fig. 3. Dependence of mean STD with respect to STD of errors in bearings. Distance 12 miles

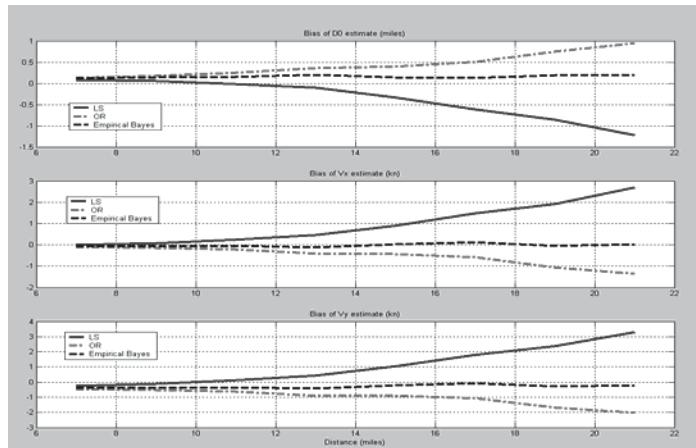


Fig. 4. Dependence of mean biases with respect to distance. STD of errors in bearings is $10'$

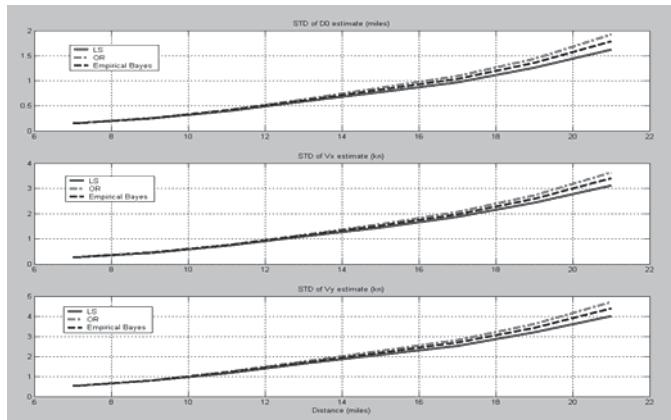


Fig. 5. Dependence of mean STD with respect to distance. STD of errors in bearings is $10'$

8 Conclusion

1. In all cases the highest accuracy is provided by the estimates based on aggregated data, at distances over 10 miles they have no alternatives.
2. Under long enough observation time the OR estimate has bias 2-3 times less than LS and are more stable to observation conditions (inadequate choice of the course at the second tack).

3. When observation time exceeds 10 minutes the effect of divergence due to specifics of regressors may occur. Estimates based on aggregated data are more stable; for them this effect occurs later.
4. LS and OR are oriented to different properties of the model (2), so random combination of stochastic factors may give advantage to any of them.
5. At medium observation time LS and OR estimates have biases of different signs that makes plausible the use of their combination based on empirical Bayes approach to accumulate their best features (Fig.2, 4).
6. All regarded schemes allow simple identifying of target maneuvering.
7. Proposed techniques for covariance estimating give adequate results coinciding with imitations at known “true” values of parameters. Though biases stay uncontrollable, the only way to estimation procedure improving is an implementation of the combined estimates with reduced bias.

References

1. Bulitchev Y.G. et al. Determining target coordinates on the base of bearings from mobile receiver. – In: Radiotekhnika, 1992, №4.
2. Makuhina T.P. et al. On-line coordinate estimate of mobile object based on bearings. – In: Questions of Radio Electronics, ser. ASUPR, 1992, Is. 2.
3. Efron B. Nonstandard methods of multidimensional statistical analysis.- Moscow: Science, 1988.
4. Demidenko E.Z. Linear and nonlinear regression. – Moscow: Finance and Statistics, 1981.
5. Krianev A.V., Lukin G.V. Mathematical methods of treating indefinite data. - Moscow: PHISMATLIT, 2003.
6. Mudrov V.I., Kushko V.L. Methods of treating measurements. - Moscow: Science, 1976.
7. Makshanov A.V. and coauthors. Robust filtering of time series of navigational bearings. – In: System analysis in creation and employment of ships, armaments and military techniques, Is.16, SPb: 2005.
8. Makshanov A.V. et al. Theory and methods of estimating target coordinates and parameters by orthogonal regression. – In: System analysis in creation and employment of ships, armaments and military techniques, Is.16, SPb: 2005.

Author Index

- Biermann, J. 84
Bourgeois, J. 111
Breunig, M. 234
Buford, J. 18

Cheng, T. 173
Chervatuk, O. 128
Claramunt, C. 34
Collins, G. 301
Coskun, M. 202

Devogele, T. 34
Fournier, S. 34
Frey, M. 84

Galjano, P. 45
Ganame, A. 111
Gold, Ch. 286
Goralski, R. 286

Jacobson, G. 18

Kashevnik, A. 69
Kirma, C. 202
Kotenko, I. 111, 128
Kravtsov, A. 187
Kriuchkov, A. 187
Kruger, K. 84
Kvachev, S. 252

Levachkine, S. 209, 223
Levashova, T. 69
Lewis, L. 18
Llinas, J. 1

Makshanov, A. 323
Martinez, M. 223
Moldovyan, A. 147
Moldovyan, N. 147, 160
Moldovyanu, P. 147
Moreno, M. 223

Noyon, V. 34
Pan'kin, A. 272
Petit, M. 34
Pikhtin, N. 194
Popovich, V. 45
Potapychov, S. 272
Pozarek, W. 96
Prokaev, A. 323

Ray, C. 34
Schade, U. 84
Schrenk, M. 96
Sergushev, A. 316
Shilov, N. 69
Shishkin, I. 316
Sidelnikova, E. 128
Smirnov, A. 69
Sokolova, L. 252
Sorokin, R. 263
Sotikova, E. 187

Tarakanov, A. 252
Tarasov, I. 194
Thomsen, A. 234
Tishkov, A. 128
Torres, M. 209, 223

Ulanov, A. 111
Unen, H. 202

Wang, J. 173
Yilmaz, E. 202

Printing: Krips bv, Meppel
Binding: Stürtz, Würzburg