Informational Measures of Aggregation for Complex Systems Analysis

Robin Lamarche-Perrin
Laboratoire d'Informatique de Grenoble – Université de Grenoble
Robin.Lamarche-Perrin@imag.fr

Jean-Marc Vincent
Laboratoire d'Informatique de Grenoble – Université Joseph Fourier
Jean-Marc.Vincent@imag.fr

Yves Demazeau
Laboratoire d'Informatique de Grenoble – CNRS
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 report 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

This paper is interested in the analysis of distributed systems, either natural or artificial. 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. For example, in case of large-scale complex systems, an analysis based on the complete knowledge of system's entities requires a very large amount of resources. Therefore, 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. Aggregation thus avoids the analysis to become ressource-greedy while the size and the complexity of the studied system increase. Our 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 complexity* (depending on the properties the analyst wants to address and the expected level of details [13]). (Note that system's complexity is necessary for descriptions complexity, but not sufficient.)

Section 2 of this report 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 [16, 13] 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, 13]), 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 in order to describe and explain the social dynamics related in newspapers articles.

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. Its purposes, its difficulty, its results thus depend on a set of grounding 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 $E_1, \ldots, E_{|V|}$ 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?

The design of descriptions consists in extracting data regarding systems' events and organizing them according to these possible dimensions. One can thus build the *spatial* distribution of events, their *temporal* distribution, their *topical* distribution, *etc.*

Multidimensional Descriptions

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

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

(Space × Topic) The 2-dimensional distribution of locations regarding the type of related events.

As events can take several values of the same property, multidimensional descriptions can also present relations between values of the same property. For example:

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

(Time × Time) The distribution of events which occur at two different dates.

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 × Space × Topic) The distribution of events' spatial relations according to events categories.

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

(Agent × Time × Time) The description of agents' temporal relations: Two dates in the history of an agent are related by an event.

(Agent × Time × Topic × Topic) The distribution of agents' topical relations over time: *etc.*

Remarks on Dimensions

We do not pretend that the above-mentioned list of possible dimensions is sealed or complete. The notion of descriptions we build is meant to be generic. In particuler, new properties 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). Values of different dimensions can thus be *a priori* related. 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 multi-dimensional 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. It can be interesting to first provide abstractions in order to simplify such analyses.

This report *does not* propose a multivariate analysis method for dimension reduction. In order to scale such methods, it focuses on a preliminary step of the analysis process: the *aggregation* of properties' values (see section 3). This step is designated to reduce the difficulty of classical statistical analyses. 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

This subsection discusses some measures to define the complexity of descriptions. Afterwards, section 3 presents an aggregation process in order to reduce such measures.

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. Indeed, millions of events classified under one value among millions do not make a complex description.

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: the minimal memory space and its associated computation time. 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 and finding the algorithmic complexity of a given description is a NP-complete problem [9]. Therefore, they are not suitable for direct application [12].

Shannon Entropy

In Shannon's probabilistic information theory [16], entropy gives a good approximation of the expected Kolmogorov complexity for a fixed distribution [9]. Beside the description size $|E|$ and its variety $|V|$, the entropy H of a description depends mostly on the events' distribution (as for algorithmic complexity):

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

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: The lower the entropy, the less memory we need. (Note that entropy really gives an average measure. The actual quantity of information needed to encode the complete description is $|E| \times H$.)

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. Entropy is then maximum when events are equally distributed ($H = \log_2 |V|$) and minimum when all the events are classified under the same value ($H = 0$). 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 3.4). This useful property characterizes other informational measures, as Kullback-Leibler divergence [11] and information loss (see subsection 3.4). Shannon entropy is also *recursive* [16, 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* [15], do not have such mathematical properties [4]. Henceforth, even if they may be adapted to evaluate the randomness or the diversity of descriptions, their are not suitable to capture the notion of aggregation.

3 Aggregation of Descriptions

Entropy of descriptions depends on represented properties and their accuracy [13]. When entropy increases, the computational resources needed to handle a description increases as well. Adjusting the accuracy of descriptions thus allow to reduce their complexity and to scale-up the analysis process. The following section presents tools for such an entropy reduction.

3.1 Reducing the Entropy

By looking at formula (1), we identify three ways of reducing the entropy of a description:

1. Suppress specific events, *i.e.* describe a much ordered subset of system's events.
2. Reorganize events distribution, *i.e.* distort the description to order it.
3. Aggregate specific events or specific values, *i.e.* reduce the accuracy of the description.

Thus, if one wants to describe the whole system without biasing events' properties, aggregation of values or events is the only way for entropy reduction. In order to preserve the size $|E|$ of the observed system' dynamics, this report focuses on aggregation of values.

3.2 Aggregation of Values

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 $E'_1, \dots, E'_{|V'|}$ 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. Since entropy gives the size of the best lossless compression, an entropy reduction necessarily implies an information loss and an accuracy loss. The analyst can then be interested in monitoring their informational qualities according to several criterion: amount of reduced complexity, information and accuracy lost during the process, *etc.* In the rest of this section, measures from probabilistic information theory are exploited to do so.

3.3 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.

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

The entropy gain is maximum when all (non-empty) values are aggregated together ($G = H$). In case of full-aggregation however, we loose all information about the original distribution: The aggregated description indeed represents only one very imprecise value.

Note that, if the entropy of the source description is very low, the entropy gain cannot be very high. Thereby, no aggregation can really improve an already ordered description.

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.

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

Information loss thus depends on the size $|V'_k|$ of the aggregated values and their number of events $|E'_k|$.

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. It is maximum when events are equally distributed. A low-divergence aggregation thus indicates that aggregated values have similar events distributions. This property is very interesting to build semantically coherent abstractions. In section 4.4, low-divergences are interpreted as *behavioral similarities*.

Statistical Complexity

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. Indeed, in case of homogeneous distributions (maximal randomness), the entropy gain of any aggregation offsets the information loss ($D = 0$). In case of ordered distributions, aggregations cannot reduce much the entropy, so the best aggregations have low divergences.

This account of complexity is yet not semantically equivalent to logical depth since it focuses on the *difficulty of description* instead of the *difficulty of creation* [12].

Log-likelihood

If we consider the source description as a set of independent and identically distributed observations of a random variable representing events' values, then An aggregated description can be interpreted as a statistical model. The *goodness of fit* of such a model is given by its *likelihood*, that is the probability of generating the source description using the aggregated description as a random variable distribution.

$$\mathcal{L} = \prod_{i \in V'} \left(\frac{|E'_i|}{|E||V_i|} \right)^{|E'_i|} \tag{5}$$

A simple calculus show that $D = \widehat{\log_2 \mathcal{L}} - \widehat{\log_2 \mathcal{L}'}$ where $\widehat{\log_2 \mathcal{L}}$ is the *average log-likelihood*. Divergence then measures the "loss of fit" induced by an aggregation. A low divergence means that the analyst is likely to accurately estimate the source description from the aggregated one.

Simple calculus show that:

- The opposite of the average log-likelihood of the source distribution (the probability of generating the exact distribution from the model) is equal to the entropy: $-\widehat{\log \mathcal{L}} = \frac{-\log_2 \mathcal{L}}{|E|} = H$. Indeed, the higher is the entropy, the more the statistical model will have a important variance.

- The opposite of the average log-likelihood of an aggregated distribution is the sum of the divergence and the entropy of the source description: $-\widehat{\log \mathcal{L}'} = D + H$.

- So we have: $D = \widehat{\log \mathcal{L}} - \widehat{\log \mathcal{L}'}$.

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: $AIC = 2|V| - \log \mathcal{L}$.

The *average relative log-likelihood* evaluates an aggregated description by the mean between the average AICs of source and aggregated descriptions. A simple calculus gives:

$$relative \log \mathcal{L} = \frac{\widehat{AIC} - \widehat{AIC'}}{2} = \frac{|V| - |V'|}{|E|} - D \tag{6}$$

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. Indeed, as we showed in subsection 2.3, the variety $|V|$ of a description does not implies its complexity. We propose to use a refined information criterion using the entropy H to represent complexity: $IC = 2|H| \times |E| - \log \mathcal{L}$. We thus define a *Relative Information Criterion* (RIC) expressing the tradeoff between entropy gain and divergence:

$$RIC = \frac{\widehat{IC} - \widehat{IC'}}{2} = G - D \tag{7}$$

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.

3.4 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:

$$G = \sum_{k \in V'} g_k \qquad L = \sum_{k \in V'} l_k \qquad D = \sum_{k \in V'} d_k \qquad RIC = \sum_{k \in V'} ric_k$$

where:

$$g_k = \frac{|E'_k|}{|E|} \log_2 \frac{|E'_k|}{|E|} - \sum_{i \in V_k} \frac{|E_i|}{|E|} \log_2 \frac{|E_i|}{|E|}$$

$$l_k = -\frac{|E'_k|}{|E|} \log_2 \frac{1}{|V_k|} \qquad d_k = l_k - g_k \qquad ric_k = g_k - d_k$$

These decompositions allow to evaluate specific aggregates instead of the whole aggregation (see sections 4.3 and 4.4).

Distributions of Reference

Entropy can be defined as the Kullback-Leibler divergence regarding the *homogeneous distribution* [11]. All above-mentioned measures of aggregation are defined relatively to this distribution of reference. However, one may want to work with other bases to define and evaluate aggregations.

For example, if a metric is available, one may want to work with *normal distributions*.

- Entropy is then defined as the Kullback-Leibler divergence from such normal distribution.
- Divergence is defined as the Kullback-Leibler divergence between the aggregated distribution, where aggregates are approximated with Gaussian functions, and the source distribution.
- Information loss is still defined as the sum of entropy gain and divergence. If the analyst knows that events follow a normal law within the aggregates, she needs less information to recover the source distribution than if they follow an homogeneous law.

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 newspaper *Le Monde*. Two geographic hierarchies (**UNEP** and **WUTS**) are then evaluated for aggregating such descriptions.

4.1 Purposes of Media Information Analysis

Three classes of objects should be distinguished: (1) the *observable dynamics*: the social activities that are covered by news media, (2) the *observation devices*: the media themselves, and (3) the *generated descriptions*: the content of produced news. Artificial Intelligence works on the third class which contains informational objects. Yet, the **analysis of class (3)** can help the social sciences analyst in fulfilling three objectives:

Analysis of class (1) Representation, explanation and prediction of the observed dynamics.

Analysis of class (2) Evaluation of the observation devices (the media themselves): specific orientations, bias, extent of covered news, *etc.* The analysis focus on the influence of social dynamics on news media.

Analysis of class (1+2) Evaluation of the perturbations induced by the observation devices on the observed systems, also known as *probe effect*. The analysis focus on the impact of news media on social dynamics.

4.2 The Data

Sources

The GEOMEDIA project aims for an analysis platform to design, process and visualize media information. It results from a collaboration of computer sciences and social sciences: the CIST (*Collège International des Sciences du Territoire*, Paris) and the LIG (*Laboratoire d'Informatique de Grenoble*). The GEOMEDIA project currently builds its own database of articles' abstracts extracted from on-line 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. The average number of daily-collected articles for the 10 most prolific newspapers is close to 60 articles per day. So, as an example, an analyst working on a 5-years basis, from these 10 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 have been conducted on a subset of these data. They 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). For each dimension, one can be interested in the *indirect* occurrences of values (within the content of abstracts) or in their *direct* occurrences (regarding the articles themselves: places and dates of publication, authors, sources, *etc.*). Here, 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, 146 of which were actually mentioned at least once (see [8] for a geographical justification of the micro-states elimination). 7132 occurrences have been found.

Time *When* did the events occur? Publication dates of articles simply distribute them over time. We used a preliminary aggregation to the week-level in order to work with less temporal values. Our temporal dimension thus contains 47 weeks, from May 2nd 2011 to March 25th 2012. (Note that more temporal tokens may be found *within the content* of abstracts. We are currently working with specialized temporal tagging applications to thus enhance our temporal dimension.)

Two Descriptions from *Le Monde*

The co-occurrence of several values within the same abstract induce inter-dimensional relations that may be important for the analyst. For example, the very frequent co-citation of United States and Afghanistan in May 2011 supposes that their mutual relations should be taken into consideration for the analysis of international relations during this period. The extracted relations are presented to the analyst in the form of multidimensional descriptions. The hereafter-presented experiences use two 2-dimensional descriptions, designed from the dataset, which are commonly exploited by geographers to explain social dynamics:

(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% (most of the observed states pairs were never co-cited within the abstracts), but the analyst still has to deal with $|E_1| = 4408$ events. The entropy of this description is $H_1 \approx 9.2$ bits per events, meaning that we need at least $|E_1| \times H_1 \approx 40,600$ bits to encode the whole description.

(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. Its entropy is $H_2 \approx 10.3$ bits per events, that is only 2.6 bits less than the maximum entropy ($H_{max} = \log_2(162 \times 47) \approx 12.9$ bits per events).

Hence, **(Space × Time)** is in average a little more complex than **(Space × Space)** ($H_1 < H_2$) and globally more difficult to handle ($|E_2| \times H_2 \approx 72,500$). Depending on the available resources, its analysis can be easy or difficult. In the case of a human analysis, no expert has the cognitive skills to easily integrate and handle such an heterogeneous description.

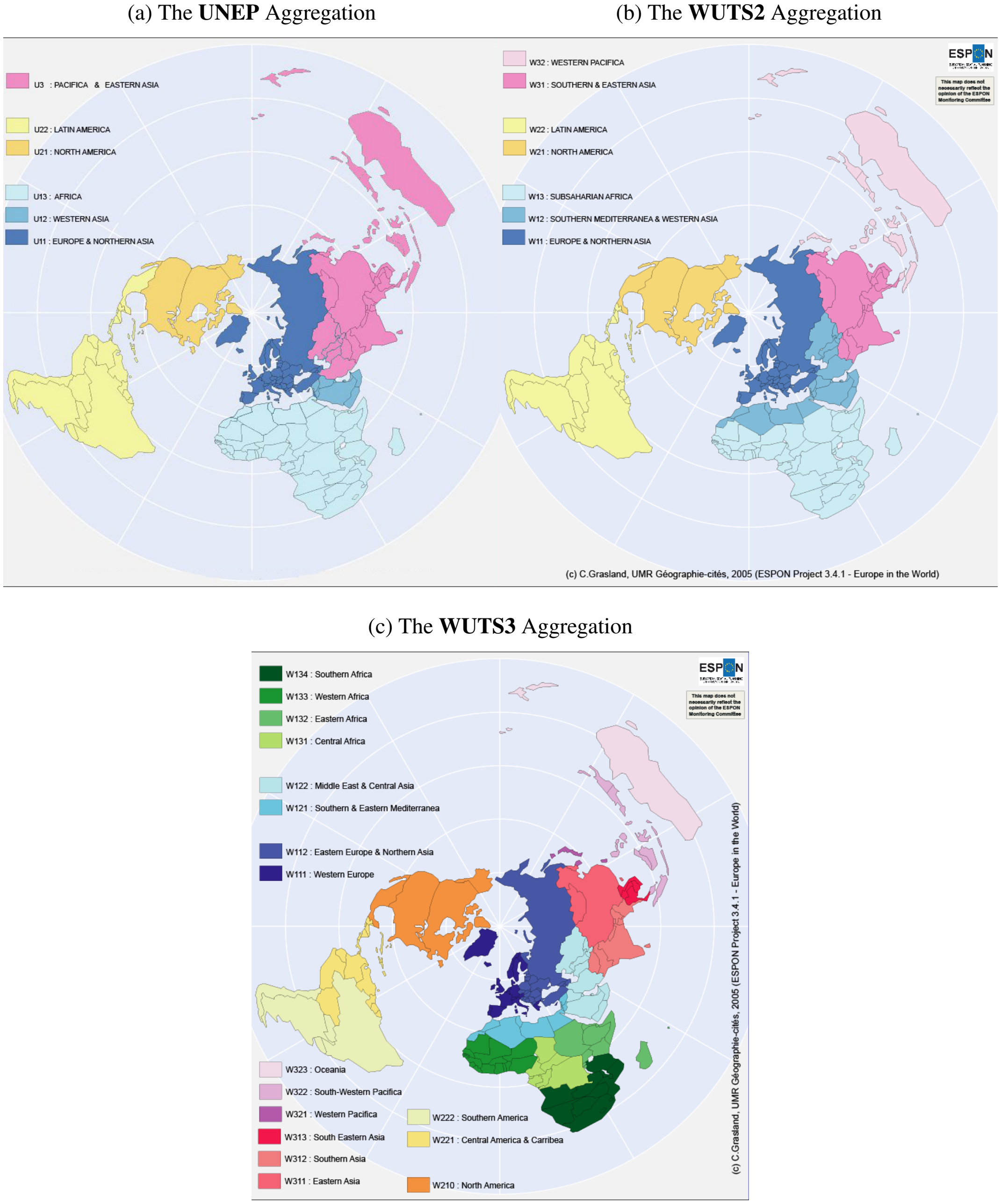
Two Aggregation from Geographic Analysis

The purpose of our experiments is to evaluate two *ad hoc* spatial aggregation used by geographers to understand world's dynamics:

The UNEP hierarchy is used by the United Nations Environment Programme in the Global Environment Outlook report (GEO) [14]. It divides the world into 6 regions (see figure 1a).

The WUTS hierarchy is proposed by the *Europe in the world* project of the ESPON 2013 Programme [8]. This World Unified Territorial System proposes a uniform breakdown of states "for the production and analysis of regional statistics". **WUTS2** divides the world territories into 7 regions and **WUTS3** into 17 regions (see figures 1b and 1c).

Figure 1: Spatial Aggregations Borrowed from Geography [14; 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

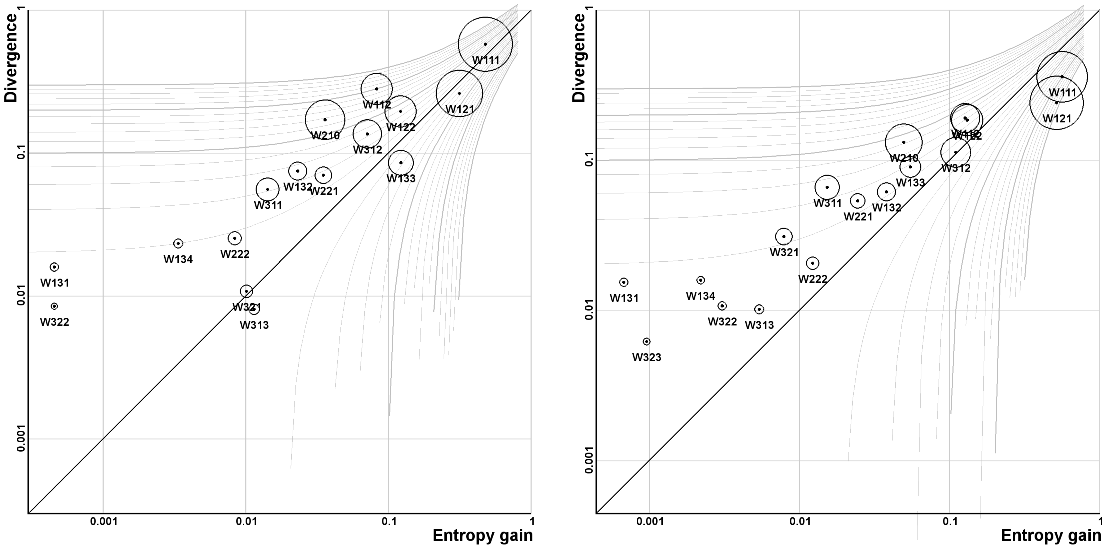
(c) The **WUTS3** Aggregation

4.3 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.

- The overall entropy gain of such a process is $G \approx 1.33$ 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. (Indeed, we need $L = 3.33$ bits per event to recover the source description.)
- $RIC = G - D < 0$: 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.



(a) On the **(Space** × **Space)** Description

(b) On the **(Space** × **Time)** Description

Figure 2: Evaluation of **WUTS3** Aggregates

Figures 2a and 2b 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. The more a circle is below this axis, the more its entropy gain offset its divergence (beware of the distances induced by the logarithmic scale). In case of international relations (figure 2a), this is the case for three aggregates of very different sizes:

Aggregates		$\lvert E'_k \rvert$	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 .

The biggest aggregate "Western Europe" (**W111**) has very poor results ($ric_{W111} = -0.1032$). This can be explained by the fact that, as the observed media is French, France is over-mentioned. Indeed, it is co-cited 934 times out of 4408 states couples! Any aggregation including France then induces an important divergence. By disaggregating France from **W111**, its interest for the analysis increases ($ric_{W111} = 0.03107$). The analyst is thus informed that the "Western Europe" abstraction can be used, on condition that we keep local information about France's special behavior.

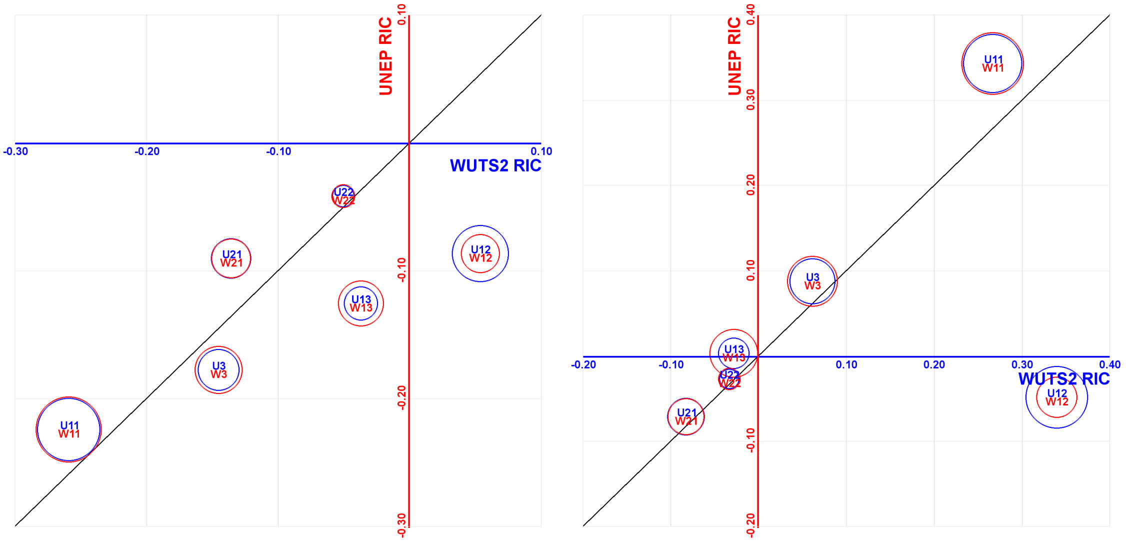
4.4 Comparison of UNEP and WUTS2

The aggregates of **WUTS2** are globally similar those of **UNEP** (see figures 1a and 1b). 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. Israel is not located in Europe (contrary to **UNEP**).
3. Northern African countries (Morocco, Algeria, Tunisia, Libya and Egypt) are not aggregated with other African countries.
4. Some countries of Central Asia (Iran, Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan and Uzbekistan) are not aggregated with Eastern Asian countries.
5. The aggregate "W. Asia & N. Africa" (**W12**) contains Western Asian countries (as for **U12**), but also the 5 previous Northern African countries and the 6 previous Central Asian countries.
6. Eastern Asia and Western Pacific form two different regions.

In the following, we use aggregation measures two evaluate these choices for the analysis of territorial relations **(Space** × **Space)** and weights variations **(Space** × **Time)** over the last 11 months.

Figures 3a and 3b present the compared RIC-plots of **UNEP** and **WUTS2** aggregates. The size of the blue and red circles represent the number of aggregated events for both aggregations. 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, one can easily spot the RIC-positive aggregates (on the left of the vertical red axis or/and above the horizontal blue axis). One can also tell which aggregation better defined particular aggregates: Those below the diagonal line are better defined by the **WUTS2** aggregation; Those above the line are better defined by the **UNEP** aggregation.



(a) On the **(Space** × **Space)** Description

(b) On the **(Space** × **Time)** Description

Figure 3: Comparison of **UNEP** and **WUTS2** Aggregates

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. 3a and 3b). Henceforth, within the news from *Le Monde*, the Mexican international relations are closer to those of Latin American countries, than to those of USA and Canada. **UNEP** is then better than **WUTS2** regarding the place of Mexico: An analyst should use **U21** and **U22** rather than **W21** and **W22**.

The "Western Asia & Northern Africa" Aggregate

In case of international relations (fig. 3a), **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. 3b), **W12** is event better ($ric_{W12} = 0.340$), but it is not the only RIC-positive aggregate (**U11**, **W11**, **U3** and **W3** also are). 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.

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 2 and 3). 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 report 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 4.4). An important work is on-going 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 guarantee 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 (Collège International des Sciences du Territoire, Paris) for their collaboration on the GEOMEDIA project and their advices regarding the 4th section of this report.

References

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

[2] C. Bennett. *The Universal Turing Machine: A Half-Century Survey*, chapter Logical Depth and Physical Complexity, pages 227–257. Handbooks of the Philosophy of Science. Oxford University Press, 1988.

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

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

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

[6] J.-L. Dessalles and D. Phan. Emergence in multi-agent systems: Cognitive hierarchy, detection, and complexity reduction. *Artificial Economics*, 564:147–159, 2005.

[7] B. Edmonds. *The Evolution of Complexity*, chapter What is Complexity? - The philosophy of complexity *per se* with application to some examples in evolution. Kluwer, Dordrecht, 1995.

[8] C. Grasland and C. Didelon. *Europe in the World - Final Report*. December 2007.

[9] P.D. Grünwald and P.M.B. Vitányi. *Handbook of the Philosophy of Information*, chapter Algorithmic Information Theory, pages 281–320. Handbooks of the Philosophy of Science. North Holland, 2008.

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

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

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

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

[14] United Nations Environment Programme. *Global Environmental Outlook: environment for development*, volume 4. Nairobi, 2007.

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

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