Informational Measures of Aggregation for Complex Systems Analysis[*]

Robin Lamarche-Perrin[1] and **Jean-Marc Vincent**[2] and **Yves Demazeau**[3]
Laboratoire d'Informatique de Grenoble, France
[1]Université de Grenoble, [2]Université Joseph Fourier, [3]CNRS
{Robin.Lamarche-Perrin,Jean-Marc.Vincent,Yves.Demazeau}@imag.fr

Abstract

The analysis of systems' dynamics lies on the collection and the description of events. In order to scale-up classical analysis methods, this paper is interested in the reduction of descriptional complexity by aggregating events' properties. *Shannon entropy* appears to be an adequate complexity measure regarding the aggregation process. Some other informational measures are proposed to evaluate the qualities of aggregations: *Entropy gain*, *information loss*, *divergence*, *etc.* These measures are applied to the evaluation of geographic aggregations in the context of news analysis. They allow determining which abstractions one should prefer depending on the task to perform.

Keywords: Data aggregation, macroscopic descriptions, news analysis, Shannon entropy, information loss, Kullback-Leibler divergence.

1 Introduction

The analysis of systems' dynamics can fulfill various purposes (*e.g* representation, explanation, prediction). It relies on a specific knowledge regarding the systems' events and it benefits from a precise amount of resources (either computational or cognitive). The difficulty of an analysis depends on the adequacy between (1) the task to perform [3], (2) the knowledge to handle and (3) the available resources. Therefor, in case of large-scale complex systems, classical analysis methods may be hard to scale-up.

In order to maintain the adequacy between data and resources, this paper proposes a formal process of *data aggregation* in order to produce scalable macroscopic descriptions out of microscopic knowledge: This data account of complexity so focuses on the *difficulty of description* of a system [12]. As for other relativist accounts [7; 3; 6], it never directly deals with the *system's inner complexity* (size, heterogeneity, openness, interactions number, *etc.*), but with its *descriptions complexities* (depending on the properties the analyst wants to address and the expected level of details [14]).

Section 2 of this paper generally defines *descriptions* as distributions of observed events regarding some selected properties. Appropriate complexity measures for such descriptions are discussed. Among measures from information theory, we retain *Shannon entropy* for its interpretation (in term of order) and its mathematical properties [17; 14] which make it coherent regarding the aggregation process. Section 3 defines *aggregations* as simplifications of descriptional properties. We are not interested in a decrease of the system's own entropy over time, but in a decrease of entropy between two descriptions of the same system's state. Henceforth, contrary to many works on complexity reduction (*e.g.* [3; 6; 14]), we actually measure a shift between two abstraction levels. An aggregation can be evaluated according to the amount of reduced complexity (*entropy gain*), the amount of information lost during the process (*information loss*) and the accuracy of the generated description regarding the source description (*divergence*). These measures are used to evaluate aggregations and select the best one according to the analysis context (available resources, expected accuracy, *etc.*). Section 4 evaluates two spatial aggregations borrowed from geography and applied to news analysis. Informational measures allow determining which abstractions one should prefer for the analysis purposes.

2 The Notion of Description

2.1 Designing Descriptions

We call *description* any formal feature that represents the system's events according to one or several properties. They are the description's *dimensions*. Any system analysis is driven by one or another kind of such descriptions.

Unidimensional Descriptions

Let E be a set of observed events of the system's dynamics. An *unidimensional description* classifies these events according to a set V of descriptional values. We note $E_1, \dots, E_{|V|}$ the subsets of events associated to these values. They form the *events distribution* of the description.

Several dimensions of great interest can be generically identified for systems analyses:

Space *Where* does the events took place?

Time *When* did they occurs?

Agents *Which* system's entities were involved?

Topic *What* kind of events were they?

Source Where does the information come from?

[*]This paper summarizes a forthcoming research report [13].

Multidimensional Descriptions

Multidimensional descriptions present relations between properties' values. For example:

(Space × Time) The 2-dimensional distribution of events' locations over time.

(Space × Space) The distribution of events which take place in two different locations.

As we work with a fixed set of events, adding a dimension to such descriptions consists in disaggregating the current distribution to introduce a discriminatory property. For example:

(Space × Space × Time) The distribution of events' spatial relations over time.

(Space × Time × Topic) The description of *punctual events*: Something happened somewhere sometime.

Remarks on Dimensions

We do not pretend that the above-mentioned list of possible dimensions is sealed or complete. New property can easily be added from the moment that a set of precise values can be identified and observed.

These dimensions are not *a priori* independent. For example, an agent can also be characterized by places (*e.g.* main location) and dates (*e.g.* birth, death). It may be important to distinguish *a priori* inter-dependences (resulting from the mere definitions of values) from *a posteriori* inter-dependences (resulting from the observation of events).

2.2 Analyzing Descriptions

The purposes of analysis consist in explaining the designed events distributions, in revealing particularities and eventually in providing models for statistical inference. The analysis of multidimensional descriptions thus allow to answers the question: *how* the observed events occurred?

To that end, analysis methods from multivariate statistics appropriately reveal important correlations between events' properties: *e.g.*, multivariate regressions, dimension reductions, Principal Components Analysis (PCA), Correspondence Analysis (CA), *etc.* However, in case of complex systems, these tools can be very expensive to use. In order to scale such methods, this paper focuses on a preliminary step of the analysis process: the *aggregation* of properties' values (see section 3). It does not reveal *inter*-dimensional correlations, but simplifies the *intra*-dimensional representations of events by reducing the descriptional precision.

2.3 Measuring Descriptions Complexity

Complexity Measures from Information Theory

The *complexity* of a description loosely designates the number of parameters one should deal with to process to its analysis. As pointed out in [7], the size $|E|$ and the variety $|V|$ of the observed system's dynamics cannot constitute good complexity measures. They are necessary for complexity, but yet not sufficient.

Information theory proposes measures which also consider the particular distribution of events. A description can be coded as an ordered string of $|E|$ characters taken among $|V|$ possible values. In algorithmic information theory, the Kolmogorov complexity measures the size of the best lossless compression of such a string [10]. Bennett's logical depth measures the time needed to decompress such a lossless compression [2]. They really evaluate the computational resources needed to handle a description. More theoretically, Kolmogorov complexity measures the incompressible "randomness" of a distribution (*deterministic complexity*) and logical depth measures its structural complexity (*statistical complexity*) [12]. These are interesting properties conveying the fact that a fully ordered description is easy to grasp, while it is more difficult to handle a complex algorithmic structure or a totally random distribution.

These complexity measures are yet not computable in general [9] and therefor not suitable for direct application [12].

Shannon Entropy

In Shannon's probabilistic information theory [17], entropy gives a good approximation of the expected Kolmogorov complexity for a fixed distribution [9].

$$H = -\sum_{k \in V} \frac{|E_k|}{|E|} \log_2 \frac{|E_k|}{|E|} \tag{1}$$

As for Kolmogorov complexity, entropy is interesting for its interpretation in term of quantity of information. It gives the minimum quantity of information (in bits per event) needed to encode the properties' values. It is generally used as a measure of disorder or randomness: The higher the entropy, the more uncertain we are about the properties of a random event. This constitutes an important feature of macroscopic (*i.e.* low-complexity) descriptions: They introduce order in our representations of systems.

Entropy has good mathematical properties regarding the aggregation process. In particular, the *sum property* [4] shows that entropy can be defined as the sum of a local function on values' probabilities. Thus, the entropy of an aggregated description is the sum of the entropies of its aggregates (see subsection 3.3). Shannon entropy is also *recursive* [17; 4]: It can be defined according to hierarchical partitions of the distribution. The entropy of a description is then equal to the entropy of the aggregated distribution, plus the wighted sum of local aggregates' entropies.

Shannon's entropy is thus coherent with the aggregation process. *Generalized entropies* [4; 5], based on parametric information measures such as the *Rényi entropy* [16], do not have such mathematical properties [4]. Henceforth, their are not suitable to capture the notion of aggregation.

3 Aggregation of Descriptions

Entropy of descriptions depends on the selected properties and their accuracy [14]. By adjusting it, one can adapt the data complexity to the computational resources available for analysis.

3.1 Reducing the Entropy

An *aggregation* is defined according to a *source description* (basically the most precise description of events one can perform) and induces a shift in the abstraction level. It consists in partitioning the set of source values V in a set of aggregated values V', thus inducing a simpler distribution of events (and so an increase of order). We note $E'_1, \ldots, E'_{|V'|}$ the subset of events associated to aggregated values. (Note that entropy can also be reduced (1) by suppressing specific events or (2) by reorganizing the events distribution. However, such transformations either bias the size $|E|$ of observed dynamics or the original values of events.)

In opposition to algorithmic complexity (subsection 2.3), aggregations are not lossless compressions. They induce information and accuracy losses. The analyst can then be interested in monitoring their informational qualities.

3.2 Evaluating Aggregations

Entropy Gain

We note H' the entropy of the aggregated description. The *entropy gain* G of an aggregation measures the quantity of information (in bits per event) that is saved by encoding the aggregated description instead of the source description. It evaluates the amount of complexity reduced during an aggregation.

$$G = H - H' \tag{2}$$

Information Loss

The more a description is aggregated, the less it contains information about the original events distribution. We define the *information loss* L as the minimum quantity of information necessary to recover the source description from the aggregated one (*i.e.*, the cost of disaggregation). It represents the uncertainty induced by an aggregation regarding the precise values of events.

$$L = -\sum_{i \in V'} \frac{|E'_i|}{|E|} \log_2 \frac{1}{|V'_i|} \tag{3}$$

Divergence

In information theory, the Kullback-Leibler divergence measures the difference between two distributions [11]. The *divergence* thus represents the accuracy of an aggregation: The closer is the aggregated description from the source description, the lower is the divergence.

$$D = -\sum_{i \in V} \frac{|E_i|}{|E|} \log_2 \frac{|E'_i|}{|E_i||V'_i|} \tag{4}$$

Divergence evaluates the similarity of events distribution within the aggregates. This property is very interesting to build semantically coherent abstractions. In section 4.3, low-divergences are interpreted as behavioral similarities.

A simple calculus shows that $D = L - G$. The divergence can thus be interpreted as a compromise between information loss and entropy gain. The *statistical complexity* of a description (as opposed to entropy, which is a *deterministic complexity* [12]) can then be defined as *the divergence of the best aggregations* in term of entropy gain: A description is then complex when it is hard to compress it without making very rough approximations. As for Bennett's logical depth [2], this account based on divergence locates complexity between order and randomness. It is yet not semantically equivalent to logical depth since it focuses on the *difficulty of description* instead of the *difficulty of creation* [12].

Information Criteria

The *Akaike Information Criterion* (AIC) is a well-known measure for statistical models selection [1]. It describes a tradeoff between the model's complexity and its *goodness of fit*. A low-AIC model is thus a simple model with a good accuracy. In our case: $AIC = 2|V| - \log \mathcal{L}$.

Although AIC represents a good compromise between complexity, defined as number of parameters ($|V|$), and accuracy (D), one may want to use a more adequate notion of complexity. In our case, entropy: $IC = 2|H| \times |E| - \log \mathcal{L}$. We thus define a *Relative Information Criterion* (RIC) expressing the tradeoff between entropy gain and divergence:

$$RIC = \frac{\widehat{IC} - \widehat{IC'}}{2} = G - D \tag{5}$$

If $RIC > 0$, then we consider that the complexity gain offset the accuracy loss. The aggregation thus constitutes a *good abstraction*. In section 4, we use this composite measure to evaluate and compare spatial aggregations. Other criteria for model selection, such as the *Bayesian Information Criterion* (BIC), or the more general *Deviance Information Criterion* (DIC), can be expressed and exploited according to these informational measures.

3.3 Remarks on Aggregation Measures

The Sum Property

As subsection 2.3 points out, the *sum property* [4] shows that Shannon entropy can be defined as the sum of aggregates' local entropies. Entropy gain, information loss, divergence and RIC can also be defined as sums of aggregates' local measures (noted g_k, l_k, d_k and ric_k). These decompositions allow to evaluate specific aggregates instead of the whole aggregation (see sections 4.2 and 4.3).

Distributions of Reference

Entropy can be defined as the Kullback-Leibler divergence regarding the *homogeneous distribution* [11]. However, one may want to work with other distributions of reference, *e.g.* with *normal distributions*: *Entropy* is then defined as the divergence from such normal distribution. Aggregates are approximated with Gaussian functions. *Information loss* and *divergence* are defined regarding these approximations.

4 Evaluation of Spatial Aggregations for News Analysis

This section presents an application of the informational measures presented in the previous section to the context of news analysis. Two descriptions are elaborated from the content of articles published by the French newspapers *Le Monde*. Two geographic hierarchies (**UNEP** and **WUTS**) are then evaluated for aggregating such descriptions.

4.1 The Data

Sources

The GEOMEDIA project, from the CIST (*Collège International des Sciences du Territoire*, Paris) aims for an analysis platform to design, process and visualize media information.

It currently builds its own database of articles' abstracts extracted from online newspapers in the RSS format. Within 11 months, from May 2011 to March 2012, we collected 392,000 abstracts (400 characters on average) from 40 different newspapers. As an example, an analyst working on a 5-years basis, from the 10 most prolific newspapers, will need to cover 1,090,000 articles. The representation, organization and displaying of such an amount of data constitute a real challenge.

The experimentations presented in this section focus on the "International Section" of the well-known French newspaper *Le Monde*, consisting in 7076 abstracts.

Spatial and Temporal Dimensions

Each abstract relates an event which can be described according to the generic properties presented in section 2.1. They correspond to the famous 5 Ws of journalism (Who, What, Where, When, Why). We focus on two dimensions:

Space *Where* does the events related in the articles took place? Spatial tokens are extracted from the abstracts. In our experiments, we focused on the names and demonyms of 162 states [8].

Time *When* did the events occur? Publication dates of articles simply distribute them over time. The temporal dimension contains 47 weeks. (Note that more temporal tokens may be found *within the content* of abstracts.)

Two Descriptions from *Le Monde*

(Space × Space) is interpreted as *the weights of territorial relations*. The description generated from *Le Monde* is a 162 × 162 distribution of states co-citations. It is only filled at 5.1%, but the analyst still has to deal with $|E_1|$ = 4408 events.

(Space × Time) is interpreted as *the variation of territorial weights* over time. The generated description is a 162-states ×47-weeks distribution filled at 26.1% and containing $|E_2|$ = 7069 events.

Two Aggregation from Geographic Analysis

The UNEP hierarchy is used by the United Nations Environment Programme [15]. It divides the world into 6 regions (see figure 5(a)).

The WUTS hierarchy is proposed the ESPON 2013 Programme [8]. This World Unified Territorial System proposes a uniform breakdown of states: **WUTS3** divides the world territories into 7 regions and **WUTS2** into 17 regions (see figures 5(b) and 5(c)).

4.2 Evaluating the WUTS3 Aggregation

By applying the **WUTS3** aggregation to the first dimension of the **(Space × Space)** description, we generate a 17-regions×162-states distribution.

- $G \approx$ 1.33 bits per event: We saved $|E_1| \times G \approx$ 5,900 bits out of the $|E_1| \times H \approx$ 40,600 needed to encode the source description.
- $D = L - G \approx$ 2.00 bits per event: The entropy gain does not offset the information loss. The saved bits are not sufficient for disaggregation, which cost $|E_1| \times L \approx$ 14,800 bits.
- $RIC = G - D < 0$: The entropy gain does not offset the accuracy loss.

Globally, **WUTS3** has bad results. However, by looking at local measures, we can refine this evaluation.

Figure 1: **WUTS3** on **(Space × Space)**

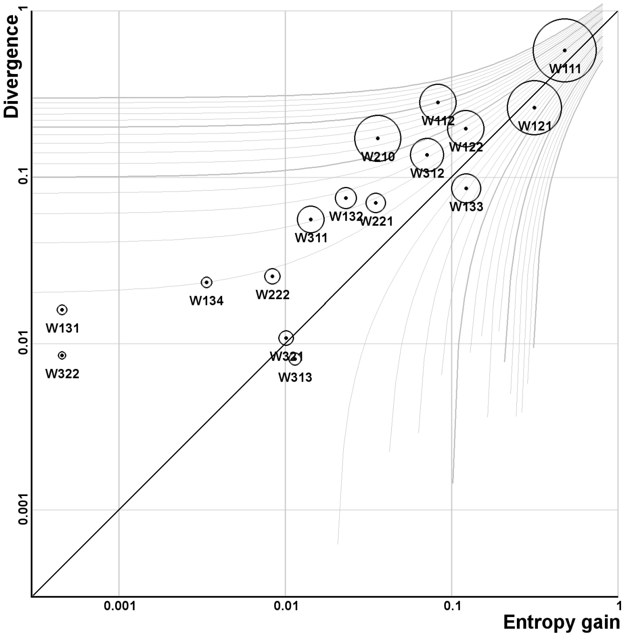
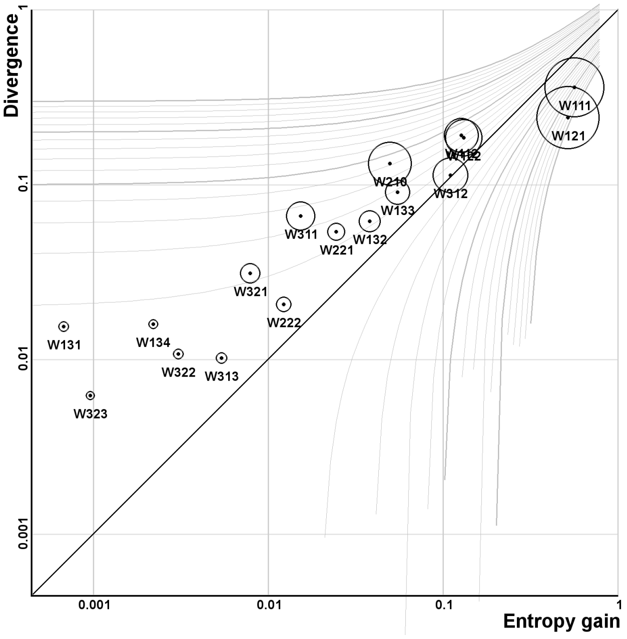


Figure 2: **WUTS3** on **(Space × Time)**



Figures 1 and 2 present the **WUTS3** aggregates positioned with respect to their entropy gains g_k (abscissa) and their divergences d_k (ordinates) on logarithmic scales. The area of circles is proportional to the number of aggregated events $|E'_k|$. The diagonal axis $d_k = g_k$ represents the limit beneath which an aggregate become RIC-positive (its entropy gain offset its divergence). In case of international relations (figure 1), this is the case for three aggregates of very different sizes:

Aggregates		$\|E'_k\|$	ric_k
W121	S.-E. Mediterranea	801	0.0507
W133	W. Africa	234	0.0363
W313	S.-E. Asia	43	0.00336

These abstractions are thus interesting for the analysis of world's territorial relations as they are related by *Le Monde*. In other words, aggregated countries behave similarly in term of international relations .

Because France is co-cited 934 times out of 4408 states couples, the biggest aggregate "Western Europe" (**W111**) has very poor results ($ric_{W111} = -0.1032$). By disaggregating France, its interest increases ($ric_{W111} = 0.03107$). The "Western Europe" abstraction can thus be used, on condition that we keep information about France's special behavior.

4.3 Comparison of UNEP and WUTS2

The aggregates of **WUTS2** are globally similar those of **UNEP**. They yet present some interesting particularities, two of which are hereafter evaluated:

1. In **WUTS2**, Mexico is located in North America (not in Latin America).
2. The aggregate "W. Asia & N. Africa" (**W12**) contains Western Asian countries (as for **U12**), but also 5 Northern African countries and 6 Central Asian countries.

Figures 3 and 4 present the compared RIC-plots of **UNEP** and **WUTS2** aggregates. Each one is positioned according to its ric_k value within the **WUTS2** aggregation (blue abscissa) and its ric_k value within the **UNEP** aggregation (red ordinates). In that way, we can easily spot the RIC-positive aggregates and tell which aggregation better defined them. (The size of the blue and red circles represent the number of aggregated events for both aggregations.)

The Place of Mexico

Should Mexico rather be aggregated with Northern America (**W21** and **U21**) or with Latin America (**W22** and **U22**)? We directly see that, in both descriptions, these aggregates are slightly better defined within the **UNEP** aggregation ($ric_{W21} < ric_{U21}$ and $ric_{W22} < ric_{U22}$, see fig. 3 and 4). In these cases, Mexico is closer to Latin America: An analyst should then use **U21** and **U22** rather than **W21** and **W22**.

The "Western Asia & Northern Africa" Aggregate

In case of international relations (fig. 3), **W12** is the only RIC-positive aggregate of both **WUTS2** and **UNEP** aggregations ($ric_{W12} = 0.0537$, on the left of the red axis). It thus constitutes the only abstraction whose accuracy loss is compensated by its entropy gain. In case of weight variations (fig. 4), **W12** is event better ($ric_{W12} = 0.340$). However, for both descriptions, **U12** is RIC-negative (below the blue axis).

Henceforth, the behavior of Northern African and Western Asian countries, both in term of international relations and weights variations, are quite similar. This can obviously be explained by the Arab Spring that took place in these countries since early 2011 and which take an important place in news media. This **WUTS** aggregate is thus interesting to represent the world's global behavior over the observed period.

Figure 3: **UNEP** and **WUTS2** on (**Space** × **Space**)

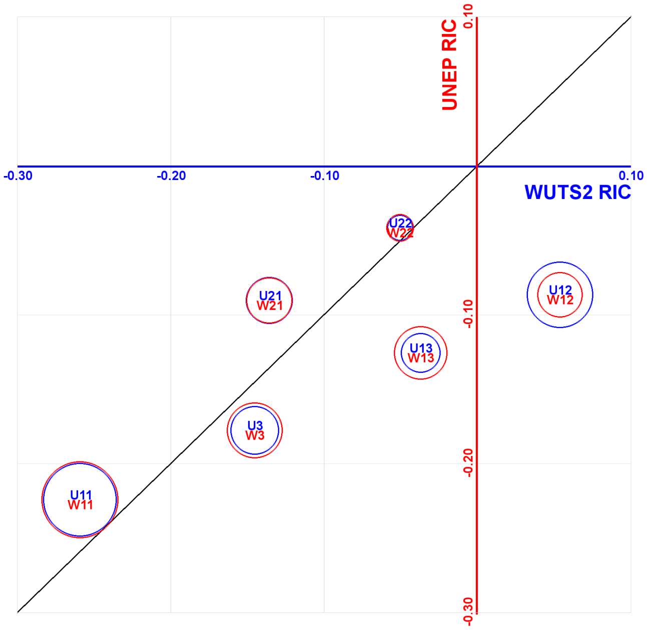
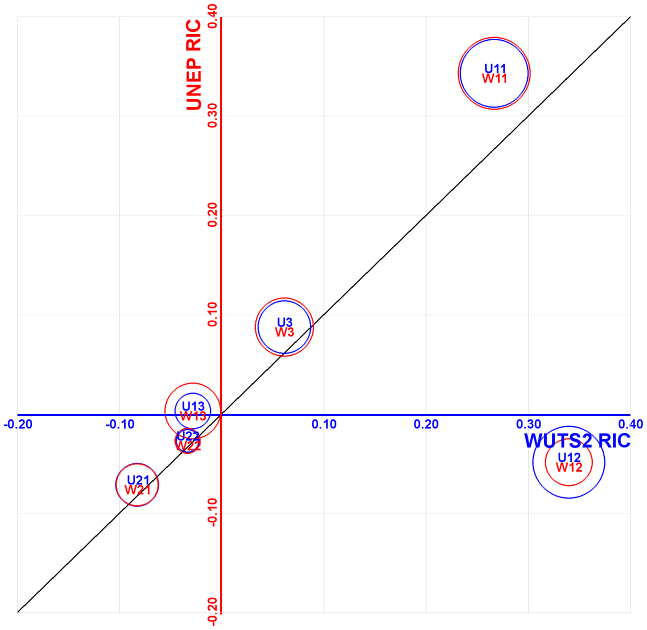


Figure 4: **UNEP** and **WUTS2** on (**Space** × **Time**)



5 Discussion and Perspectives

The very generic notions of *description* and *aggregation* presented in sections 2 and 3 are exploited for news analysis in section 4. It shows that compositions of informational measures, among the *entropy gain*, the *information loss* and the *divergence*, can be used to easily evaluate and compare aggregations (see figures 1 to 4). They allow to answer questions such as: What are the most interesting abstractions for a given description? Does a value should be integrated to a given aggregate? On what periods, or within which regions, does an abstraction can be used the more interestingly?

The experiments presented in this paper have been conducted on a rather small dataset. They illustrate some possible uses of the informational measures, but not yet constitute strong affirmations about news content. Indeed, given the small observed period and the source uniqueness (*Le Monde*), these experiments mostly capture short-term phenomena (as the Arab Spring in subsection 4.3). An important work is ongoing to indicate which properties a dataset should verify to adequately use such aggregation measures.

We also work on informational measures specific to given analysis methods. For example, in case of Principal Component Analysis (PCA), one may want to garantee that an aggregation preserves the information regarding the variances of events. A survey of the computational complexities of such statistical methods will also allow to adapt our informational measures to specific performed tasks.

Acknowledgments

We would like to thank Claude Grasland, Marta Severo and Timothée Giraud from the CIST for their collaboration on the GEOMEDIA project and their advices regarding the 4th section of this paper.

References

[1] Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] Bennett. *The Universal Turing Machine*. Oxford University Press, 1988.

[3] Bonabeau and Dessalles. Detection and Emergence. *Intellectica*, 25(2):85–94, 1997.

[4] Csiszár. Axiomatic Characterizations of Information Measures. *Entropy*, 10(3):261–273, 2008.

[5] Dehmer and Mowshowitz. Generalized Graph Entropies. *Complexity*, 17(2):45–50, 2011.

[6] Dessalles and Phan. Emergence in multi-agent systems. *Artificial Economics*, 564:147–159, 2005.

[7] Edmonds. *The Evolution of Complexity*, chapter What is Complexity? Kluwer, Dordrecht, 1995.

[8] Grasland and Didelon. *Europe in the World – Final Report*, volume 1. ESPON 3.4.1, December 2007.

[9] Grünwald and Vitányi. *Handbook of the Philosophy of Information*. North Holland, 2008.

[10] Kolmogorov. Three Approaches to the Quantitative Definition of Information. *Problems Information Transmission*, 1(1):1–7, 1965.

[11] Kullback and Leibler. On Information and Sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[12] Ladyman, Lambert, and Weisner. What is a Complex System? *European Journal of Philosophy of Science*, (forthcoming), 2012.

[13] Lamarche-Perrin, Vincent, and Demazeau. Informational Measures of Aggregation for Complex Systems Analysis. Technical report, *Laboratoire d'Informatique de Grenoble*, (forthcoming), 2012.

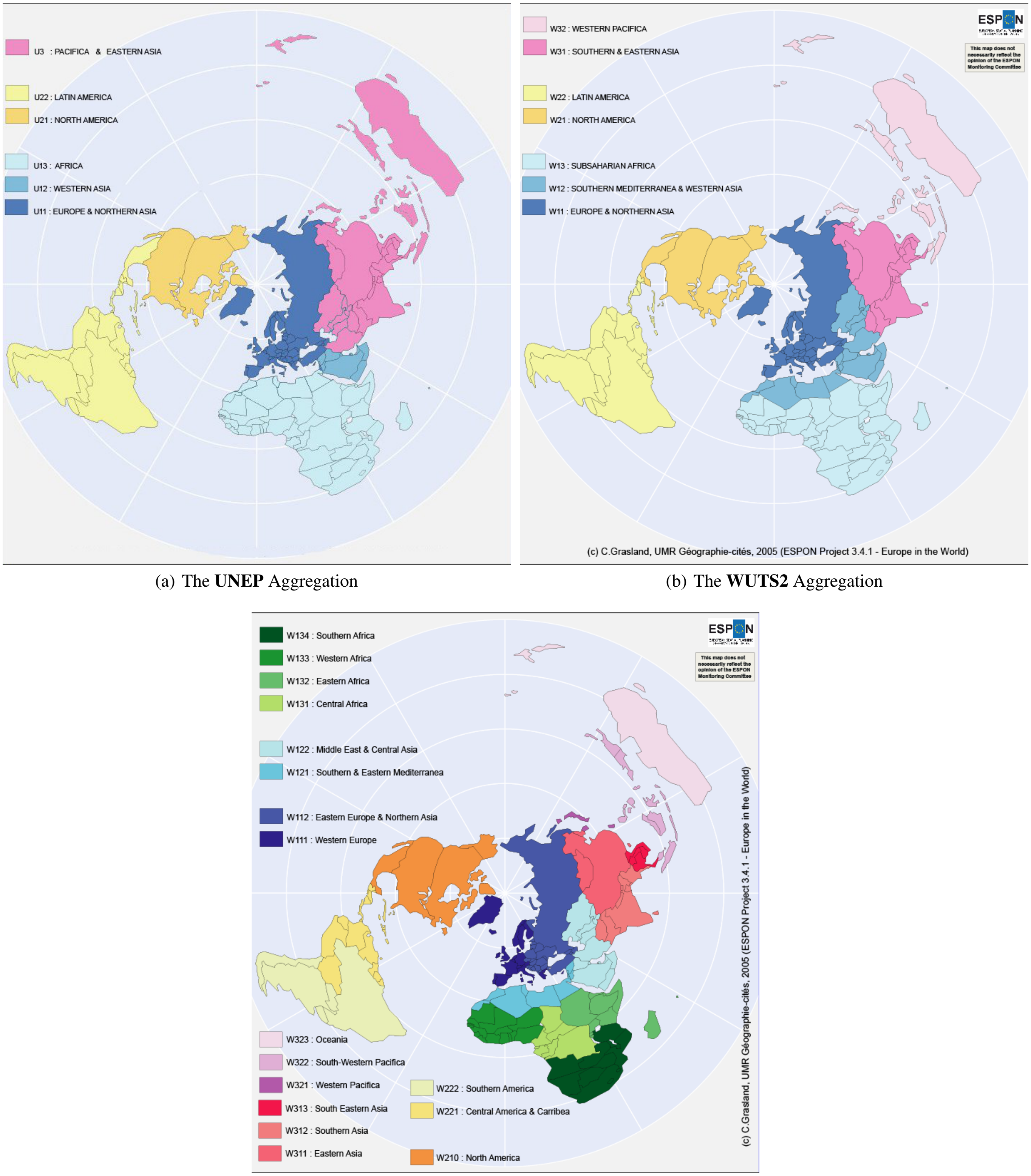
[14] Mnif and Mller-Schloer. Organic Computing – A Paradigmatic Shift for Complex Systems. *Quantitative Emergence*, pages 39–52, 2011.

[15] United Nations Environment Programme. *Global Environmental Outlook*, volume 4. Nairobi, 2007.

[16] Rényi. On Measures of Information and Entropy. In *4th Berkeley Symposium on Mathematics, Statistics and Probability*, pages 547–561, 1961.

[17] Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423/623–656, 1948.

Figure 5: Spatial Aggregations Borrowed from Geography [15; 8]


U3 : PACIFICA & EASTERN ASIA
U22 : LATIN AMERICA
U21 : NORTH AMERICA
U13 : AFRICA
U12 : WESTERN ASIA
U11 : EUROPE & NORTHERN ASIA
W32 : WESTERN PACIFICA
W31 : SOUTHERN & EASTERN ASIA
W22 : LATIN AMERICA
W21 : NORTH AMERICA
W13 : SUBSAHARIAN AFRICA
W12 : SOUTHERN MEDITERRANEA & WESTERN ASIA
W11 : EUROPE & NORTHERN ASIA
ESPON
This map does not necessarily reflect the opinion of the ESPON Monitoring Committee
(c) C.Grasland, UMR Géographie-cités, 2005 (ESPON Project 3.4.1 - Europe in the World)
W134 : Southern Africa
W133 : Western Africa
W132 : Eastern Africa
W131 : Central Africa
W122 : Middle East & Central Asia
W121 : Southern & Eastern Mediterranea
W112 : Eastern Europe & Northern Asia
W111 : Western Europe
W323 : Oceania
W322 : South-Western Pacifica
W321 : Western Pacifica
W313 : South Eastern Asia
W312 : Southern Asia
W311 : Eastern Asia
W222 : Southern America
W221 : Central America & Carribea
W210 : North America
ESPON
This map does not necessarily reflect the opinion of the ESPON Monitoring Committee
(c) C.Grasland, UMR Géographie-cités, 2005 (ESPON Project 3.4.1 - Europe in the World)

(a) The **UNEP** Aggregation

(b) The **WUTS2** Aggregation

(c) The **WUTS3** Aggregation