# A text-mining and possibility theory based model using public reports to highlight the sustainable development strategy of a city

Benjamin Duthil, Abdelhak Imoussaten, Jacky Montmain

# A text-mining and possibility theory based model using public reports to highlight the sustainable development strategy of a city

Benjamin Duthil
Laboratoire L3I
Université de La Rochelle
Avenue Michel Crépeau,
17042 La Rochelle Cedex 01, France
benjamin.duthil@univ-lr.fr

Abdelhak Imoussaten
Centre de recherche LGI2P
École des mines d'Alès, site ERIEE
Parc scientifique G. Besse,
30035 Nîmes cedex 01, France
abdelhak.imoussaten@mines-ales.fr

Jacky Montmain
Centre de recherche LGI2P
École des mines d'Alès, site ERIEE
Parc scientifique G. Besse,
30035 Nîmes cedex 01, France
jacky.montmain@mines-ales.fr

***Abstract*—Nowadays, ecology and sustainable development are priority government's actions. In Europe, and more specifically in France, sustainable development (SD) is generally broken down into several distinct evaluation criteria. Each criterion is a requirement imposed by the government and corresponds to strategic stakes. When SD improvement actions are financed in an economic region or a city of the French territory by the government, a set of measures is usually set up to assess and control the impact of these actions. More precisely, these measures are used to check whether the region or the city has efficiently invested its budget in respect to the SD strategy of the government. This assessment process is a complex task for the government. Indeed, evaluations are only based on reports provided by the financed regions. These very numerous reports are written in natural language and thus, it is a thorny and time-consuming task for the government to efficiently identify the meaningful information in a plethora of reports and then objectively assess all the expected priorities. This project aims at automating the assessment process from the huge corpus of documents. Text-mining and segmentation techniques are introduced to automatically quantify the attention the region or the city pays to a given criterion. Obviously, this quantification can only be imprecisely determined. Then, the possibility theory is used to merge the information related to each criterion prioritization from all the documents. Finally, an application on the 265 largest cities in France shows the potential of the approach.**

## I. Introduction

The incredible increase in the web knowledge resources is supported by any internet user, any company or local authority, etc. Internet users provide a huge amount of comments through recommender websites, companies can publish their product catalogues directly on the WWW, and local authorities publish their own self evaluation reports and specific subjects or policies. This information is a windfall for the decision makers (DM) who are interested to buy a product, to analyze the preferences of customers etc. Decisions are often based on several criteria and Multiple Criteria Decision Analysis (MCDA) seems to be a necessary tool to support DMs to exploit so abundant information. We have already proposed several works based on MCDA and text-mining techniques have been proposed to this aim. For example, in [1] and [2] a MCDA recommender system is proposed to help Internet users to choose a movie. In [3] an index to determine realistic and prioritized targets in tourism applications is introduced. In addition to this exploitation for entertainment purposes, one can imagine also to extract useful information from local authorities reports for control purposes.

Indeed, many reports are published by local authorities about their priority actions in order to ensure public transparency. Looking for precise information in these reports may be very tedious for an individual citizen when considering the huge amount of official reports that are published by local authorities. But the task is much more thorny for the government who tries to be in control of policy and operation in all cities in the country.

To illustrate this problem, this paper focus on the serious concern of the government faced with sustainable development (SD) [4], [5]. The government is interested to determining the extent to which regions benefiting from financing meet the state's expectations with respect to sustainable development (SD) [4], [5]. Indeed, nowadays, sustainable development are priority government's actions [6], [7]. In Europe, and more specifically in France, the assessment of the respect of the protocol on the application of the government SD policy is generally broken down into several distinct evaluation criteria. Each criterion is defined by the government or scientific community (United Nations Conference on Sustainable Development (UNCSD)) and corresponds to strategic stakes that are [1] *Agenda 21, energy transition, the protection of soil biodiversity, organic farming, sustainable consumption, waste management, water quality, air quality*. The government enforces the law by giving instructions to the different territorial elected representatives (*i.e.* regional president, prefect, mayor) and provides the necessary funds to local authorities for the application of SD actions. However, it is very difficult for policy makers to drive the implementation of their policy be-

[1] http://www.developpement-durable.gouv.fr/

cause it is difficult to measure its genuine implementation and impact.Supporting the decision maker in this tedious process is one of the current scientific challenges [4], [8].Today, political DM has reports at his disposal, written in natural language, describing all actions cities have implemented in order to meet SD government policy objectives.

The aim of our proposal is to automatically analyse these reports in order to check the priority given by local authorities to the criteria related to SD assessment. Natural language processing remains a complex challenge [9], [10], [11]. Because of the imprecision inherent to language, it is difficult to automatically identify and extract topics from a text in natural language (NL). The idea of this proposal is to evaluate the extent to which a city is mobilized to follow governmental recommendations *w.r.t* the criteria related to SD assessment. Our assumption states that: the more documents published by the city deal with a given criterion, the higher priority this criterion seems to be for the city. Then, our issue is to automatically scroll all the documents a city publishes and extract from them all the pieces of knowledge that refer to the SD criteria. The density of segments related to each criterion assesses the importance the city grants to each criterion. Nevertheless, dealing with imprecision in NL prevents this automatic count to be so precise and specific tools of the possibility theory are required to carry out this study.

Our approach conjointly uses text-mining techniques and possibility theory notions to evaluate as accurately as possible the extent to which a SD criterion is recurrent in the city's publications. On one hand, text-mining [12], [13], [14], [15] techniques are now mature enough to answer the problem of extraction of information and will be used to extract SD criteria in published reports; on the other hand, possibility theory [16] [17] offers the adequate framework to handle and merge multiple imprecise data provided by the analysis of multiple reports. Finally, we obtain a fuzzy prioritization of criteria that allows the government checking whether the SD policy of the city is consistent with its recommendations.

The paper is organized as follows: Section 2 describes the text-mining approach. Then, section 3 introduces the merging process of imprecise data. Section 4 describes the complete processing pipe of our approach. Section 5 presents the case study.

## II. Text-mining

In this section, we describe the text-mining process used to extract information about the importance given to the SD evaluation criteria in the local authorities communication reports written in natural language. As mentioned above, natural language processing remains a complex challenge. This complexity is related to the words that compose it. Indeed, a word has three dimensions [18]: the first one is syntactic, it depends on the grammatical construction of the sentence, the second one is semantic, *i.e.*, the meaning of a word, and the third one is pragmatic, it is related to the individual's personal experience and perception. This pragmatic dimension confers its imprecision to language. This imprecision must be taken into account in the topics extraction process. Hence, detecting the presence of a topic (a criterion here) in a text passage is not so easy. In conclusion, the ratio of the text that can actually be assigned to this criterion can only be imprecisely assessed. In [19], an unsupervised approach is proposed to detect and extract this imprecise information. This approach called *Synopsys* allows to extracting text segments related to specific criteria and consists on 3 steps: *1) Automatic construction of a training corpus used to learn characteristic words, called descriptors, for a criterion of interest; 2) Automatic learning of these descriptors and construction of a lexicon associated with the criterion; 3) Text segmentation using the lexicon associated with the criterion. Synopsys* approach uses a set of seed-words. On one hand, they serve to semantically characterize the criterion of concern, and on the other hand, to initiate the learning of descriptors for the criterion [19]. The training corpus is built automatically. Schematically, the more frequently a word is found in the neighbourhood of a seed word of the criterion (counting on sliding window), the greater the membership function of this word to the lexical scope of the criterion. In the following the *Synopsys* steps are presented in figure 1 and described in details.

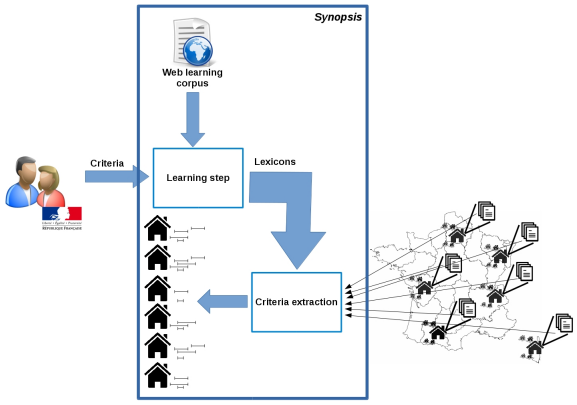


Fig. 1. Description of Synopsis steps

### A. Document acquisition

*Synopsis* builds its training corpus from web documents. Step 1 thus consists on defining the domain and the criteria. For each criterion, the user just needs to give a set of words called seed words to specify what this criterion intends to be. A corpus of documents is built for each seed word of a specific criterion (relatively to a domain) by querying a search engine for documents containing at least one occurrence of this seed word. The corpus of documents formed by the union of all the corpora related to criterion's seed words is named the class of the criterion. Similarly, a second corpus is built for this criterion: this time, the search engine gathers documents containing none of the criterion's seed words. Intuitively, this corpus named anti-class intends to give a better characterization of the criterion. The HTML tags, advertisings

and so on... are then taken out from the criterion corpus documents. These documents are then transformed using a morpho-syntactic analyser.

### B. Lexicon building

Step 2 intends to identify the word representative (resp. non-representative) of the criterion from the lemmatized texts from Step 1. This step identifies the representative (respectively non-representative) words for a criterion denoted $C$. This is achieved by occurrence frequency analysis, assuming that the probability that a word characterizes a criterion is proportional to its occurrence frequency in the immediate neighbourhood of one of the criterion's seed words. This occurrence frequency is computed over the whole corpus of criterion $C$ and is used to quantitatively capture the representativeness score $Sc$ of a word $W$ *w.r.t.* $C$. At the end of this step, lexicon $L$ related to a concept $C$ is formed with a set of words and their representativeness w.r.t. $C$. Two categories of words are distinguished: *i.e.* those prevailing in the class and those prevailing in the anti-class. More formally, the words in the immediate neighbourhood of a criterion's seed word are first selected inside a window $\mathcal{F}$ of size $sz$ in a document $d$ [20]:

More formally, a window $\mathcal{F}$ is :

$$\mathcal{F}(r, sz, d) = \{w \in d / d_{noun}(r, w) \leq sz\} \quad (1)$$

where $w$ is an instance of word $W$ in the text, $r$ is the seed-word and $d_{noun}(r, w)$ is the distance corresponding to the number of nouns (considered as meaningful words) separating word $w$ from $r$ in the document $d$ [20].

It is now possible, for each word of the corpus, to define its representativeness in the class of the criterion $C$. It is defined as the sum of occurrences of a word in a window $\mathcal{F}(r, sz, d)$ for all the seed words of $C$ and all the documents of the corpus. From the representativeness of a word in the class and in the anti-class, a score is established for this word using a discrimination function. A lexicon of relevant words is finally built for each criterion.

### C. Text annotation

Finally, Steps 3 explains how the achieved lexicon can be used to obtain thematic segmentation of any document. A sliding window $\mathcal{F}$ is introduced: it is successively centred on each occurrence of nouns in the processed document $d$. From lexicon $L$ of a criterion $C$, a score is computed as follows for each sliding window $\mathcal{F}$ in a document $d$:

$$Score(\mathcal{F}) = \sum_{w \in \mathcal{F}} Sc(w, sz) \quad (2)$$

In document $d$, the sliding window $\mathcal{F}$ is said to be related to a criterion $C$ as soon as its score is higher than a predetermined threshold. Roughly speaking, the higher this threshold, the more reliable the matching between the selected sliding windows and the criterion $C$. Several segmentations depending on the threshold value can then be achieved. The number of words that can be linked to the criterion $C$ is a function of the threshold value. The number of words slowly evolves with the threshold value, except for some singular values that correspond to rough changes in pragmatic points of view, *i.e.* significant breaks in the granularity description. The choice of the threshold can be supported by sensitivity analysis of the function. In practice, for any document, we compute the arithmetic mean of the scores of the selected windows in each possible segmentation: the segmentations with the lowest and the uppermost means provide the lower and upper bound of the interval that characterizes the value range of the degree to which the text talks about the criterion. Next, this computation must be processed for all the documents to the scale of a whole city. The intervals achieved from the documents are finally merged to compute a fuzzy degree that estimates the place of the criterion in the SD communication of the city. This fusion process is now presented.

## III. Merging Imprecise Statistical Data

The more a criterion is evoked in communication reports, the more it is a matter of public interest in the city policy. Natural language may be unclear. Even human beings may be in trouble to precisely identify what the text is talking about in some ambiguous cases. A topic is a vague notion and should be modelled as such. Hence, automatic topics detection cannot be precisely performed; topics extraction must be seen as an imprecise process. In addition, the information obtained is also uncertain because of the multiplicity of reports collected speaking about the criterion. In summary we are faced to the problem of merging imprecise statistical data. We propose to deal with these data using possibility theory which provides an appropriate framework to represent and merge imprecise and uncertain data. Indeed, classical uncertainty theories, *e.g.*, probability theory, meet their limits when they have to deal with imprecise information. This is due to the additivity property required by these theories. Possibility theory in a non additive theory which makes it more flexible [16]. The degree of probability assigned to an event, in the framework of probability, is replaced here by two dual degrees: possibility degree and necessity degree. These two degrees are determined from the information obtained from the reports and having a link with the event. In the following a short reminder on possibility theory is proposed.

### A. Possibility Theory

Let $\Omega$ represents a universal set of elements $\omega$ under consideration that is assumed to be finite and let $2^{\Omega}$ represent the power set of $\Omega$. A possibility distribution $\pi$ is a normalized function $\pi : \Omega \longrightarrow [0, 1]$ (i.e. $\exists \omega \in \Omega$, such that $\pi(\omega) = 1$). $\pi$ assigns to each element $\omega$ in $\Omega$ a degree of possibility $\pi(\omega)$ of being the correct description of a state of the world. From $\pi$, possibility and necessity measures are respectively defined for all subsets $A \subseteq \Omega$: $\Pi(A) = \sup_{\omega \in A} \pi(\omega)$ and $N(A) = 1 - \Pi(A^c)$. $\Pi(A)$ quantifies to what extent the event $A$ is plausible while $N(A)$ quantifies the certainty of the event $A$. Conversely, possibility distribution can be obtained from possibility measure as follows: $\pi(\omega) = \Pi(\{\omega\})$. An interesting concept defined in possibility theory is the $\alpha$-cut of

a possibility distribution. For a possibility distribution $\pi$, an $\alpha$-cut is the classical subset: $E_\alpha = \{\omega \in \Omega : \pi(\omega) \geq \alpha\}$, $\alpha \in ]0,1]$. Finally, two subsets are still important in the representation of a possibility distribution: 1) the support of a distribution: $support(\pi) = \{\omega \in \Omega : \pi(\omega) > 0\}$; and 2) the kernel of a distribution: $kernel(\pi) = \{\omega \in \Omega : \pi(\omega) = 1\}$.

### *B. Building Possibility Distributions From Imprecise Statistical Data*

Let consider a set of distinct intervals $\{I_j, j = 1, nbi\}$ as the data obtained after the Synopsis analysis of the city communication reports and their occurrence $m(I_j)$.
When intervals are nested, *i.e.* $I_1 \subset I_2 \subset ... \subset I_{nbi}$, a possibility distribution $\pi$ may be built from possibility (plausibility) measure, as proposed in [16] [21]: $\forall \omega \in \Omega$, $\pi(\omega) = \Pi(\{\omega\}) = \sum_{j=1,nbi} m(I_j).\mathbb{1}_{I_j}(\omega)$ where $\mathbb{1}_A$, $A \subseteq \Omega$ denotes the characteristic function.
When intervals are not nested but only consistent, *i.e.* $\bigcap_{j=1,nbi} I_j = I \neq \emptyset$ (all reports share at least one value), two possibility distributions $\pi_1$ and $\pi_2$ are built: First, we set $\pi_1(\omega) = \sum_{j=1,nbi} m_1(I_j).\mathbf{1}_{I_j}(\omega)$, $\forall \omega \in \Omega$. Second, $r$ nested focal elements $\{E_s, s = 1, r\}$ are obtained from the $\alpha$-cuts of $\pi_1$: $E_1 = I$ and $E_s = E_{s-1} \cup E_{\alpha_s}(\pi_1)$ $(s = 2, r)$. The new occurrences $m_2$ assigned to intervals $E_s$ are computed as proposed in [16]: $m_2(E_s) = \sum_{\{I_j \ related \ to \ E_s\}} m_1(I_j)$ (each assessment $I_j$ being *related* in a unique way to the smallest $E_s$ containing it).

Then a possibility distribution $\pi_2$ can be defined as: $\forall \omega \in \Omega$, $\pi_2(\omega) = \sum_{s=1,r} m_2(E_s).\mathbf{1}_{E_s}(\omega)$. Membership functions $\pi_1$ and $\pi_2$ are mono modal possibility distributions since $\bigcap_{j=1,nbi} I_j = I \neq \emptyset$ holds. Furthermore, they are the best possibilistic lower and upper approximations (in the sense of inclusion) of assessment sets $\{I_j, j = 1, nbi\}$ [16]. It can be seen easily that $\pi_1 \subseteq \pi_2$ (inclusion of fuzzy subsets) as $\forall \alpha \in ]0,1]$, $E_{1,\alpha} \subseteq E_{2,\alpha}$.
The consistency constraint may not be satisfied in practice, i.e. $\bigcap_{j=1,nbi} I_j = \emptyset$ due to divergence expressed by reports, *e.g.,* some reports seem to promote some sustainable development aspects, whereas these aspects may be totally absent in some other ones. To cope with this situation, groups of intervals, *maximal coherent subsets* (*MCS*), with a non-empty intersection are built from original intervals. This is made by finding subsets $K_\beta \subset \{1, ..., nbi\}$ with $\beta \in \{1, ..., g\}$ such that: $\bigcap_{j \in K_\beta} I_j \neq \emptyset$, with g being the number of subsets $K_\beta$ [17] [22]. For each group $K_\beta$, lower and upper possibilistic distributions $\pi_1^\beta$ and $\pi_2^\beta$ are built (as in the previous case when elements are consistent). Let possibility distribution $\pi_1$ (resp. $\pi_2$) be the union (denoted $\widetilde{\bigcup}$) of possibility distributions $\pi_1^\beta$ (resp. $\pi_2^\beta$):

$$\pi_1 = \widetilde{\bigcup_{\beta=1,g}} \pi_1^\beta \ (resp. \ \pi_2 = \widetilde{\bigcup_{\beta=1,g}} \pi_2^\beta) \quad (3)$$

then $\pi_1$ and $\pi_2$ are the multi-modal ($g$ modes) possibilistic lower and upper approximations of original intervals.

### *C. Matching Between Gradual Linguistic Information and Imprecise Statistical Data*

We assume in this work that the prioritization required by the government can be expressed in natural language. For example, it may be expressed as follows: "the criterion is very important". This expression is a gradual linguistic information and it is represented in our approach by a fuzzy subset in form of a trapezoidal distribution that can be defined by its support and kernel. Let $\Omega$ be the interval $[0, 100]$, a possible representation of the expression "the criterion is very important" may be the distribution having $[80, 100]$ as support and $[85, 95]$ as kernel.
The aim of this subsection is to quantify to which degree the priority of a SD criterion recommended by the government (*e.g.*, Water quality is of major importance) is adequate to the city communication policy. More formally, the trapezoidal distribution that characterizes the linguistic requirement of the government is compared to the possibility distribution that results from the merging of the voting documents. Let still consider two possibility distributions $\pi$ and $\pi'$ defined on $\Omega$.We define the degree of inclusion of two mono-modal possibility distributions $\pi$ in $\pi'$ as:

$$incl(\pi, \pi') = (\int_\Omega (\pi' \wedge \pi)) / \int_\Omega \pi \quad (4)$$

Let consider possibility distributions $\pi_1$ and $\pi_2$ as the the possibilistic approximations built in the previous section of a set of intervals $\{I_j, j = 1, nbi\}$. Furthermore, let represent the gradual linguistic information by a trapezoidal possibility distribution $\pi_w$. In case of consistent data, we define the degree of matching of $\pi_w$ to data $\{I_j, j = 1, nbi\}$ as:

$$match(\pi_w, \{I_j\}) = [incl(\pi_1, \pi_w) + incl(\pi_w, \pi_2)]/2 \quad (5)$$

In case of inconsistent data, we define the degree of matching of $\pi_w$ to data $\{I_j, j = 1, nbi\}$ as:

$$match(\pi_w, \{I_j\}) = \max_{\beta=1,g} match(\pi_w, \{I_j, j \in K_\beta\}) \quad (6)$$

where $g$ is the number of maximal coherent subsets $K_\beta$ computed from the imprecise data $\{I_j, j = 1, nbi\}$. At the end of this step, the importance the local authority grants to a given SD criterion can be linguistically described with the labels the government proposes.Then, the data $\{I_j\}$ can be seen as a discrete fuzzy set over the linguistic labels that characterize the priority. After some normalization process, the degrees of match in equations (5) and (6) then represent the membership degree of the data to each of the labels.

## IV. General process

After the technical descriptions of our both major data processing tasks, let us come back to the basic principle of their conjoint use.

1) Defining the government's criteria for assessing the way standards of sustainability are derived after local implementation. Evaluating public policy on sustainability does not require a detailed understanding of the underlying technology, but rather a willingness to weigh the issues raised by the SD in a broader social context.
2) Gathering all the reports in which local authorities communicate about their SD actions;
3) Extracting knowledge pieces with the Synopsis approach to assess how often reports pick up on any SD criterion. This imprecise measure gives some idea of the importance local authorities place on the matter;
4) Merging these imprecise measures into a possibility distribution that captures the overall expression of interest local authorities grant to a SD criterion;
5) Identifying the best match of the resulting possibility with gradual linguistic labels;
6) Checking whether the prioritization granted to a criterion in the communication reports of local authorities corresponds to the expected objective defined in the SD policy of the government.

## V. Case study

It is nowadays common practice that financial committee states its view that eligibility for project funding under the various programs and budget headings should be based on the condition of sustainable development. Sustainable development must be broken down into tangible objectives that can be related to elementary criteria: *Agenda 21, protection of soil biodiversity, energy transition, organic farming, sustainable consumption, waste management, water quality, air quality*. Indicators can then be associated to these objectives. These indicators can be used by government to control and supervise the way its SD policy is carried out within the economic territory. However, these indicators represent only a tiny fraction of the overall SD policy. Most advances in SD policy cannot be quantitatively measured, but qualitatively appraised through the interpretation and validity of various implementations in practice. SD is now seen as a matter of substantial importance within public health and economic policies; this explains its central role in the public documents local authorities publish for communication purposes.

The approach we propose here is a decision-support system that helps central government to assess to which extent its policy is implemented on the national territory. Each communication report the city publishes is seen as a potentially useful measurement for supervision purposes. As stated above, appraisals can only be qualitative for a matter of interpretation. Our approach has been designed to deal with this imprecision. We thus imagine that government defines a subset of linguistic levels to qualify the importance local authorities grant to a given criterion: "not important", "not enough important", "moderately important", "enough important" and "very important". We represent these levels as trapezoidal possibility distributions over the value range [0,100]. Table I gives the kernels and supports of those distributions.

TABLE I
Definition of importance levels

| liguistic expression | kernel | support |
|---|---|---|
| 1: "not important" | $[5, 15]$ | $[0, 20]$ |
| 2: "not enough important" | $[25, 35]$ | $[20, 40]$ |
| 3: "moderately important" | $[45, 55]$ | $[40, 60]$ |
| 4: "enough important" | $[65, 75]$ | $[60, 80]$ |
| 5: "very important" | $[85, 95]$ | $[80, 100]$ |

We gathered reports related to the 265 France's biggest cities available on the web. For each city, we collected about fifty documents, which constitute a dataset of 13794 documents in total. These documents are published by either municipalities or big corporations. The general processing from text analysis to imprecise statistical data fusion is applied for each city. For instance, we propose in table II the result for the city of Paris. To see the results of other cities, the reader is invited to use the following url: http://navidomass.univ-lr.fr/CIVEMSA.html. Figure 2 is a screenshot of the application.

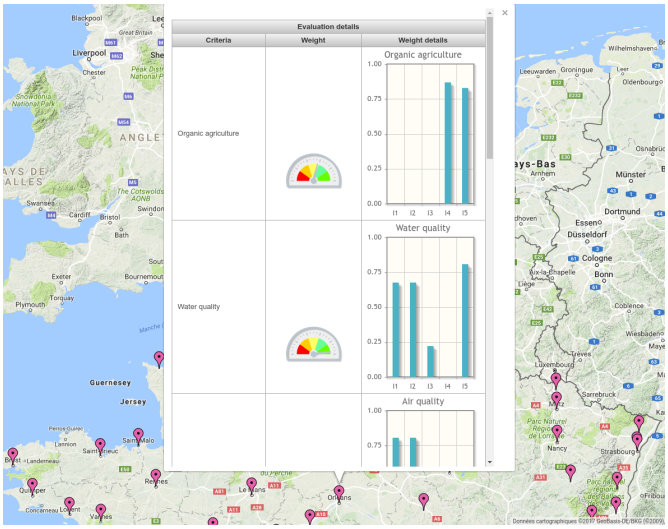


Fig. 2. Example details for Orléans city

Concerning Paris results, we can see that no criterion is actually considered *highly important* in the city communication reports. Three criteria are rather considered *enough important*: energy transition, protection of soil biodiversity and waste management. In the opposite *Agenda 21* and *air quality* seem to be ignored and there is a clear lack of communication about them: it should be discussed soon.

Another man machine interface available in our decision-support system shows a map of France. It is then possible to represent the priority of each criterion in the communication reports of all the 265 cities in France. For example, figure 3

TABLE II
GENERAL PROCESS APPLIED TO PARIS CITY: DEGREES OF MATCHING FOR THE 8 CRITERIA TO PRIORITY LEVEL

| liguistic expression | energy transition | protection of soil biodiversity | Agenda 21 | water quality |
|---|---|---|---|---|
| 1 | 0.15 | 0.87 | 0.7 | 0.8 |
| 2 | 0.85 | 0 | 0.8 | 0.73 |
| 3 | 0.6 | 0.76 | 0 | 0.4 |
| 4 | 0.5 | 0.57 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 |
| liguistic expression | air quality | sustainable consumption | waste management | organic farming |
| 1 | 0.77 | 0.85 | 0.29 | 0.42 |
| 2 | 0.55 | 0.79 | 0.78 | 0.78 |
| 3 | 0 | 0.22 | 0.56 | 0.32 |
| 4 | 0 | 0 | 0.26 | 0 |
| 5 | 0 | 0 | 0 | 0 |

illustrates the results for the criterion *sustainable consumption*: the larger the round, the greater the criterion priority for the city. We can state that the disparities are flagrant over the French territory for this criterion. For more details, an online application to visualize the assessment of each French city w.r.t each of the eight SD criteria has been developed (http://navidomass.univ-lr.fr/CIVEMSA.html). It offers a view dedicated to each of the criteria (by clicking on the marker of a city).

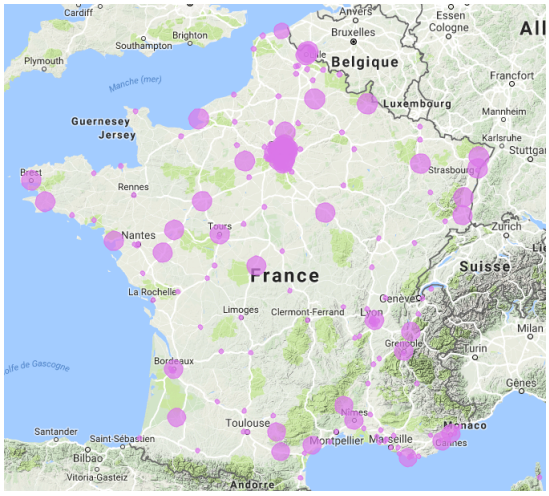


Fig. 3. Example for the sustainable consumption criterion

## VI. CONCLUSION

We have proposed an approach based on text-mining and uncertainty theory to automatically assess the priority given by local authorities in France to governmental sustainable development requirements. We cannot use this analysis to check whether a city implements or not some SD actions, but merely check that this city is careful to increase awareness of sustainable development *w.r.t* to the SD priorities of the government. This work could be a useful support of mediation between local authorities and central government: this will help to early identify the critical issues, to identify the challenges and to construct remedial actions as quick as possible.

## REFERENCES


[1] A. Imoussaten, B. Duthil, F. Trousset, and J. Montmain, "Possibilistic interpretation of a sentiment analysis for a web recommender system," LFA'2013, Reims, 2013.

[2] ——, "A highly automated recommender system based on a possibilistic interpretation of a sentiment analysis," in Information Processing and Management of Uncertainty in Knowledge-Based Systems. Springer, 2014, pp. 536–545.

[3] ——, "Automated extraction of multicriteria assessments for allocating realistic and prioritized targets in tourism applications," LFA'2016, La Rochelle, 2016.

[4] OECD, "Environment at a glance 2015."

[5] L. P. M. BORZA, "The sustainability dashboard: An appraisal tool for business area competitiveness," 2016.

[6] "Growth, competitiveness, employment: The challenges and ways forward into the 21st century - white paper. parts a and b. com (93) 700 final a and b, 5 december 1993. bulletin of the european communities, supplement 6/93," 1993.

[7] "A sustainable europe for a better world: A european union strategy for sustainable development. commission's proposal to the gothenburg european council. com (2001) 264 final, 15 may 2001," 2001.

[8] A. Scipioni, A. Mazzi, M. Mason, and A. Manzardo, "The dashboard of sustainability to measure the local urban sustainable development: The case study of padua municipality," Ecological Indicators, vol. 9, no. 2, pp. 364 – 380, 2009.

[9] M. Delgado, M. D. Ruiz, D. Snchez, and M. A. Vila, "Fuzzy quantification: a state of the art," Fuzzy Sets and Systems, vol. 242, pp. 1 – 30, 2014, theme: Quantifiers and Logic.

[10] M. Chatzigeorgiou, V. Constantoudis, F. Diakonos, K. Karamanos, C. Papadimitriou, M. Kalimeri, and H. Papageorgiou, "Multifractal correlations in natural language written texts: Effects of language family and long word statistics," Physica A: Statistical Mechanics and its Applications, vol. 469, pp. 173 – 182, 2017.

[11] M. Selway, G. Grossmann, W. Mayer, and M. Stumptner, "Formalising natural language specifications using a cognitive linguistic/configuration based approach," Information Systems, vol. 54, pp. 191 – 208, 2015.

[12] B. Duthil, F. Trousset, G. Dray, J. Montmain, and P. Poncelet, "Opinion extraction applied to criteria," in DEXA, 2012, pp. 489–496.

[13] K. Chen, Z. Zhang, J. Long, and H. Zhang, "Turning from tf-idf to tf-igm for term weighting in text classification," Expert Systems with Applications, vol. 66, pp. 245 – 260, 2016.

[14] H. J. Escalante, M. A. Garca-Limn, A. Morales-Reyes, M. Graff, M. M. y Gmez, E. F. Morales, and J. Martnez-Carranza, "Term-weighting learning via genetic programming for text classification," Knowledge-Based Systems, vol. 83, pp. 176 – 189, 2015.

[15] P. Rosso, S. Correa, and D. Buscaldi, "Passage retrieval in legal texts," The Journal of Logic and Algebraic Programming, vol. 80, no. 3, pp. 139 – 153, 2011.

[16] D. Dubois and H. Prade, "Fuzzy sets and statistical data," European Journal of Operational Research, vol. 25, 1986.

[17] A. Imoussaten, G. Mauris, and J. Montmain, "A multicriteria decision support system using a possibility representation for managing inconsistent assessments of experts involved in emergency situations," IJIS, vol. 29, no. 1, pp. 50–83, 2014.

[18] C. Morris, Foundations of the theory of signs, ser. International encyclopedia of unified science. University of Chicago Press, 1938.

[19] B. Duthil, F. Trousset, M. Roche, G. Dray, M. Plantié, J. Montmain, and P. Poncelet, "Towards an automatic characterization of criteria," in DEXA, 2011, pp. 457–465.

[20] S. Ranwez, B. Duthil, M. F. Sy, J. Montmain, P. Augereau, and V. Ranwez, "How ontology based information retrieval systems may benefit from lexical text analysis," in New Trends of Research in Ontologies and Lexical Resources. Springer, 2013, pp. 209–231.

[21] G. Mauris, L. Berrah, L. Foulloy, and A. Haurat, "Fuzzy handling of measurement errors in instrumentation," IEEE Transactions on instrumentation and measurement, vol. 49, no. 1, pp. 89–93, 2000.

[22] S. Destercke, D. Dubois, and E. Chojnacki, "Possibilistic information fusion using maximal coherent subsets," IEEE Transactions on Fuzzy Systems, vol. 17, no. 1, pp. 79–92, 2009.