Geoinformatica (2018) 22:383–410
DOI 10.1007/s10707-017-0308-z

OSCAR: a framework to integrate spatial computing ability and data aggregation for emergency management of public health

Danhuai Guo [1,2] · **Yingqiu Zhu** [1,2] · **Wenwu Yin** [3]

Received: 1 February 2017 / Revised: 17 July 2017 / Accepted: 29 August 2017 /
Published online: 26 September 2017

Abstract Spatial computing has emerged a critical issue in emergency management of public health. Due to the complexity of spatial data structure and disperse character of spatio-temporal data, when emergency event of public health occurs, it is difficult to get the needed data and analysis it then make quick decision in a short time. In this paper, OSCAR: an Open Spatial Computing and data Resource platform were introduced including its components, framework, elements and two implementations. OSCAR provides a data resource aggregation platform to retrieve data from official statistic agencies through data service and database, scrawl related data from BBS and social media and mirror the environment data from earth observation data sites. All the dataset are arranged in data cubes according to their spatial and temporal dimensions. This mechanism ensures the feasibility and timeliness of time-sequence analysis of specific regions. The algorithms of spatial computing of public health are usually complicated and depend on particular computing environment, which is usually not default configuration of computer of nowadays. OSCAR deploys a series of computation images in a cloud-computing environment. The computation ability can be extended on-demand and thus the time of the computation can be shortened and limited in several minutes when it is needed. The two implementation of human rabies of China and H7N9 in China show the convenience of our platform.

Keywords Spatial computing · Data integration · Emergency management · Cloud computing · Public health · Framework

✉ Danhuai Guo
guodanhuai@cnic.cn; guodanhuai@gmail.com

[1] Computer Network Information Center, Chinese Academy of Sciences, Beijing, China

[2] University of Chinese Academy of Sciences, Beijing, China

[3] Chinese Center For Disease Control And Prevention, Beijing, China

1 Introduction

Public health events are increasing in recent years and showing a trend to cause greater harm in greater range with shorter duration. For instance, H5N1 caused hundreds of death among Asia just in several months in 2003 [1, 2]; H7N9 spread across numerous provinces in China and cause huge economic loss in 2013 [3]. The leading cause is the increasing intense of human activities in both spatial span and temporal duration, which constantly change ecological environments [4]. Many literature show the spreads of infection disease are highly related with human migrations. The economic globalization brings more frequent and convenient human interaction and brings more frequent and more severe outbreak and spread of global public health events in last decades [5–7]. Emerging infections (EI), which is defined as newly appeared in population or ever existed previously but are rapidly increasing in incidence or geographic range, have caused incalculable misery and death in human history [8]. In recent years, may re-emergences have catalyzed by occasional regional armed conflicts, Refugee flows, loss of cohesion, and natural disaster like earth quakes and floods, indicating the importance not only of microbial and viral factors, but also of social and environmental determinants [9, 10].

It is known that the spreads of infectious disease are highly related to geographic and environmental conditions. Therefore, the challenge to emerging infection disease researchers is to explore the correlations between diseases and relevant social and geographic environments then conduct comprehensive analyses involving geographic or environmental factors [8, 11].

Human activities as well as change of geographic conditions and ecological environments play significant roles in the spread of diseases. Spatial analysis involving those factors can directly contribute to prevention and control of diseases. Unfortunately, the human activity, which is efficient to emerging infection disease analysis, can not be directly obtained from statistical data due to lack of reliable data source or spatio-temporal accuracy. It is difficult to detect the human behavior in physical space, but it is much easier to detect them in cyber space. The internet behavior, which include internet search, e-Commerce, and social medial etc., is verified as reflection of physical behavior and its importance is being realized by public health researchers [12–15]. But for public health researchers, it remains a big challenge to collect web data and transform these data to environment factor being used in spatial epidemic analysis. It is needed daily run platforms to aggregate data from thousands of news of web site, discussion from internet forum, transactions from e-commerce, post from social medial. These huge volume data source should be processed such as cleaning, geo-coding and aggregating before they are used in spatial epidemic analysis.

With the increase of data volume and consideration factors number of emergency response of public health, parallel computing and clouding computing resources is urgently needed. Due to the emergence of new viruses, which have not yet been well studied and the infection routes are still not clear, simulation of the spread and the progression of diseases is increasingly significant. When new viruses break out, there isn't much time for researchers to find out the exact pathogenic mechanism and infection route before decision-makings on disease control. To minimize the effects of a disease, medical staffs and the government must have the ability to quick respond to the disease. To support the quick response, simulation based on spatial computing is utilized as a practical method to analyze factors, including biological, environmental and socio-economic factors, that lead to the outbreak or promote the spread, and thus provide reference for predicting the spread and finding optimal measures to prevent and control the disease. With the help of spatial computing, epidemiologists and public health

researchers can exploit the rich spatial content of geographical data and find explicit or implicit spatial relations and interactions among spatial objects, which contributes to the analysis of diffusion of diseases, especially infectious diseases.

The primary objective of spatial analysis for public health issues is to unravel credible causes of diseases, which conduce to better understanding of characteristics of the diseases as well as support forecasting possible outbreaks of the diseases. Therefore, the spatial analysis is supposed to provide valuable inferences for emergency management of public health affairs. In this paper, we focus on how to systematically raise both the effect and the efficiency of the spatial analysis process, hence optimize the predication of outbreaks of diseases and better support the emergency management.

In this paper, we aim to discuss a decision support system based on GIS for emergency management on public health issues with respect to the above challenges. For this purpose, we develop a comprehensive an Open Spatial Computing and dAta Resource (OSCAR) platform for spatio-temporal analysis to provide one-step services for relevant researcher and decision makers. The main contributions of the framework are as follows:

1. Data aggregation. The system gains abilities to automatically search from open data sources and collect relevant information, extract corresponding geographical locations as well as organize collected data in a specific form, called as spatio-temporal cube, which facilitates the storage and retrieval of the data and well conveys the information in the temporal dimension and and the spatial dimension.
2. Model integration. The system incorporates different models, each of which is built with a specific data assumption. The system provides a solution to integrate models at the code level and offer unified service interfaces for users of the system. To ensure high efficiency of the analysis processing involving different models, the system is designed to exploit cloud computing resource adaptively and compress the time of calculating.
3. Integrated visualization portal. The system provides a web-based integrated visualization portal to provide data aggregation, data processing, model select and computation resource.

The following of the paper is organized as follows. Section 2 introduces related works that concentrate in emergency management, GIS and researches of public health issues. Section 3 describes in detail the proposed framework along with its specifications to handle public health issues. Section 4 demonstrates two case studies to verify the performance of the proposed system using data of human rabies and H7N9 in China. The final section concludes the whole paper and provides discussions on future development.

2 Related works

FEMA publishes and updates detailed criteria for considering the dispatching and management during emergencies, which is aimed at protecting property, public health, and safety and lessens or averting the threat of an incident becoming a catastrophic event [16]. System for emergency management has been studied for many years and numerous sophisticated systems have been already developed to deal with natural and technological hazards, including hurricane, flood, and other threats to communities. [17–19] Spatial analysis based on GIS is a critical tool for emergency management. GIS is well embedded into the system of emergency

management and has a number of successful applications on management for disasters like earthquake, chemical disasters, [20–24]. GIS expands dimensions for the analysis supporting decision-making during emergency management tasks. Geospatial information is exploited and integrated with expert experience, web data. Based on the integration of information, a number of GIS-based information systems that contributes to emergency management have been developed [25–27]. GIS also shows its ability to help resolve public health issues. Geoinformatics technology can aid in epidemiological investigation and outbreak response and thus reducing the health hazards in the communities before, during and after epidemic episodes. Integration of GIS and GPS improves the quality of spatial and non-spatial data for analysis and contributes to decision-making through providing integrated approach to disease control and surveillance. Moreover, GIS mapping and modeling enable spatial and temporal analyses to enhance understanding of the population's access to health services and the emergency referral system, which leads to better management during emergencies [28]. To support emergency management on public health issues, ArcGIS server, a typical GIS tool, is widely used in the shortest path algorithm and displaying of maps. Researchers build web based public health emergency systems to simplify operations for the dispatch of multi-resource [29]. Typical GIS tools like ArcGIS and GeoData provide reliable assistance for the spatial analysis on public health issues. However, a comprehensive analysis involving understanding of epidemiology calls for functions exceeding what a single GIS tool can provide. Although GIS has been applied in research works on diseases, there are few mature GIS-based frameworks or systems proposed for emergency management on public health issues. A precaution management system based on GIS for public health issues has been designed but the functions were rather limited. [30] Integration of data and models can contribute to the comprehensive analysis. Open source platforms like LabKey [31] help researchers to retrieve, organize and share heterogeneous disease data, and then conduct cooperative analysis. Several frameworks have been proposed to integrate models in spatial decision support systems. The model integration is implemented using reusable toolkits with specific management aimed at avoiding conflicts. Problem-oriented methods have been incorporated in those integrated systems to support the deployment and management of different models as well as enable a declarative description of models and the conceptual relationships. [Taylor, 1999] The experience on the integration of multi-source data and models can provide reference for designing a comprehensive integrated system for spatial analysis on public health issues.

To make a deep spatial analysis for public health issues, more specifications based on professional knowledge of diseases or other health issues are necessary. Numerous studies have been conducted to investigate the epidemiology and transmission dynamics of public health issues, especially zoonotic infectious diseases, across different temporal and geographical scales [32–40]. Studies investigating the spatial patterns of some infectious diseases have shown obvious variations due to the differences in geographical, climatic and environmental attributes: distance from major roads, presence of river or lake, land cover such as deciduous forest, average temperature and neighborhood to enzootic zones [41–43]. In order to prevent and control public health issues, those factors are worthy of consideration. A number of international studies have investigated the contribution of risk factors using in the spatial distribution of diseases using various types of data [44–48]. To take more factors into consideration, it is advisable to integrate data of different types into a comprehensive model for the analysis. Among numerous factors, unsuccessful control of pathogenic animal and inadequate post-exposure prophylaxis (PEP) of patients are thought to be critical and intuitive

signals leading to the high incidence of diseases [37, 38]. Some studies have tried to find the correlation between disease exposure risk and socioeconomic status as well as ecological variables using the records of receiving PEP measurements [49–51]. Surveillance data have been recently exploited in spatial models to forecast raccoon rabies emergence [46]. To take more factors into consideration, it is indispensable to integrate data of different types into models for the analysis. However, these studies did not consider the spatiotemporal variation in environment factors. The heterogeneity of local regions was also ignored. In addition, variation derived from spatially lagged geographical variables was inadequately accounted for. The only spatial epidemiological study of canine rabies has shown that the spatial and temporal distribution of canine rabies was not evenly distributed across China [52]. With more factors allowed for, spatial effects can no longer be ignored. Spatial principles, which exist among physical phenomenon and reflect solid patterns in spatio-temporal dimensions, shed light on specifications for optimization of spatial analysis. To investigate subjects with spatial structure, approaches customized according to spatial principles are more appropriate while analyses through a single model or simple algorithm that fails to consider spatial effects become less worthy. Besides application of spatial principles, another solution is to integrate different models or algorithm, each one may be built on a specific assumption, to analyze the issues from different perspectives. On the other hand, an aggregation of different models can output comparable results, which are expected to contribute to the design of better interventions in the context of governmental and medical decision making, which is aimed at reducing cases of diseases.

The data aggregation idea and platform are often found in gene related research and applications [53]. These platforms collect and aggregate sequence gene data for organization and analysis. Its worth increases with the data volume it aggregate. Some crowd-sourcing platforms provide data resource services through share among users. But these data service remain in data resource download level as most of data not being processed or organized for a general purpose. Big data quality problem remain in these systems.

3 Framework of OSCAR

3.1 The D-M-V three tiers architecture

There are three designed goals of OSACAR. Firstly, OSCAR is a reliable public health related data aggregation platform and provide data services of data gathering, data organizing, geocoding and data sharing. Secondly, it is a emerging high performance spatial computation platform of public health issue by integrating different spatial analysis model, providing on-demand spatial computing cloud service combining data resource. Finally, OSCAR is designed and implemented as a decision support system portal by providing data-driven analysis and simulation services. Researchers on public health issues, medical staff, and policy makers are target users of the system. Although the system is designed to provide user-friendly interface and visualization for users to obtain intuitive results, basic knowledge of spatial analysis and statistics are required for users who exploit the system. The system is designed as both application-driven and user-centric, providing user-friendly interfaces to implement spatial analysis.

Traditionally, the data analysis system usually is designed as a stand-alone system, in which the function of data processing, computation and visualization are implemented in a single

computer or computation node with same or similar configuration. This architecture works well when the data size is small, the computing complexity is low, and there is no need to integrated with multi data source to implement complex comparative analysis. With the growth of data size and increasing demand of complex analytics algorithm and distribution of data and user, it is more and more difficult to implement a complex visualization task in a stand-alone architecture. To meet the three designed goals, a D-M-V architecture [54, 55], which has been proven efficiently in collaborative visualization system, was adopted in OSCAR. Therefore, harnessing parallel and distributed resources turns to become an advisable solution. In addition, user-centric approach, which presents uniform interfaces instead of detailed management or processing of distributed resources for users, is exploited to facilitate the analysis through reducing cumbersome works.

3.2 Service-oriented architecture design

The system also utilizes Service-oriented architecture (SOA) to integrate GIS and spatial analysis models into the workflows. SOA, which is based on loosely coupled software services, allows for flexible and scalable aggregation of sophisticated applications from individual services and manages applications with common interfaces and standards. SOA is widely used during the phase of systems development and integration. In this paper, we design a three-tiers services-oriented architecture, which is shown in Fig. 1, to provide a scalable platform for spatial analysis supporting emergency management on public health issues. Workflows can interact with individual services to make the most of the capabilities of high-performance computing, distributed resources, data management, and visualization.

Owing to the supports by SOA, the system gains benefits at least three aspects:

1. Loose coupling of services makes it available to distribute different components of the system, like data storing, data modeling and data visualization, into different computers. Based on the characteristics of distributing and loose coupling, SOA provides the possibility to employ parallel computing resource for analysis tasks.
2. Unified service access portal add to the feasibility to provide uniformed data access and services access for users, regardless of the sources and formats of the original data, variant analytic algorithms and different operating system.
3. Integration in web-based environment platforms conduces to the deployment of different Web-based applications and map services, which may be offered by other system or applications. The system illustrates comparative analyses for users to find hidden rules from spatio-temporal data.

Data integration and model integration serve as the most important roles in the whole system. Spatial principles [56], which reveal spatial and temporal constraints and interactions among spatial objects, are considered to optimize the integration of data and model. To support the analysis tasks, a complete mechanism covers data collection, data organization, spatial analysis models, job scheduling, visualization and several detailed procedures is necessary. As is shown in Fig. 2, the whole framework is comprised of 5 general levels: automatic data collection, data organization and integration, model integration, job scheduling and integrated visualization.

The most pressing works here are to reasonably process and organize the data with the spatial structure well preserved, to incorporate models revealing the implicit patterns among

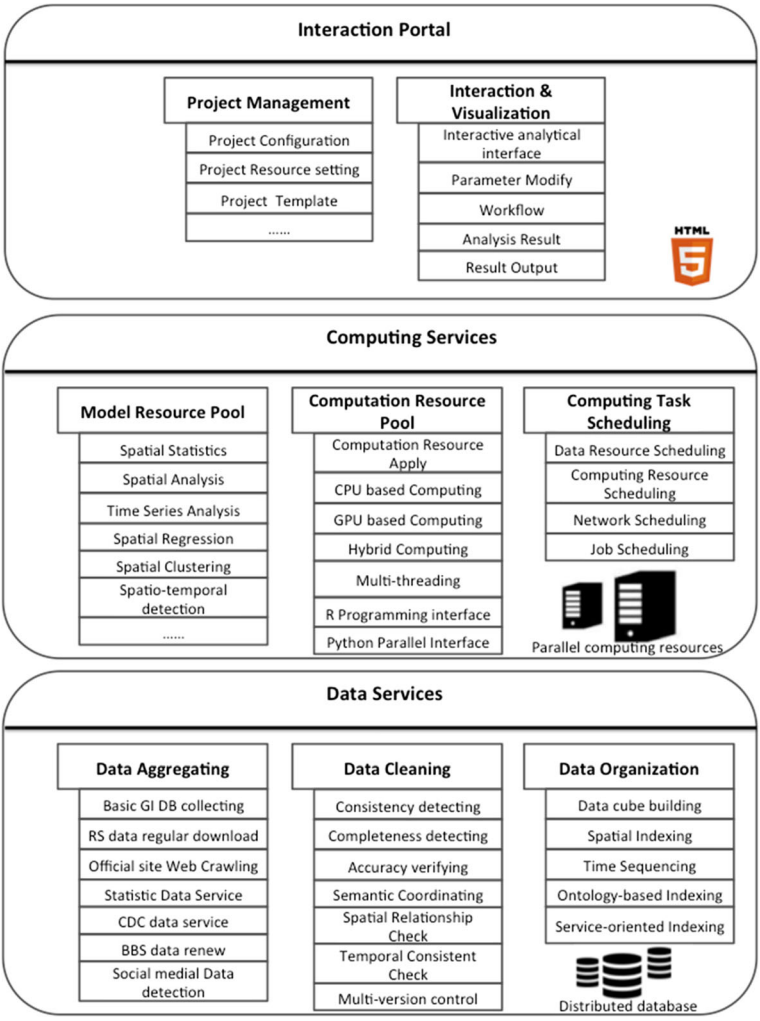

Fig. 1 OSCAR: three-tiers services-oriented designed architecture

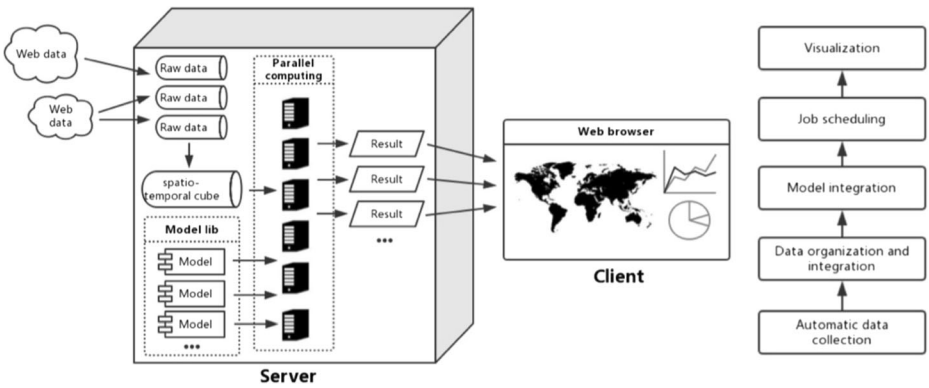

Fig. 2 Five levels of the framework of OSCAR

the data and to synthesis a set of methods for decision makings, which are tailored to GIS and spatial analysis. The system is developed to capture essential spatial characteristics of problems. Besides, to fulfill requirements of emergency management, the system must be developed to be fast in response to any sudden emergencies. Therefore, the system has to perform efficiently in data modeling and analysis as well as ensure the correctness of its indicators. A balance between the efficiency and the effect is critical to the framework. Technical aspects of each level are introduced in detail in following subsections.

3.3 On-demand data aggregation

Automatic data collection is the foundation of the whole system. Specifically, it contains the detection for emergent events, automatic data acquisition and the storage of raw data.

The data needed for analysis can be classified into two types: static data and dynamic data. Static data is the data that mainly released on official websites of the government or institutions, where the data is updated relatively unfrequently or regularly. Static data provides basic information including digital maps and other geographic data, climatic data, remote sensing data that records environmental characteristics and socio-economic statistic data. Dynamic data is the data that collected from dynamically updated web media and social networks. News websites, blogs, web forums and twitter messages are set as the targets of automatic crawlers.

There are two concerned challenges to our data discovery process: to properly deploy automatic crawlers for different data sources with totally different data structures, and to keep the latest information detected and stored timely, which means the automatic process cannot be too time consuming.

The deployment of crawlers varies for the static data sources and the dynamic data sources. The static data sources provide well structured data with explicit geographical coordinates or locations. Official websites often release data at regular intervals. To receive those static data with a controllable overhead, crawlers collecting static data are assigned with timed tasks to check for updates of websites and request the latest released data. Owing to the constant structure and explicit geographic labels, the collected data does not need much complex processing before storage. For the dynamic data, the automatic collection is more sophisticated. To obtain the real time information, crawlers have to continuously monitor web pages of news and social media and search for special keywords, including descriptions on emergencies, names and acronyms of diseases, which can be seen as signs of occurring emergencies.

Text data with themes relevant to emergencies on public health issues are extracted to detect cases. The contents of web texts are usually informal and random, thus make it confusing to detect cases and find sufficient information. In order to obtain enough information to describe a disease case, the context is also collected. Within the context, an ontology network, which consists of a set of ontologies that denote diseases and other threatens to the public health and semantic relations among ontologies, is exploited to help determine the topic described, and then rapidly identify cases of emergencies. The ontology network can be generated from professional understandings and experiences of experts in emergency management. The crawlers for dynamic data are deployed on demands of users and can be specified according to manual settings.

When it comes to the data for analysis, the records of disease cases are of the most importance. In our system, users can initialize the search targets by selecting research topics such as rabies, H7N9, Ebola, etc. The cases data come from both records in static tables on the

official websites and texts from web. A serious problem is that a substantial part of texts from web are represented without geographical coordinates and thus bring obstacles to further spatial analysis. To address it, we conduct reasoning based on the additional information extracted from the texts. When there are no explicit coordinates, the system seeks phrases that mention locations or provide some intimations, for example, names of places, restaurants, buildings or tourist spots. With such phrases recognized, it is readily to match them with corresponding geographical regions and then label the cases with coordinates.

In light of the further spatial analysis, the raw data from official websites and web texts need to be formatted using unified geocoding, which is conducted using scripts calling Google or Baidu map APIs. Once we have mapped the locations into a consist coordinate system, the data of other variables, including ecological and socio-economic explanatory variables, can be indexed according to their corresponding positions in the system.

The collection is implemented using a distributed search engine automatically. Distributed databases such as Hbase, which enable the system to handle massive data in limited time, are embedded in the system with traditional relational databases like MySQL. Kafka, a widely used distributed streaming platform, is incorporated to publish streams of newly collected data. With distributed components, the system gains the ability to store large amounts of data and well process real-time data streams. Besides, the flexibility of distributed components makes it practically facile to extend the scale of clusters of computers used. Users can adjust the scale of the system to fit the task of data collection while keep a high efficiency. The framework of the collection system is illustrated in Fig. 3.

3.4 Spatio-temporal based data organization and integration

Data organization, in broad terms, refers to the method of classifying and organizing data sets to make them more useful [12]. In the field of spatial analysis for public health, different types of data such as raster data, vector data need to be restructured for further analysis and integration. Raster data, vector data and other spatial data collected are transformed using the GDAL library, a translator library for geospatial raster and vector data formats, released under an X/MIT style open source license by the Open Source Geospatial Foundation. Data with spatial structure in different types are mapped into a unified coordinate system. Feature information of the cases are extracted from the geo-data using GDAL and related with

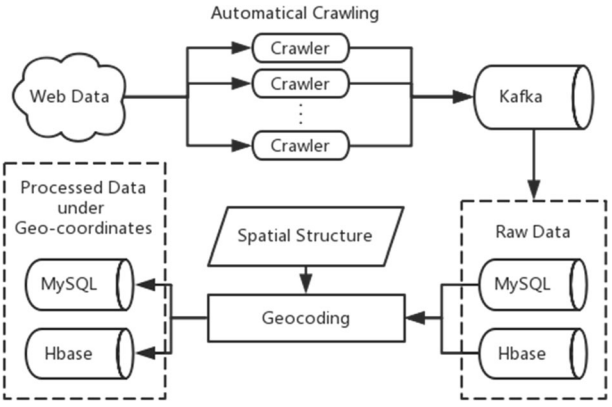

Fig. 3 The flow of data processing and organization in OSCAR

corresponding positions. In our system, all the geo-data can be processed automatically by executing program scripts developed based on the GDAL library.

The processed data are stored in a spatial data warehouse using a spatial data cube model [57], or called a spatial multidimensional database model. To combine spatial data with non-spatial features, a star schema model is adopted, which consists of spatial dimensions and non-spatial dimensions. As a data warehouse, the storage components are organized as cuboids, which are composed of numerous cells. Each cell describes a spatial object obtained from merging of spatial objects that share same non-spatial properties. The merging can be implemented through spatial clustering while explores materialization at finer granularity. Closely connected small areas, which have same or similar characteristics and environments, are treated as an individual entity. The cell serves as the primary unit for any operations on the warehouse. The merging, which can be seen as a pre-computation, can not only contribute to fast response for OLAP operations on the database, but also provide a better format of data for spatial analysis. Moreover, with those frequently used and shared spatial objects clustered, we can save a huge amount of storage space.

The data warehouse adopts a star schema model, which has been widely applied in data warehouse. As is shown in Fig. 4, the star schema model consists of a large fact table, which stores basic information of spatial objects, and a set of dimension tables, which are linked to the spatial objects and store corresponding features or explanatory variables. The dimension tables can be further split to finer granularity and thus present data at different hierarchies. With the star schema model incorporated, the data warehouse is available for typical OLAP operations like drill-down, roll-up, dicing, slicing, pivoting, etc.

In consideration of OLAP operations on the data warehouse with a spatial structure, the dimensions and measures, which may involve spatial components, should be specified. The dimensions are classified to three types:

1. non-spatial dimension: a dimension that contains data without spatial structure, such as temperature. The generalizations of non-spatial dimensions do not depend on spatial distributions.
2. Spatial-to-nonspatial dimension: a dimension that is spatial at fine granularity or low hierarchies while presents to be non-spatial at high hierarchies. For instance, administrative division is a spatial dimension at levels of villages, cities and provinces with respect to

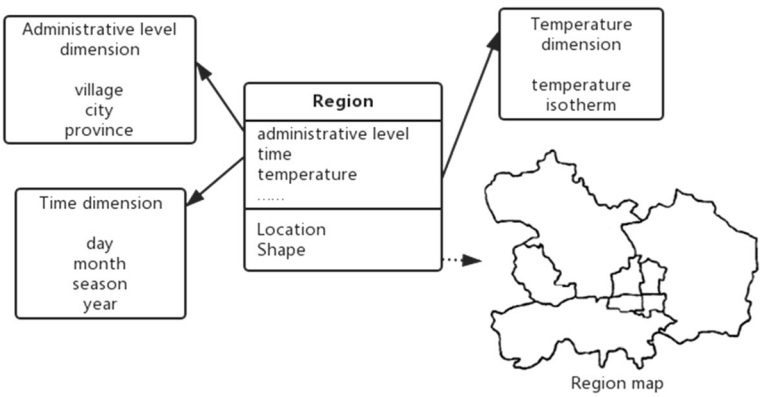

Fig. 4 The star schema model for spatial objects

neighboring correlations between regions. However the dimension comes to be non-spatial when we roll-up to a high level such as Western Provinces, where spatial correlations are contained inside but have no influences on measurements or operations on the whole generalized objects.

3. Spatial-to-spatial dimension: a dimension that remains spatial at any granularities, such as region in the same isotherm, which always lies on the geographical structure regardless of the change of temperature.

The measures in the cube are also classified according to whether spatial structure is involved:

1. Numerical measure: a measure that works with only numerical computations, involving little spatial information. It can be conducted by cube partition and distributed aggregation or algebraic manipulations and well fit the non-spatial dimensions.
2. Spatial measure: a measure involves spatial objects. When handling objects under a spatial structure, measurements and operations, like roll-up, are based on the geographical structure, where we can obtain corresponding geographical locations and the adjacency between the spatial objects. It may require several additional processes like spatial clustering.

In the process of the data organization, the raster data and vector data are usually very big, such as the temperature, NDVI and DEM, each of which generates a large TIF file every day. The total amount of such a feature can be more than 50GB within a year. Fortunately, in the era of big data, the computing resource becomes more powerful and cheaper nowadays, and therefore the distributed computing is used in order to reduce the computing time and satisfy the need of quick response in our system.

3.5 Hybrid model integration

Due to the diversity of algorithms for spatial analysis and different perspectives they focus on, it is often expedient to make comparative analyses through different algorithms with the same datasets and obtain comprehensive inferences. However, it is difficult to implement all relevant algorithms in a stand-alone computer, even if the source codes can be downloaded freely. Specifications, configurations and requirements for environments of different algorithms or models will bring endless obstacles to the integration. For a system consists of numerous computers, keeping the consistency of configurations of models within the system is also troublesome and has a high overhead. Moreover, as a result of various structures of GPU and CPU, the same algorithm may perform sharply different under two specific computation structures. From the perspective of users, usually, the case is that users only hope to use a finite number of algorithms for a certain problem. Therefore, calling numbers of models may be redundant for users to handle an analysis task and lead to a waste of computing resources.

To address it, it is recommended to deploy models on distributed nodes within the system using a service-oriented cloud architecture. In our framework, different models are integrated in a unified model library. Within the model library, models are separately installed on different virtual machines according to their specific requirements for coding languages or environments. Although different virtual machines may be physically bound together at a physical machine, they are supposed to have few impacts on each other during most of their work. The

decentralized deployment of the model library contributes to conflict avoidance and access control of the system. Each model is properly set and work steadily as in a stand-alone computer. Commonly used algorithms or models, such as logistic regression, decision tree, neutral networks and SVM, can be readily incorporated into the library and function correctly. R software environment, a free and open collection of tools for statistical analysis, is also embedded in the model library. Based on the R environment, packages for linear modeling, time series, multivariate methods are incorporated for users as alternatives. For the spatial analysis, major commercial GIS software such as ArcGIS® Server, ArcIMS®, Envinsa®, GeoPublisher®, GeoMedia® and Oracle MaperViewer® are also accessible.

To optimize the space allocation and improve the efficiency of the aggregation of models, a service-oriented cloud architecture is adopted, which is conductive to the management of distributed models and corresponding resources. As is shown in Fig. 5, the cloud computing architecture integrates Infrastructure as a service (IaaS) and Software as a Service (SaaS):

The infrastructure Layer covers the hardware facilities that support the whole system, including computers, storage and network devices. The infrastructure layer serves as the physical basis of the system. To manage numerous devices with different capabilities and different settings, devices incorporated must follow uniform standards. Moreover, request dispatchers using Virtual Machine Monitor and Service Governance Service are recommended to optimize the allocation of requests to available recourses. In order to ensure the coherence between working processes on the storage, the reading function is implemented using several accessors. In light of the fact that the multi-source data integrates data with different and complex structure, each of the accessors is specified to fit certain format or structure. When the

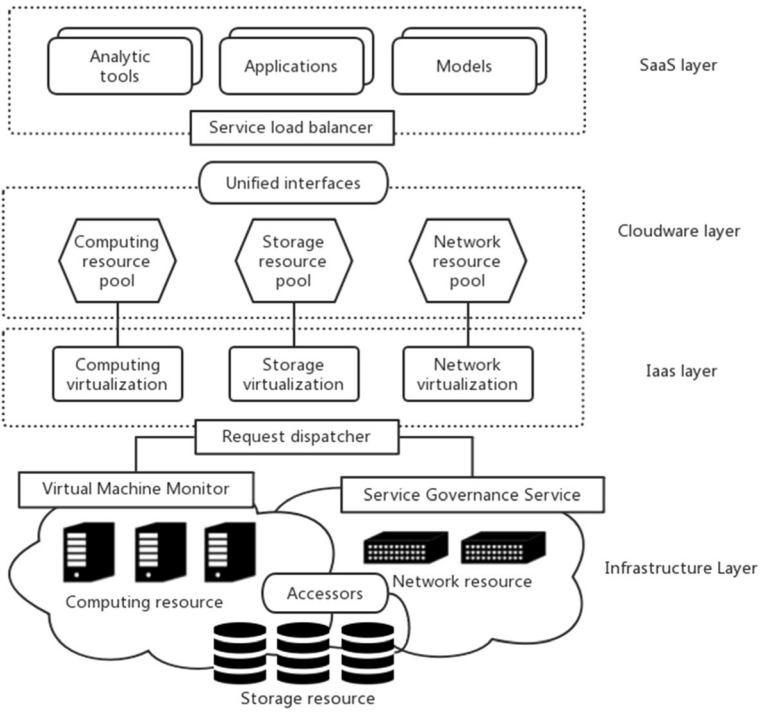

Fig. 5 Model integration of OSCAR in cloud architecture

data is to be accessed, the system calls proper accessor that matches the structure of the data to make retrieval and sampling. Therefore, users are not required to have knowledge about the internal data structure. Moreover, operations on data are uniformly arranged by the system and thus keep the feasibility and consistency of the storage.

In the IaaS layer, virtualization technologies, including computing virtualization, storage virtualization, network virtualization, etc., are exploited to shield the differences and complexities of the hardware and models. Through virtualization layer, resources are presented as a standardized, flexible, extensible virtual resource pools.

The cloudware layer serves as a management platform for the dispatching and combination of resource pools. The pools are linked with virtualized resources through ontology mappings. In the cloudware layer, unified interfaces are generated to automatically allocate resources according to the demands of analyses tasks or applications. The arrangement of the computing resources and storage resources is invisible for users. Users only need to call well encapsulated interfaces to initialize, start up, pause or suspend models regardless of what operations are implemented on machines.

In the SaaS layer, applications, applications, analytic tools and models are presented to end users as services. Users can utilize available services on demand. The interactive interface can be implemented in a browser, though which users can get rid of troubles of deployment and management of numerous models. SaaS can be further integrated with more applications and thus make the system more extensible.

Compared with traditional architectures, the cloud architecture achieves high utilization, high availability with low cost through allowing applications to share resource pools. The architecture contributes to rapid deployment of different algorithms, easy expansion, intelligent management of resources.

3.6 Job scheduling of parallel computation

In the section of the integration of models, a service-oriented cloud architecture is described and the cloud layer is introduced as an intermediate layer for dispatching resources for analysis tasks. To keep the whole system work smoothly and achieve high efficiency of the spatial analysis, reasonable job scheduling is critical.

All computing resources, including multi-computers that have tens to thousands of processors, which can support large diverse workloads of parallel jobs of spatial analysis, are divided into partitions, each of which consists of different numbers of processors. Partitions are devoted to service tasks or set aside to support interactive work through time-slicing. The pool of computing resource, which is mentioned above, is comprised of the partitions.

When analysis tasks are given to the system, the corresponding spatial domain is abstracted for estimating the computational intensity. The computational intensity serves as an indicator that approximately suggests how many computing resources are required, namely how many partitions are supposed to assigned to. Basically, the definition of computational intensity concerns the ratio of the number of computing operations executed or the number of memory accesses. However, due to the uneven distribution of spatial objects, the computational intensity is sensitive to spatial distributions and varies in different sub-areas in the spatial domain. Following the spatial principle of heterogeneous phenomena, it is necessary to decompose the original analysis to small jobs on different sub-areas and thus improve the overall efficiency. The original spatial domain is mapped to grids and the spatial distribution of computational intensity is obtained using region quadtrees and space filling curves. The spatial

distribution of computational intensity reflects sufficient information to support task-scheduling algorithms. According to values of intensity assigned to the grids, it is available to make optimal deployment of computing resources and scheduling of parallel operations. Task scheduling heuristics are exploited to balance the loads of computing among subdomains.

3.7 Interactive visualization based on HTML5

To illustrate the geo-data and the outputs of the spatial analysis models to users, especially nontechnical users, in comprehensive analytical tasks, an intuitive visualization is necessary. The system is supposed to highly integrate multi-source data and different analysis algorithms and offer unified interfaces. OSCAR implements its visualization using interfaces and components provided by HTML5. Based on the HTML5-based interfaces, we can present the indicated information, which may derive from numerous data source and be generated asynchronously, in a uniform visualization system. A visual analytic tool based on Canvas, a practical HTML5 component for illustrating 2D content, is embedded in the system to dynamically draw a set of digital maps and label the maps, showing explicit or implicit spatial patterns with the outputs or predictor variables. Various information, including geographical structures, original data of infection cases and results of spatial analysis models, are integrated in the canvas and thus assist users such as policy makers to deepen their understanding of the results of the analysis and generate hypotheses and ideas about what may be going on in both the macro level and the micro level. Besides HTML5, our proposed framework mainly uses the web based technique such as Openlayers, which is a completely free Open Source JavaScript tool, to provide heat maps visualizing the spatial distribution of disease, relevant variables and the risk of the disease. The heat maps of a specific disease can be created according to requests of users in real time.

As further support to the visualization, a number of web based visualization technologies, such as Scalable Vector Graphics (SVG), a language for building rich graphical content, d3.js, a JavaScript library for manipulating SVG objects, and Bootstrap, a front-end Web development framework, are exploited. All of the intermediate data used for painting graphs are stored in Java-Script Object Notation (JSON), a lightweight web-friendly data-store and data-interchange format. With the standardized format, different analysis results are well assembled in the same visualization system. The unified interfaced and unified format of intermediate data greatly facilitate the processing for drawing heat maps.

4 Experimental implementation

Since the first version was issued in 2015, OSCAR has showed its advantage in data aggregation and comparative analysis for various public health emergency management issues, especially spreads of infectious diseases. Cases of infectious diseases have been systematically analyzed through OSCAR with different algorithms and models, thus resulted in more comprehensive understanding of the variation tendency of the diseases. Typical statistical models, GIS analytical tools and specified models with spatio-temporal structures were contained in the model library. Allowing for spatial principles, the system provided less biased estimations intuitive visualizations for users. Herein two cases of public health issues, human rabies and H7N9 in China, are presented to demonstrate the applicability of OSCAR. The two cases illustrate the workflow of the framework, showing

how data integration and model integration synergistically optimize the spatial analysis of public health issues.

4.1 Human rabies analysis in case view

Human rabies, a widely distributed zoonotic infectious disease which is estimated to cause hundreds of thousands of fatalities each year worldwide [40, 58, 59], was investigated using spatial analysis under our framework. Phylogenic analysis of Chinese rabies viruses from 1969 to 2009 has demonstrated that infection had been transmitted intra-provincially and extra-provincially due to human-related activities. Based on statistical data, many existing works suggested that numerous factors present explicit or implicit correlations with the incidence of the disease. However, due to the lack of access to sufficient data of cases, especially variables of social environmental factors, existing works have seldom conducted analysis from the view of cases. As many as dozens of factors may affect the distribution and the spread of human rabies while most of them only have limited access. Table 1 lists major factors that are considered to have direct or indirect effects on the incidence of human rabies, including environmental variables, socio-economic variables and transportation variables. As is shown in the table, data of the factors have been stored and managed in specific databases and researchers often have to request them through specific data sources. The sources of data for spatial analysis are so multiple and decentralized that it was far from readily for researchers to incorporate different factors into their analyses or make quick response with comprehensive considerations. Moreover, processing on data of some factors requires expert knowledge and specific tools or software. For instance, data of NDVI, which reflects local environment status and food supply for animals that may carry viruses, can hardly be directly measured through statistical methods. NDVI data has high requirements on the accuracy of spatial dimension and temporal dimension and therefore need specific processing such as Kriging interpolation. Instead of statistical measure, NDVI is often obtained through spatial computation on remote sensing images. The specific processing calls for expert experience and adds to the obstacles for general researchers to conduct comprehensive analyses involving NDVI data. Furthermore, the amount of required remote sensing data and that of NDVI data are often extremely large. Data of NDVI and land cover in China obtained via spatial computing on raw data from MODIS, a common data source, can be more than 1 T bytes. Retrieval operation on so large a data set is of significant difficulty. When researchers request data with the location and the date specified, how to find the data efficiently is a great challenge. To improve the efficiency, proper data organization is necessary.

4.1.1 Data aggregation on-demand

OSCAR was motivated by demands for multi-source data, which could support comprehensive analyses, and further data integration. OSCAR automatically collected and integrated data to cover most of above factors. Then the massive data were well organized through spatio-temporal data cube, which greatly optimized operations on data with spatio-temporal structures. OSCAR provided sufficient data of cases, environmental variables, socio-economic variables and other relevant data to perform spatial analytical approaches.

Moreover, OSCAR provided HTML5-based visualization tools for users to present data and interact with the system. Owing to advantages of HTML5, the visualization was highly interactive and flexible. Different from traditional visualizations, which illustrated data through

Table 1 An example of major data sources with direct or potential effects on the spread of human rabies

Category	Description of dataset	Abbreviation	Unit	Data source
Environmental variables	digital elevation	DEM	m	USGS
	digital slope	SLOPE	degree	USGS
	Average Temperature	AT	$^{\circ}C$	MODIS
	Human Population Density 2000	POPDENS	p/km^2	National Statistics Bureau
Social-economic variables	Human Population Density 2005	POPDENS	p/km^2	National Statistics Bureau
	Human Population Density 2010	POPDENS	p/km^2	National Statistics Bureau
	Ratio of illiteracy	ROI	p/million	National Statistics Bureau of China
	Ratio of middle school and above	RMS	p/million	National Statistics Bureau of China
	Yearly Gross Domestic Production	GDP	10^4RMB	National Statistics Bureau of China
	early per Capital Gross Domestic Production	PCGDP	10^4RMB	National Statistics Bureau of China
	Distance to road network	DTRN	km	National Administration of Surveying, Mapping and Geoinformation
	Distance to city centre	DTCC	km	National Administration of Surveying, Mapping and Geoinformation
	Distance to county centre	DTCNC	km	National Administration of Surveying, Mapping and Geoinformation
	Distance to nearest hospital	DTHSP	km	China's Health and Family Planning Commission
Transportation variables	Distance to nearest clinic	DTCLC	km	China's Health and Family Planning Commission
	Minimum Spatio-temporal distance to nearest case	MSTDNC	Km/day	China CDC Rabies Surveillance data
	Minimum Spatial distance to nearest case	MSDNC	km	China CDC Rabies Surveillance data
	Minimum Temporal distance to nearest case	MTDNC	day	China CDC Rabies Surveillance data

static images, OSCAR provided access for users to dynamically adjust resolution and other parameters to conduct visualization analyses at different granularities and different hierarchies. Fig. 6 shows the visualization page of OSCAR. Users of OSCAR could select those variables as input data and exploit services like data retrieval and geocoding to efficiently process the data. The integration of multi-source data, which cover different kinds of variables, could contribute to deeper understanding of traits of the disease and reduce bias derive from the insufficiency of considered factors. In Fig. 6, the visualization page displays numbers of rabies notifications from 2005 to 2013. A downtrend of the count of cases, which first occurred in 2008, is represented in the figure. However, the annual number of cases remains in a high level and the corresponding damage cannot be simply ignored. The page represents the spatial distribution of dog rabies cases in China. According to the scatterplots, cases appear more in eastern and southern part of China.

Data integration harvested available information for the spatial analysis. However, expansion of features for regression brought challenges to the regression models while boosting the analysis. The expansion itself often fluctuates the performance of regression models, i.e. raising typical evaluations such as R^2 regardless of whether the added features actually contribute to the regression. More importantly, a cumbersome model with a number of features would make it elusive to distinguish definitive factors from those have slight effects. To catch more critical factors, feature selection was necessary. Multiple backward stepwise logistic regression was carried out to select the significant explanatory variables. The stepwise process was repeated 1000 times using different training subsets. The top 20 regression models are picked out and the corresponding variables were ranked by AIC and mean *P*-value. Variables yielding non-significant effects (mean P-value >0.05) were discarded.

The remained features mainly included: the longitude, the average temperature, the distance to county center, the distance to road network and the minimum spatial distance to the nearest case.

4.1.2 Model integration

Existing spatial analyses of epidemics had two limitations. The first limitation was that those analyses generally focused on one or two algorithms or models and spared most of their efforts

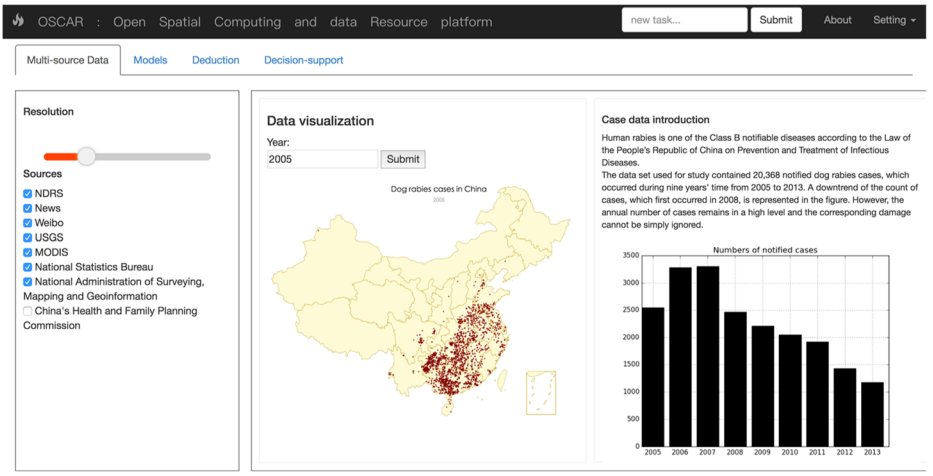

Fig. 6 The interactive analysis interface of OSCAR

to optimize the application of the models instead of bagging different models to perform a comprehensive analysis. Integration of different models was rare in existing spatial analyses on public health issues. The integration indeed could boost the spatial analysis through comparison between effects of different models. Based on the comparison, researchers could explore public health issues from various perspectives and find models that were relatively more ideal for their issues. Another limitation was that existing works could hardly trace the provenance of their processing of analyses. Retrospective analysis was of required to make the selection of models and adjustments of parameters more open and transparent, thus reveal the effects of the adjustments and lead to deeper understandings of the processing. OSCAR was designed to handle the problems discussed above. In order to support comprehensive spatial analyses and trace the processing of analyses, OSCAR integrated multi-source data with different models and incorporated cloud computing architecture to control data flows.

For the case study on human rabies, numerous tools and models were utilized to implement all phases of the spatial analysis. In the first phase, researchers had to verify the spatial autocorrelation, which served as the precondition of models such as SEM. In order to verify the spatial autocorrelation, Moran's I, which is widely used in evaluation of spatial autocorrelation, is adopted. The Moran's I from 2005 to 2013 is illustrated in Fig. 7. The global index is statistically significant and provides evidence of spatial autocorrelation in China. With the spatial autocorrelation demonstrated, effectiveness of conventional regression models like ordinary least squares (OLS) might be suspected for bias brought by spatial interactions. Therefore, a comprehensive analysis integrates outputs from different models, especially models with spatial specifications, would be more reliable. The graph also represents the downtrend of Moran's I, which may be a consequence of the continuous drop of rabies cases in China from 2008.

With the spatial autocorrelation verified, regression models were exploited to explore correlations between the incidence of human rabies with relevant factors and then predict possible outbreaks. Individual models, algorithms, analytic software and more services were integrated in the model library and organized as task-oriented. Components of the model library were separately deployed and each component implemented not only the spatial model

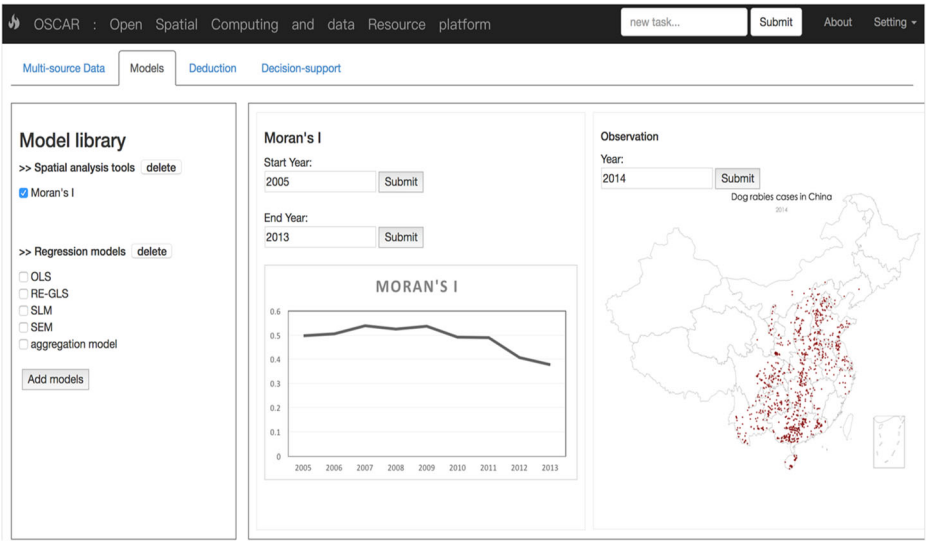

Fig. 7 The screenshot of OSCAR analysis: Moran's I of human rabies in China from 2005 to 2013

but also associated functions such as specified processing on the input data and deployments of available computational resources. To forecast possible outbreaks of dog rabies, regression models were extracted from the model library to predict counts of rabies cases in different regions of China. The predicted count of possible cases in an individual region reflected the corresponding risk of the disease, which is supposed to be taken into account for disease preventions. Different regression estimators were adopted and integrated in the analysis. Each estimator was based on a specific assumption and focus on some characteristics of the distribution of the data. In the case of dog rabies in China, we specially allowed for individual heterogeneity and spatial effects, which were verified by Moran's I and had significant influence on the estimation. Estimators included but were not limited to:

(1). A fundamental OLS estimator, which derived from pooled OLS equations. The basic OLS model concentrated on measuring the associations between dependent variable and explanatory variables with the heterogeneity and spatial autocorrelation ignored.
(2). A random effects-generalized least squares (RE-GLS) estimator. The RE-GLS estimator incorporated individual heterogeneity with respect to spatial variation among different regions.
(3). Estimator of spatial lag model (SLM). SLM, namely spatial autoregressive model, assumed that the dependent variable followed a spatial autoregressive process and concerned spatial substantive dependence.
(4). Estimator of spatial error model (SEM). SEM, also called spatial autocorrelation model, simulated the spatial dissemination of random effects from factors that were not covered by the set of explanatory variables. Like SLM, SEM also assumed a spatial autoregressive process while it suggested spatial error dependence among spatial objects.
(5). A feasible generalized least squares (FGLS) estimator proposed by Baltagi and Pirotte, which allowed for both individual heterogeneity and spatial effects. The BP-FGLS estimator was built on panel data and considered interactions among spatial objects like spatial spillover effect.

To support models with residuals or error terms that depicted spatial effects, general method of moments (GMM), an approach that can provide unbiased estimations for those models, was imported from the model library. Due to the cloud architecture, different estimators worked on different nodes of the system at the same time. Each model was assigned with exclusive control of computational resources and services. The exclusive deployment of resources was managed by the job scheduling system. Accordingly, the outputs had no interference to each other and performed a comprehensive result together.

Due to the integration of multi-source data, which provide information from different perspectives, the system gained ability to simultaneously apply different models to conduct an overall analysis. The input of the regression models was a balanced panel that contains 271 cross sections with 9 time periods. Then 2439 samples were generated, each of which consisted of the normalized incidence and explanatory variables of an individual section in a specified year. The result was illustrated in the visualization pages. Figure 8 shows the scatterplots of observed counts against predicted counts. Users could select the model with best performance on temporal data or comprehensively take all result as references.

Figure 9 shows the prediction of integrated model for dog rabies cases in 2014. The color depth denotes the possible extent of the disease and darker colors suggest more possible cases in the corresponding regions. Compared with the scatterplot of observed cases in 2014, it can be found that the prediction well fit the observation in most of regions in China.

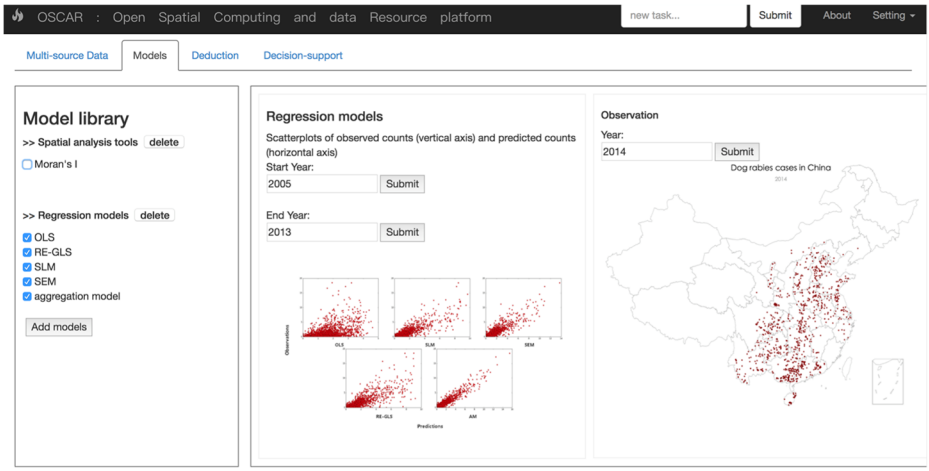

Fig. 8 Scatterplots of observed counts (vertical axis) and predicted counts (horizontal axis) of different models

Although models control for the individual heterogeneity and spatial autocorrelation perform better than rough models like OLS, they may require more computing resources and become more time-consuming. To ensure the fast response of the system, predictive results of all models, including models that can output predictions in a very short time but may lose some information from the original data, are given in succession to support the users. Decision-makers are available to begin analysis with results from fast models and continuously make accurate adjustments based on results from complex models, which fully utilize the inputted data but may take more time.

4.2 Distribution and spread of H7N9 in China

In 2013, H7N9 has stricken eastern provinces of China and cause twenty cases in less than one month. Most human cases of infection have occurred in four major Chinese cities: Shanghai, Hangzhou, Huzhou, and Nanjing. Soon after the initial outbreak, the government began to

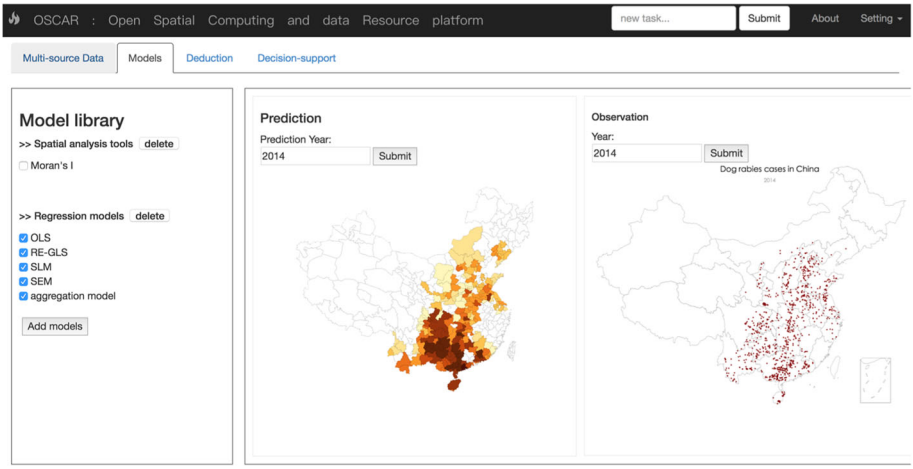

Fig. 9 Predicted counts of human rabies cases in 2014

close LPMs in April, 2013, as a precautionary measure. However, decision-making of measures to control and prevent the disease, including closure of LPMs, was facing challenges then. On the one hand, the spread of H7N9 was so fast that the government and medical staffs had to make response as quick as possible to avoid more human infections. An effect but rough solution was to temporarily close all LPMs. However, on the other hand, closure of LPMs and culling of poultries would cause great financial damages to traders and farmers. The social influence of the measure was complex and could not be ignored. Therefore, the object of the case study was to verify and estimate the direct correlation between H7N9 and LPMs. If the direct correlation was verified and well quantified, accurate decision-makings would be available.

To support the study on LPMs, data of LPMs, especially locations of LPMs, were necessary. However, none of official authorities could provide those data at 2013. To collect data of LPMs, OSCAR timely launched distributed crawlers to search relevant data from web map services like Openstreet map, Baidu Map and AutoNavi Map. Then OSCAR added unified geographical coordinates to detected LPMs using geo-coding. As the data were collected from multiple sources, OSCAR merged same LPMs based on the similarity between information of LPMs from different map services. The similarity was measured by the name, the longitude and the latitude of LPMs and LPMs that were highly similar to each other were identified as one LPM. Such specific processing, including geo-coding and unified mapping of ontologies, was beyond functions GIS could provide. OSCAR well satisfied the demands for dynamic acquisition of specified data and data consolidation.

4.2.1 Influenza a H7N9 data

Every confirmed human case of avian influenza A H7N9 virus infection reported in the four cities by June 7, 2013 were extracted from the database of China CDC. Information of LPMs was obtained from official websites using automatic crawlers. All data were mapped into a uniform geographic coordinate system and well organized in spatio-temporal cubes.

Within the 130 individuals with avian influenza A H7N9 virus infection reported by June 7, 2013, 123 (95%) had serious disease with pneumonia and 44 (34%) had died. 85 (65%) were identified in Shanghai, Hangzhou, Huzhou, and Nanjing. As 88 of the 95 cases with illness onset confirmed by April 12 were identified in or near these areas, the four cities were marked as the epicenter. Of the 85 confirmed cases in the four cities, 67 were urban residents and 18 were residents of suburban townships neighboring but outside these cities, who were nonetheless treated in tertiary referral centers in the cities. It was obvious that urban areas suffered most in the influenza. For poultry species, the urban environments were not suitable for living or propagation. LPMs, where poultries were artificially clustered, were considered as the major channel through which poultries, including virus carriers, had been brought to the urban areas.

4.2.2 Deduction of LPM closure

To quantify the direct correlation between H7N9 and LPMs and further balance effects of closure of LPMs, including both the effect on the control and prevention of H7N9 and the financial influence on farmers and traders, simulation of the spread H7N9 and closure of LPMs was needed. The simulation concerned the range and the intensity of the spread of H7N9 as well as concrete arrangement of the closure of LPMs. The closure of LPMs might follow different principles and thus perform different effects. However, GIS alone failed to provide

the function of simulation. Researchers of public health issues also found it difficult to build a specific platform to conduct the simulation and efficiently manage such platform. To achieve a efficient and comprehensive simulation, OSCAR integrated 16 CPUs and implemented a parallel computing environment to support public health researchers. The system provided Python interfaces to call models for spatial analyses and GUI for real-time visualization. In less than one week, OSCAR performed more than 1000 times simulations for researchers to estimate impacts from the spatial distribution and heterogeneity of LPMs. Based on the simulation, closure of LPMs and relevant effects were deducted. For policy makers or medical staff, the spatial deduction on LPMs was of considerable significance as it could help refine controlling measures, suggesting where to control or close certain LPMs.

The deduction was conducted to directly generate instructions for the control of the influenza. The orientation was to assess the effect of LPM closure using existing records and make deductions about the precautionary measure. To assess the effect of LPM closure on the number of human cases of avian influenza A H7N9 virus infection, a Bayesian statistical model was adopted to perform a population-based before-and-after analysis. The Bayesian model was extracted from the model library and assigned with sufficient computing resources. Reported cases in Shanghai, Hangzhou, Huzhou, and Nanjing were used for the analysis. Based on the report data, time series of onset of illness were generated for visual analysis. Through synthesizing the time series with records of LPM closure, researchers readily caught the correlation between the measure and the spread of H7N9.

In the deduction, the posterior probabilities of infection with LPM open and the posterior probabilities of infection with LPM closed were calculated using the records. The probability that denoted the estimated incidence was described by the number of possible infectious cases. Figure 10A shows the daily records while Fig. 11B illustrates the posterior estimations of the number of daily infectious cases outputted by the Bayesian statistical model. Based on the indication from the model, general pattern was found that the closure of LPM presented significant relationship with the influenza incidence. Within the model, when LPM closures were simulated at certain dates, the estimations in the four cities rapidly declined. The

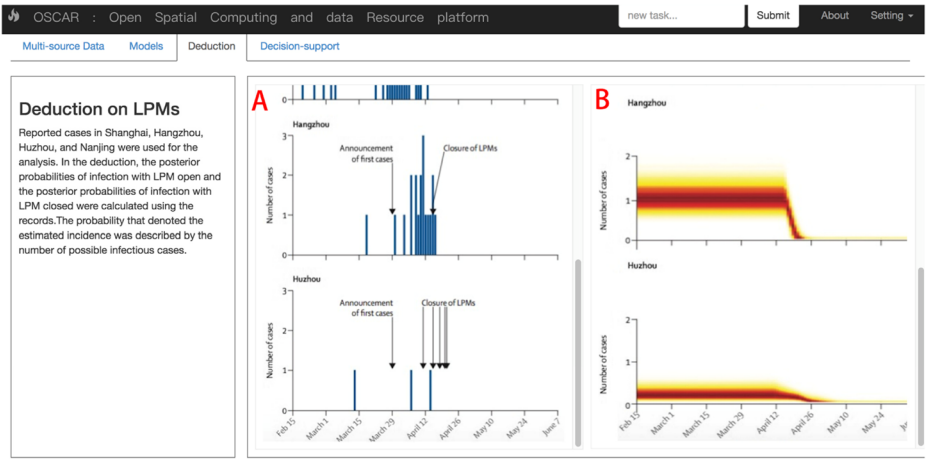

Fig. 10 Number of H7N9 Case Onset by date [3]. (**A**) Illness onset dates of 60 cases of avian influenza A H7N9 virus infection between Feb 19, and April 16, 2013. LPMs in the five districts in Huzhou were closed on different dates. (**B**) Posterior estimates of the mean daily number of people with illness onset. Darker colours indicate regions with higher posterior probability. LPM = live poultry market

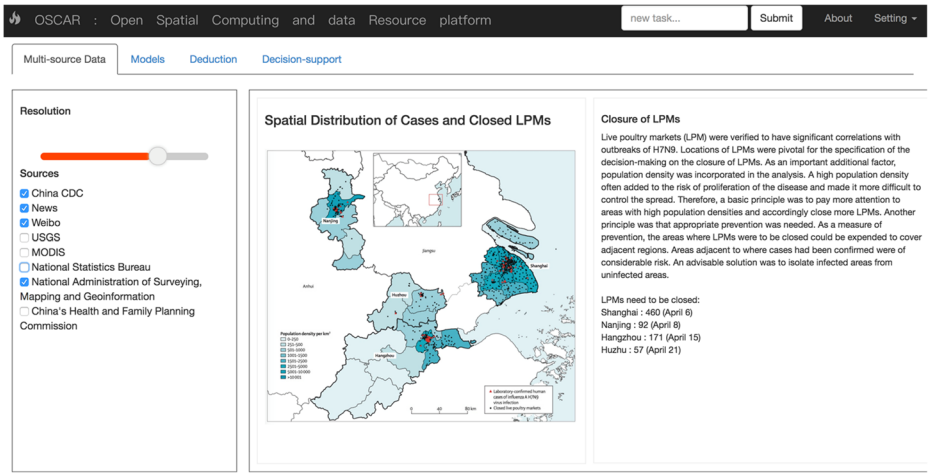

Fig. 11 Location of laboratory-confirmed cases of influenza A H7N9 virus infection and closures of live poultry markets in Nanjing, Huzhou, Hangzhou, and Shanghai [3]

estimations well matched the observed trend in Fig. 11A as the reported cases reduced after the closures of LPMs. The deduction suggested that LPMs played a critical role in the spread of the influenza in urban areas. The deduction was supported by more than 60 confirmed cases. One to three cases were confirmed almost every day in Shanghai and Hangzhou in the week before LPMs were closed (Fig. 10A). In Nanjing, one to two cases were reported every few days before LPM closures, and in Huzhou, only two cases were reported before closures (Fig. 10A). However, few new cases with onset dates after closures were officially announced by June 7 (Fig. 10A). In a sensitivity analysis, the model suggested that LPM closure reduced incidence by 81% in Nanjing, 93% in Shanghai, 92% in Hangzhou, and 91% in Huzhou.

Furthermore, locations of LPMs to be closed were pivotal for the specification of decision-makings. To support concrete decision-makings, visualization analysis, which presented intuitive interfaces for researchers, was conducted. As is shown in Fig. 11, the interface illustrated the geospatial information of the four cities. As an important additional factor, population density was incorporated in the analysis and depicted on the map. When the disease broke out, population density could never be neglected. A high population density often added to the risk of proliferation of the disease and made it more difficult to control the spread. Therefore, a basic principle was to pay more attention to areas with high population densities and accordingly close more LPMs. This was also consistent with the spatial principle that concerns individual heterogeneity.

Another principle was that appropriate prevention was needed. As a measure of prevention, the areas where LPMs were to be closed could be expended to cover adjacent regions. According to the spatial principle, closer things have more correlations between each other than farther away and dynamic processes often happen among neighbor objects, which lead to exchanges and interactions between the objects. Therefore, areas adjacent to where cases had been confirmed were of considerable risk. The detailed interactions were hard to track while they might result in further spread of the influenza. An advisable solution was to isolate infected areas from uninfected areas. The range of the closure was expanded and LPMs in areas adjacent to infected areas were also closed. Concerted with the deduction, Fig. 11 shows the spatial distribution of closed LPMs and confirmed cases. Local authorities closed the 460 live poultry

markets (LPM) in Shanghai on April 6, and the 92 in Nanjing on April 8. Most of the 171 LPMs in Hangzhou were closed on April 15, and the remainder on April 24. The 57 LPMs in the five districts of Huzhou were sequentially closed between April 11 and 21. Overall, 780 LPMs were closed, depopulated, and disinfected in the four cities. Pet bird trade was also suspended.

5 Conclusions

In this paper, we proposed a comprehensive framework for spatial analysis on public health issues. The framework exploit open source data to help find newly occurred cases of diseases, which serve as timely additions to official data released by governments and relevant institutions. The system automatically crawls web data and organizes data effectively using a data warehouse with spatio-temporal cubes. The organization of data well retains the time series information and geographical structure among the collected data. Based on the specific processing on the data, we can conduct analyses for public health issues through temporal spatio-temporal dimensions with respect to complex interactions between neighboring spatial objects. To give analyses through different perspectives, various algorithms and models are embedded and integrated in model library, where individual models can be separately called and work without conflicts. Due to the integration of data and integration of models, the system lessons the limitation for data or resources and makes it available for researchers to implement different types of models as needed. As a decision-support system, OSCAR predicts possible spread of diseases as references for further analysis and incorporate spatial principles to enhance the performance. A balance between efficiency and effect is also considered. The system is designed to provide fast results as well as one-step services for relevant decision makers. Then a case study of dog rabies, which verified the validity of our system, is illustrated. The case study demonstrates that data from different sources, in different types can be converged using our framework and organized with spatio-temporal structure well kept. Based on the integration of data, numerous algorithms and models are supported. Different models are stored in model library and deployed separately. Herein different models can work together and provide comparable results. In the spatial analysis for dog rabies, the specification of models lies on the expansion of collected data and the spatio-temporal organization of data, where spatial and temporal dimensions are sufficiently kept to enable us to integrate spatial autoregressive process into the analysis. Finally, the distributions of observations and predictions for possible infection cases are illustrated on maps as an intuitive visualization. Another case study, which is conducted using data of H7N9 cases in China, shows how to make inductions and deductions through spatial analysis using the system. The case focus on LPM closure, which is used as a concrete measure to control and prevent the influenza. Through integration of case data and statistical models, the system contributes to indicating effects of LPM closure on the spread of the disease. Furthermore, following spatial principles, the deduction can be used to support specific implementation of the measure. The framework well organizes spatial data and provides researchers and decision makers analytic tools and intuitive graphs for comprehensive decision-making. Our framework is built on distributed database, distributed computing and a service-oriented cloud architecture, therefore it gains better flexibility, extendibility, and ability to respond fast with large amounts of spatial data during emergency management for public health issues, and highly integrate different data and algorithms for spatial analysis.

However, more verifications and further improvements are also found necessary. Although our approach obtained promising results on the dataset of dog rabies cases in China, it still needs to be verified on more different dataset. Moreover, the system is expected to use machine learning

meta-algorithms for further enhancements on the aggregation model. In the future, we would continually improve the framework and test it on various datasets.

Fundings This work is partly supported by Natural Science Foundation of China under Grant No.41371386 and Beijing Natural Science Foundation under Grant No. 9172023. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Li K, Guan Y, Wang J, Smith G, Xu K, Duan L, Rahardjo A, Puthavathana P, Buranathai C, Nguyen T (2004) Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. Nature 430(6996):209–213
2. Chen H, Smith G, Zhang S, Qin K (2005) H5N1 Virus outbreak in migratory waterfowl. Nature 436(7048):191
3. Yu H, Wu JT, Cowling BJ, Liao Q, Fang VJ, Zhou S, Ni MY (2014) Effect of closure of live poultrymarkets on poultry-to-person transmission of avian influenza A H7N9 virus: an ecological study. The Lancet 383(9916):541–548
4. Sallis JF, Frank LD, Saelens BE, Kraft MK (2004) Active transportation and physical activity: opportunities for collaboration on transportation and public health research. Transp Res A Policy Pract 38(4):249–268
5. Rotz LD, Khan AS, Lillibridge SR, Ostroff SM, Hughes JM (2002) Public health assessment of potential biological terrorism agents. Emerg Infect Dis 8(2):225
6. Wilson ME (1995) Travel and the emergence of infectious diseases. Emerg Infect Dis 1(2):39
7. Martens P, Hall L (2000) Malaria on the move: human population movement and malaria transmission. Emerg Infect Dis 6(2):103
8. Morens DM, Folkers GK, Fauci AS (2004) The challenge of emerging and re-emerging infectious diseases. Nature 430(6996):242–249
9. Lederberg J (2000) Infectious history. Science 288(5464):287–293
10. Dobson AP, Carper ER (1996) Infectious diseases and human population history. Bioscience 46(2):115–126
11. Weiss RA, McMichael AJ (2004) Social and environmental risk factors in the emergence of infectious diseases. Nat Med 10(12s):S70
12. Dugas AF, Hsieh Y-H, Levin SR, Pines JM, Mareiniss DP, Mohareb A, Gaydos CA, Perl TM, Rothman RE (2012) Google flu trends: correlation with emergency department influenza rates and crowding metrics. Clin Infect Dis 54(4):463–469
13. Lazer D, Kennedy R, King G, Vespignani A (2014) The parable of Google flu: traps in big data analysis. Science 343(6176):1203–1205
14. Ortiz JR, Zhou H, Shay DK, Neuzil KM, Fowlkes AL, Goss CH (2011) Monitoring influenza activity in the United States: a comparison of traditional surveillance systems with Google flu trends. PLoS One 6(4):e18687
15. Corley CD, Cook DJ, Mikler AR, Singh KP (2010) Text and structural data mining of influenza mentions in web and social media. Int J Environ Res Public Health 7(2):596–615
16. McCarthy FX (2010) FEMA's disaster declaration process: a primer. DIANE Publishing, Collingdale
17. Abrahams J (2010) Disaster management in Australia: the national emergency management system. Emerg Med 13(2):165–173
18. Iakovou E, Douligeris C (2001) An information management system for the emergency management of hurricane disasters. Int J Risk Assess Manag 2(3):243–262(220)
19. Rodrigues AS, Santos MA, Santos AD, Rocha F (2002) Dam-break flood emergency management system. Water Resour Manag 16(6):489–503
20. Yang C, Yu M, Hu F, Jiang Y, Li Y (2017) Utilizing Cloud Computing to address big geospatial data challenges. Comput Environ Urban Syst 61:120–128
21. Kun Y, Quan-Li X, Shuang-Yun P, Yan-Bo C (2006) The design and implementation of urban earthquake disaster loss evaluation and emergency response decision support systems based on ArcGIS. In:IEEE International Conference on Geoscience and Remote Sensing Symposium, 2006. IGARSS 2006, pp. 892–895
22. Li YY, Shang XQ, Liu RF (2012) GIS-based emergency management system for chemical Industry Park. Adv Mater Res 550(553):2941–2944
23. Rashed T, Weeks J (2003) Assessing vulnerability to earthquake hazards through spatial multicriteria analysis of urban areas. Int J Geogr Inf Sci 17(6):547–576
24. Shuai XH, Cheng XP, Jiang LX (2001) Earthquake emergency response information system based on ArcView. Earthquake 21(4):94–99

25. Tanasescu V, Gugliotta A, Domingue J, Davies R, Gutiérrez-Villarías L, Rowlatt M, Stincic S (2006) A semantic web services GIS based emergency management application. In: International Semantic Web Conference, pp. 959–966
26. Tzemos S, Burnett RA (1995) Use of GIS in the federal emergency management information system (FEMIS) (No. PNL-SA–26086; CONF-9505242–1). Pacific Northwest Lab., Richland
27. Zhigang L (2010) Liangtian, Wunian Y: research of GIS-based urban disaster emergency management information system. Comput Commun Technol Agric Eng Int Confe 2:484–487
28. Bailey MJ, Dynarski SM (2011) Gains and gaps: changing inequality in US college entry and completion (No. w17633). National Bureau of Economic Research, Washington DC
29. Manole A, Fratta P, Houlden H (2014) Recent advances in bulbar syndromes: genetic causes and disease mechanisms. Curr Opin Neurol 27(5):506–514
30. Hong-zhi P, Ling-bin Y, Yong-shun H (2007) Design and development of emergency and precaution management of public health system based on GIS [J]. Sci Surv Mapp 3:045
31. Nelson EK, Piehler B, Eckels J, Rauch A, Bellew M, Hussey P, Ramsay S, Nathe C, Lum K, Krouse K (2011) LabKey server: an open source platform for scientific data integration, analysis and collaboration. BMC Bioinf 12(1):71
32. Meng S, Xu G, Wu X, Lei Y, Yan J, Nadin-Davis SA, Liu H, Wu J, Wang D, Dong G et al (2010) Transmission dynamics of rabies in China over the last 40 years: 1969–2009. J Clin Virol 49(1):47–52
33. Gong W, Jiang Y, Za Y, Zeng Z, Shao M, Fan J, Sun Y, Xiong Z, Yu X, Tu C (2010) Temporal and spatial dynamics of rabies viruses in China and Southeast Asia. Virus Res 150(1–2):111–118
34. Yu J, Li H, Tang Q, Rayner S, Han N, Guo Z, Liu H, Adams J, Fang W, Tao X et al (2012) The spatial and temporal dynamics of rabies in china. PLoS Negl Trop Dis 6(5):e1640–e1610
35. Yin C, Zhou H, Wu H, Tao X, Rayner S, Wang S, Tang Q, Liang G (2012) Analysis on factors related to rabies epidemic in China from 2007-2011. Virol Sin 27(2):132–143
36. Zhang J, Jin Z, Sun GQ, Sun XD, Ruan SG (2012) Modeling seasonal rabies epidemics in China. Bull Math Biol 74(5):1226–1251
37. Zhang J, Jin Z, Sun G, Zhou T, Ruan S (2011) Analysis of rabies in China: transmission dynamics and control. PLoS One 6(7):e20891–e20899
38. Song M, Tang Q, Wang D-M, Mo Z-J, Guo S-H, Li H, Tao X-Y, Rupprecht C, Feng Z-J, Liang G-D (2009) Epidemiological investigations of human rabies in China. BMC Infect Dis 9(1):210–218
39. Yin W, Dong J, Tu C, Edwards J, Guo F, Zhou H, Yu H, Vong S, Rabies T, Advisory B (2013) Challenges and needs for China to eliminate rabies. Infect Dis Poverty 2(1):23
40. Guo D, Zhou H, Zou Y, Yin W, Yu H, Si Y, Li J, Zhou Y, Zhou X, Magalhães RJS (2013) Geographical analysis of the distribution and spread of human rabies in China from 2005 to 2011. PLoS One 8(8):e72352
41. Smith DL, Lucey B, Waller LA, Childs JE, Real LA (2002) Predicting the spatial dynamics of rabies epidemics on heterogeneous landscapes. Proc Natl Acad Sci U S A 99(6):3668–3672
42. Lucey BT, Russell CA, Smith D, Wilson ML, Long A, Waller LA, Childs JE, Real LA (2002) Spatiotemporal analysis of epizootic raccoon rabies propagation in Connecticut, 1991-1995. Vector Borne Zoonotic Dis (Larchmont, NY) 2(2):77–86
43. Guerra MA, Curns AT, Rupprecht CE, Hanlon CA, Krebs JW, Childs JE (2003) Skunk and raccoon rabies in the eastern United States: temporal and spatial analysis. Emerg Infect Dis 9(9):1143–1150
44. Wilde H, Khawplod P, Khamoltham T, Hemachudha T, Tepsumethanon V, Lumlerdacha B, Mitmoonpitak C, Sitprija V (2005) Rabies control in south and Southeast Asia. Vaccine 23(17–18):2284–2289
45. Cliquet F, Picard-Meyer E (2004) Rabies and rabies-related viruses: a modern perspective on an ancient disease. Rev Sci Tech 23(2):625–642
46. Recuenco S, Blanton JD, Rupprecht CE (2012) A spatial model to forecast raccoon rabies emergence. Vector Borne Zoonotic Dis 12(2):126–137
47. Li X, Feng P, Du Y, Bian G, Yu X (2010) Socioeconomic status is a critical risk factor for human rabies post-exposure prophylaxis. Vaccine 28(42):6847–6851
48. Haupt W (1999) Rabies–risk of exposure and current trends in prevention of human cases. Vaccine 17(13–14):1742–1749
49. Fang LX, Ping F, Ping DY, Hui BG, Yan YX (2010) Socioeconomic status is a critical risk factor for human rabies post-exposure prophylaxis. Vaccine 28(42):6847–6851
50. Gautret P, Shaw M, Gazin P, Soula G, Delmont J, Parola P, Soavi MJ, Brouqui P, Matchett DE, Torresi J (2008) Rabies Postexposure prophylaxis in returned injured travelers from France, Australia, and New Zealand: a retrospective study. J Travel Med 15(1):25–30
51. Si H, Guo Z-M, Hao Y-T, Liu Y-G, Zhang D-M, Rao S-Q, Lu J-H (2008) Rabies trend in China (1990-2007) and post-exposure prophylaxis in the Guangdong province. BMC Infect Dis 8
52. Suzuki K, Pereira JAC, López R, Morales G, Rojas L, Mutinelli LE, Pons ER (2007) Descriptive spatial and spatio-temporal analysis of the 2000–2005 canine rabies endemic in Santa Cruz de la sierra, Bolivia. Acta Trop 103(3):157–162

53. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C (2012) Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28(12):1647–1649
54. Guo D, Wu K, Zhang Z, Xiang W (2012) WMS-based flow mapping services. In: 2012 IEEE Eighth World Congress on Services (SERVICES) IEEE Honolulu, pp. 234–241
55. Guo D, Li J, Cao H, Zhou Y (2014) A collaborative large spatio-temporal data visual analytics architecture for emergence response. IOP Conference Series: Earth and Environmental Science 18(1):012129. https://doi.org/10.1088/1755-1315/18/1/012129
56. Yang C, Goodchild M (2011) Using spatial principles to optimize distributed computing for enabling the physical science discoveries. Proc Natl Acad Sci U S A 108(14):5498–5503
57. Gray J, Chaudhuri S, Bosworth A, Layman A, Reichart D, Venkatrao M, Pellow F, Pirahesh H (1997) Data cube: a relational aggregation operator generalizing group-by, cross-tab, and sub-totals. Data Min Knowl Disc 1(1):29–53
58. Chomel B (1993) The modern epidemiological aspects of rabies in the world. Comp Immunol Microbiol Infect Dis 16(1):11–20
59. Zhang Y, Xiong C, Xiao D, Jiang R, Wang Z, Zhang L, Fu Z (2005) Human rabies in China. Emerg Infect Dis 11(12):1983–1984

Danhuai Guo is associate professor at Computer Network Information Center, Chinese Academy of Sciences and the University of Chinese Academy of Sciences. His research interest includes High performance spatial computing, spatial data visualization, GIS and Public Health, and Spatial temporal Data Mining.

Yingqiu Zhu is a Postgraduate student at Computer Network Information Center, Chinese Academy of Sciences. His research interest includes Business Intelligence and Data Mining.

Wenwu Yin is the medical doctor of Chinese Center of Disease Control and preventation. His research cover the epidemic of zoonotic disease, and prevention of infection disease.