Accepted Manuscript

A hybrid integrated architecture for energy consumption prediction

Alejandro Maté, Jesús Peral, Antonio Ferrández, David Gil, Juan Trujillo

PII: S0167-739X(16)30064-4
DOI: http://dx.doi.org/10.1016/j.future.2016.03.020
Reference: FUTURE 2994

To appear in: *Future Generation Computer Systems*

Received date: 12 June 2015
Revised date: 19 March 2016
Accepted date: 21 March 2016

Please cite this article as: A. Maté, J. Peral, A. Ferrández, D. Gil, J. Trujillo, A hybrid integrated architecture for energy consumption prediction, *Future Generation Computer Systems* (2016), http://dx.doi.org/10.1016/j.future.2016.03.020

*Highlights (for review)

Bullet points

- Energy consumption predictions based on data mining and supported by external data.
- Heterogeneous data are combined through DW and Information Extraction (IE).
- Multidimensional model integrates information extracted from Social Networks and IE.
- The scenario: consumption prediction is modified with external unstructured data.

A hybrid integrated architecture for energy consumption prediction

Alejandro Maté[a], Jesús Peral[a], Antonio Ferrández[a], David Gil[b], Juan Trujillo[a]

[a] *Department of Software and Computing Systems, University of Alicante, Spain*
[b] *Department of Computing Technology and Data Processing, University of Alicante, Spain*

Abstract

Irresponsible and negligent use of natural resources in the last five decades has made it an important priority to adopt more intelligent ways of managing existing resources, especially the ones related to energy. The main objective of this paper is to explore the opportunities of integrating internal data already stored in Data Warehouses together with external Big Data to improve energy consumption predictions. This paper presents a study in which we propose an architecture that makes use of already stored energy data and external unstructured information to improve knowledge acquisition and allow managers to make better decisions. This external knowledge is represented by a torrent of information that, in many cases, is hidden across heterogeneous and unstructured data sources, which are recuperated by an Information Extraction system. Alternatively, it is present in social networks expressed as user opinions. Furthermore, our approach applies data mining techniques to exploit the already integrated data. Our approach has been applied to a real case study and shows promising results. The experiments carried out in this work are twofold: (i) using and comparing diverse Artificial Intelligence methods, and (ii) validating our approach with data sources integration.

Keywords: Data mining, energy consumption, Information Extraction, big data, decision trees, social networks

1. Introduction

Energy resources resources have been used irresponsibly and negligently in the past five decades. Therefore, the European Union [22] [56] [9] [67] and other international organizations [1] [7]are promoting initiatives to foster responsible and efficient actions to ensure the sustainability of cities.

Several initiatives have highlighted how Information and Communication Technologies (ICTs) can be used to achieve cities' climate targets by lowering energy use and greenhouse gas (GHG) emissions from other sectors [75] [70]. Some of these initiatives include proposals such as dematerialisation and demobilisation, as well as comprehensive concepts for smart logistics and smart cities [50]. Hilty [37] describes how ICT can be seen as an enabling technology for improving or substituting processes in other sectors. According to the report 2020 [72] of the Climate Group and McKinsey which focused on the potential for reducing GHG emissions in six different sectors: power, transportation, agriculture, building, manufacturing and consumer and services), ICT is a key player in the battle against climate change and offers the possibility of 7.8 Gt reduction of CO_2 emission in 2020. In addition, the European Commission is stressing the importance of ICT for energy reduction and sustainability, and invests in research in this area [59]. Therefore, a number of research programmes combining ICT with energy have been developed.

According to the Climate Group, a smart city is a city that uses data, information and communication technologies strategically to provide efficient services to citizens, monitors policy outcomes, manages and optimises existing infrastructure, employs cross-sector collaboration and enables new business models [72]. Smart cities generate increasingly huge amounts of data which traditional data processing applications are incapable of dealing with. This is the reason why Big Data appears associated to smart cities, and represents a hot research topic [25] [76]. The overall aim of this paper is to explore the opportunities of using ICT as an enabling technology to predict energy consumption in cities.

A more efficient use of energy is to minimize the energy loss caused by inaccuracies in energy prediction. Energy prediction already makes use of large volumes of data in order to make more accurate estimations. However, improving energy prediction further requires introducing streaming and highly heterogeneous information into the estimation process. While the technologies for processing this kind of information already exist, they have not yet been adapted for improving energy consumption. Therefore, more specifically, the main aim of this paper is to explore a novel architecture for predicting energy consumption in cities, based on the combination of internal data already stored in Data Warehouses with external data gathered from Big Data sources.

This paper presents a study in which we propose a hybrid architecture that is able to integrate on the fly highly heterogeneous data sources, enabling the use of current energy information combined with contextual data which are integrated with a variety of information sources or Big Data. The benefit is to improve knowledge acquisition for better decision making. The objective of this paper is to make some predictions about energy consumption based on energy data mining and supported by knowledge that provides a torrent of information in many cases hidden across heterogeneous and unstructured data sources.

The remainder of this paper is structured as follows: in the next section, the related work is reviewed; thereafter, we design, create and implement an architecture which includes an integrated model and uses data mining to access Big Data in combination with other data sources. An explanation of the architecture follows in the next section with a case scenario. Subsequently, the analysis of the results section evaluates our proposal. Finally, we include a discussion of the advantages of the model and the difficulties related to its implementation as well as the further areas of research in this field.

2. Related Work

In the following subsection, the main ICT solutions for smart cities are overviewed. In the second one, the previous work on energy consumption prediction is summarized. In the third one, given that our approach uses ontologies, the main ontologies in the domain of energy are outlined. Finally, the contributions of our proposal are enumerated.

2.1. ICT solutions for smart cities

Although very often there is no explicit connection between smart and sustainable cities, it is obvious that ICT plays an essential role for supporting the transition to more sustainable cities, not only regarding the management of urban systems but also offering more support for sustainable urban lifestyles. Mitchell [50] has defined five main opportunities for how ICT can contribute to the reduction of energy use in cities. Four of these have direct effects and one has indirect effects on the reduction of energy use. For this purpose he makes use of opportunities. The first one is labelled as dematerialisation. Here, physical products or services are converted to digital ones (we can imagine how the CDs are now streamed music and the bank offices are online banking services mainly). According to Hilty [37] software represents the immaterial resources and the services provided represent the value that could become the pattern for a discontinued economy. The second opportunity is demobilisation, where everything that has been digitalised can be transported via the telecom network instead of being physically transported. We are now aware how transport and travel are totally or partially replaced by telecommunications. The third opportunity is mass customisation where less resource use is accomplished through intelligent adaptation, personalisation and demand management. The fourth opportunity, intelligent operation, involves more resource-efficient operations of, for instance, water, energy and transport systems. The fifth and indirect opportunity is soft transformation where the existing physical infrastructure is transformed because of new opportunities presented by the information paradigm. These principles can be applied to product design, architecture, urban design and planning at regional, national and global levels [50]. The Smart 2020 report [29] identifies ICT solutions by combining the abatement potential of ICT (called change levers) with economic end-use sectors. Somewhat similar to the opportunities put forward by Mitchell, the change levers are 1) digitalisation and de-materialisation, 2) data collection and communication, 3) systems integration and 4) process, activity and functional optimisation.

With the aim of mapping out ICT solutions which have the potential to offer beneficial environmental effects, [37] used a combination of economic sectors and environmental indicators to compile a list of ICT solutions for sustainable development: e-business, virtual mobility (teleworking, teleshopping, virtual meetings), virtual goods (services partially replacing material goods), ICT in waste management, intelligent transportation systems, ICT in energy supply, ICT in facilities management and ICT in production process management. In addition to this, The Climate Group [72] proposes a comprehensive list of possible ICT solutions that can be implemented, as well as setting out metrics in order to understand which solutions could be implemented to reach a specific city's goals.

2.2. Antecedents on energy consumption prediction

Within these ICT solutions, we are focusing on energy supply, specifically on the prediction of energy consumption in cities. In [38], a brief overview of over 100 years of energy forecasting practices is reported, which concludes that the electric power industry needs forecasts of supply, demand and price, so called energy forecasts, to plan and operate the grid. While many other industries have some form of inventory to store and buffer their products and services, those of the electric power industry, electricity, cannot be massively stored using today's technologies. As a result, electricity has to be generated and delivered as soon as it is produced. In other words, the utilities have to balance the supply and demand every moment.

The storage limitation and societal necessity of electricity lead to several interesting features of energy forecasting, such as the complex seasonal patterns, 24/7 data collection across the grid, and the need to be extremely accurate. In [4], a review and categorization of electric load forecasting techniques is presented. These techniques are classified there into nine categories: (1) multiple regression, (2) exponential smoothing, (3) iterative reweighted least-squares, (4) adaptive load forecasting, (5) stochastic time series, (6) ARMAX models based on genetic algorithms, (7) fuzzy logic, (8) neural networks and (9) expert systems. The state-of-the-art on energy consumption prediction techniques depends on the type of consumption at hand. Ideally, having all the information about the system would yield highly accurate results. However, it is the case that most often all information is not available, and thus, approximate models must be built.

The research line about Energy Informatics (EI) is related to this issue, which means to increase the efficiency of energy demand and supply systems through scientific investigations [71]. In particular, the overall domain of EI concerns itself with the appropriate analysis of available information in order to optimize the performance of these systems. EI can be represented by the equation: $Energy + Information < Energy$, in which the idea is that we can make better decisions about how to both use and conserve energy through the use of information. Starting from the most detailed level, building energy consumption accounts for 40% of global energy consumption [68]. Building consumption depends mainly on the behaviour of building occupants. However, as there is no direct way to measure their behaviour, models try to simulate it by using stochastic models [68] or advanced behavioural models [12]. When modelling large groups of buildings however, more efficient techniques are required in order to simulate the consumption of the whole sector. In these cases, distinctions are made between for either simulating residential sector consumption [62], industrial, agriculture and services consumption [34] or, at the highest level, the whole energy consumption of the country [27]. It is important to note that the largest group considered is country level, since energy demand is tailored for the characteristics of each country [61]. Techniques used for modelling large groups are usually focused on bottom-up statistical analysis, as econometrics has been shown to yield poor results [62].

The evolution of modelling techniques can be overviewed in the Global Energy Forecasting Competitions (GEFCom), which is a competition that requires participants to develop models and submit forecasts based on a given data set. One of the most common and effective techniques mentioned in literature [5] [31] is the Artificial Neural Network (ANN). ANN takes a set of input data and tries to estimate the output by adjusting the weights between neurons in the network. After several training iterations, the ANN can accurately predict the target function depending on how much information is provided by the inputs. Depending on the proposal ([34] [27] [5]) the ANN is fed with different information, from the Energy Performance Index (EPI) of the buildings to billing information, including atmospheric conditions for estimating the variation between cold and warm hours and days.

An alternative option to ANNs is to use forecasting models [61], including time series, ARIMA (autoregressive integrated moving average) models and Grey prediction among others. These techniques try to predict the aggregated demand at country level for a certain resource, whether electricity, gas, oil, etc. Inputs used for these techniques usually include GDP, income levels, temperature, and energy prices.

Once we have reached country level, further differentiation can be made across models proposed depending if they try to predict energy consumption in the short-term [3], medium-term [60] or long-term [64]. On the one hand, Short-term predictions require quick reactions against changes, like changes in weather conditions affecting renewable energy sources, natural disasters, sudden behaviour changes, etc. While inaccuracies are allowed, they have a negative impact for energy suppliers, as the energy that is not used is lost. On the other hand, Medium and Long-term predictions are used mainly for strategic planning. Medium-term demand prediction helps electrical companies to prepare for expansion and anticipate the need for new power plants. Long-term demand prediction is used by governments in order to ensure that the country will be adequately supplied in the future, and identify the need to secure and exploit new energy sources.

Given the strategic nature of these predictions, they require to be as accurate as possible as they have a strong influence on the decisions regarding the energy supply capability in the future. While most prediction techniques are fairly accurate as long as they have enough information available, medium and long-term predictions have been shown to be inaccurate [65]. The main reason found for this discrepancy has been the inability to predict the technological evolution of energy generation technologies.

2.3. Ontologies used in the domain of energy

In [20], the benefits of using ontologies in the domain of Energy Informatics are overviewed. They outline different ontologies, formalized, domain-specific taxonomies or vocabularies used in this domain. For example, the OpenEI[1] refers to itself as a free, open source knowledge-sharing platform for data, models, tools, and information related to clean and renewable energy systems [14] [77]. It is built on top of the Semantic MediaWiki platform which enables its information to be exposed as an RDF graph for semantic information interchange purposes such as Linked Open Data [46]. It is sponsored by the US Department of Energy and developed at the National Renewable Energy Lab (NREL).

The Reegle's extensive glossary links all of the terms about clean energy information that targets specific stakeholders, including governments, project developers, businesses, financiers, NGOs, academia, international organizations and civil society [57] [8].

The system architecture proposed in [36] includes an ontology model for efficient building energy management systems with concepts for sensors, equipment, zones, buildings, equipment action, zone evaluation, etc. It also includes a good description of how to build more inference rules into the ontology reasoning process via simulation.

In [42], two ontologies for agent-based modeling of energy systems, the ontology for Socio-Technical Systems (STS) and the Synthetic City (SynCity) ontology for urban energy systems, are compared. The first ontology, STS, takes a network approach to cross-domain policy (i.e., not just energy-related) modelling [66]. It has been used to develop multiple models related to energy policies and would probably serve as a good source for some policy-related concepts. The second ontology, SynCity, is interesting because it provides three major components for modeling purposes: a mixed-integer linear programming (MILP) optimization model for housing layouts, an agent-based energy demand model and a MILP optimization model for combining the other two models [43]. SynCity serves more as a knowledge base (i.e., a collection of instances of ontological concepts) than just a schema for concepts and does not take advantage of many of the reasoning capabilities and consistency checks available for ontologies.

In [21], an ontology for the ICT domain that is related to energy consumption is proposed. The novelty of this ontology is the conceptual difference between "Green Energy" (e.g., solar, wind, etc.) and "Dirty Energy" (e.g., natural gas, heating, etc.). Incorporating these kinds of relationships with the concepts being added to OEI may be worth investigating. Again, this ontology does not extend any kind of formal high level ontology to root its concepts. It is also only designed in RDF and not Web Ontology Language (OWL) and therefore lacks some of the reasoning capabilities facilitated through the use of OWL ontologies.

2.4. Contributions of our proposal

As can be drawn from this section, energy consumption predictions, both in the short, medium and long-term, require as much information about the system as possible (e.g. weather conditions or natural disasters) in order to yield highly

[1] http://openei.org (visited on 4th of June, 2015).

accurate results. The inaccuracies in the predictions have a negative impact for energy suppliers, as the energy that is not used is lost. This information is mainly obtained from structured databases and data warehouses. However, information hidden in unstructured information is usually not exploited (e.g. the Web, textual fields in databases or Social Networks). Our proposal, a multidimensional hybrid architecture, makes use of current energy data and external information to improve knowledge acquisition in order to help managers make better decisions.

There are several advantages of our proposal:

- Distribution: allowing querying one or multiple nodes of information seamlessly, thanks to the distributor/integrator module;
- Flexibility: nodes can be added or removed as necessary with minimum changes to the schema;
- Diversity: supports heterogeneous sources by transforming the data provided by each of them into a common representation, and
- Integration: Both historical and unstructured external data is queried without the risk of polluting already clean data sources. This enables us to easily include or exclude information from the analysis, thereby guaranteeing data quality from external unstructured sources and ensuring effective integration.

Furthermore, these benefits would enhance decision making by allowing the inclusion of unstructured information. This information is included by means of specialized nodes, such as the Information Extraction node, integrating information from unstructured sources such as Twitter[2] into the prediction model, and allowing us to store and compare the results of the predictions using different kinds of data.

3. Multidimensional hybrid architecture

Multidimensional hybrid architecture aims to support the Data Mining (DM) process for obtaining energy consumption predictions. The historical consumption and energy generation data stored in a Data Warehouse (DW) are the main data source for the DM process. However, there is a wide variety of available external data sources with potentially useful knowledge for generating predictions of energy consumption, which are presently known as Big Data [19] [80]. However, the quality of the data from these sources is not guaranteed, and therefore integrating it within the DW would lower the quality of the information stored.

Our proposal will overcome this obstacle by creating a specialized architecture, one that allows us to query both historical and unstructured external data

[2] http://twitter.com (visited on 4th of June, 2015).

without the risk of polluting already clean data sources, and enables us to easily include or exclude information from the analysis. For example, we can access and query data generated by a system for recovering and extracting information from data web pages, which contain facts on natural disasters. This information can be extremely valuable when facing abnormal situations, for which typically there is not enough information available to obtain a precise prediction of energy consumption. However, we may find that the system encounters errors in processing information and we may need to temporarily exclude such information from the analysis until we can validate its accuracy and incorporate it effectively.

In [49] we proposed an extended metamodel for creating multidimensional data models that support the definition of joint internal and external information. By definition, multidimensional models allow us to join information by structuring it into facts and dimensions. Facts are the center of analysis, and contain fact attributes (also called measures) that evaluate the performance of a certain activity being analyzed. Dimensions provide context information during the analysis of measures, and can be structured in hierarchies to provide varying levels of aggregation. However, multidimensional models deal poorly with partially structured information, thus constraining their applicability in certain situations where external information is required. In order to overcome this problem, our extended metamodel performs the following functions: i) decouples the information schema from the sources to be queried, thus easily allowing including or excluding sources; and ii) represents not only the multidimensional structure of the DW, but also union points with external sources. This allows us to join internal and external data into a new relation whenever a query is posed, while avoiding the pollution of the DW with low quality data.

Our architecture is based on the implementation of the Map-Reduce paradigm [23]. First, a query is posed to the different systems involved (e.g. Information Extraction, DW, etc.) using a common language. For this task we propose to use SPARQL [54], which will be translated by each node into its own language if necessary, as is the case of the DW node [49] [63]. Then, results obtained by each node are integrated. The integration is achieved thanks to the introduction of a universal multidimensional schema containing the union points to join the information distributed across the different data sources.

In Figure 1 we can see the universal multidimensional model created for our case study. It is possible to distinguish between local and external elements, marked as *. The local elements are facts, dimensions and attributes of the dimensions present on the DW node. The energy consumption is analyzed in terms of the dimensions Producer, Consumer, Power Source, and Time, all provided by the local DW node.

In addition to this information, the external dimension Event provides additional insights on important events that affected the consumption, such as natural disasters, political changes, crisis, or migratory flows. External dimensions, such as Event, are included in order to enrich the original DW information. However, their structure is not completely covered within the universal schema since they are unstable and can evolve independently of the DW. Therefore,

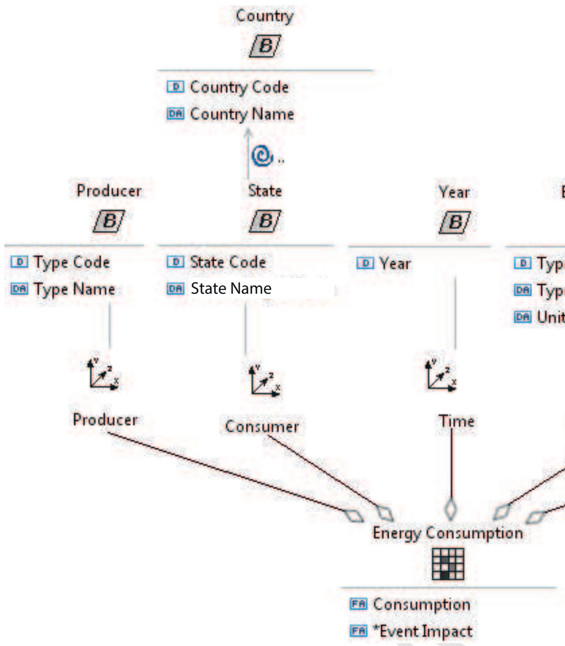


Figure 1: Multidimensional model built to support the integration of the DW data with other external sources.

only the information regarding union points is represented. In our case, Events within the event dimension are hierarchized according to their type. Since external dimensions are joined according to a common key, any changes in the data source providing the data, such as a richer hierarchy, would have no impact on the universal schema nor the system, as the union point would remain stable. Furthermore, the Event Impact Measure has been defined and it represents the influence of the specific event in energy consumption for that year, for that State, type of consumer and producer. For example, this value would represent the effect caused by a natural phenomenon or by the overall opinion on a political event (e.g. financial crisis).

3.1. Implementation of the architecture

Our current architecture is based on the one presented in our previous paper [53], with some important modifications as follows: i) we have extended it with the inclusion of an Information Extraction Node (IE Node) and the Social Networks Node (SN Node); ii) in these nodes, the information needs of the user are previously known in order to extract specific structured-information from unstructured Big Data (e.g. the Web or Social Networks such as Twitter); iii) the new extracted structured information will enrich the DM application (explained in detail in Section 4) in order to reach the final solution (to improve the prediction calculation process).

The architecture is aimed to facilitate the integration of distributed information on the fly. It requires the definition of several schemata in order to perform the integration process: universal, local, deploy, and the use of drivers for each node included in order to translate queries into the appropriate node language.

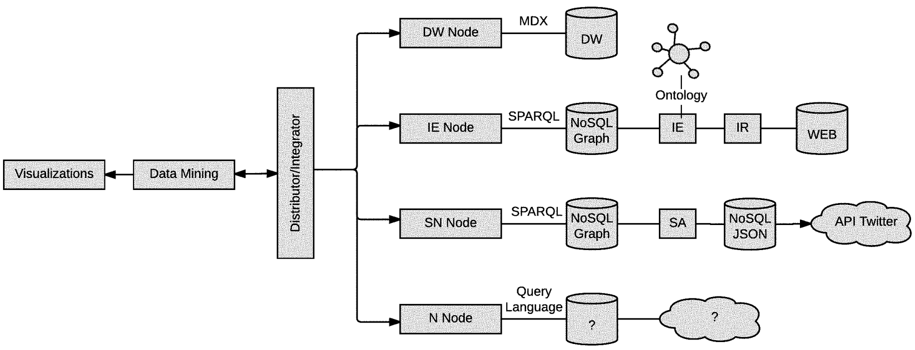


Figure 2: Overview of the proposed architecture to access and integrate structured, unstructured, internal and external data.

An instantiated example of the architecture can be seen in Figure 2. The universal schema, such as the one in Figure 1 represents the structured information stored and the union points with the unstructured/external information. It is stored in the Distributor/Integrator module, the main component in the architecture. Conversely, the local schemata describe the minimum structure required for the external information to be joined with the clean information and are stored in each node. Finally, the deployment schema (omitted) relates multidimensional entities in the universal schema to the nodes from which it should be extracted. When a query arrives at the distributor/integrator, the query is decomposed according to the universal schema and the deployment schema. Then, each subquery is sent to all the nodes that are responsible for extracting the information. Afterwards, the results are returned and integrated into a single relationship that is passed back to the corresponding user or the application (in our case it will be the DM application). Thanks to this architecture, if a new node is added or deleted, it is enough to update the corresponding schemata in order to include/exclude the information it requires from the querying process.

Our architecture is designed to allow the addition of new nodes to the system, which provide access to Big Data sources, in order to gain as much useful information for the DM process. In our particular case, we have added two new nodes to allow recovery and extraction of information from external Big Data sources. These are the SN and IE nodes, as shown in Figure 2.

3.2. Sample execution and technology selection

In order to illustrate how the architecture operates, we will provide a sample execution using the concrete architecture in Figure 2. The process starts with the user posing a SPARQL query to extract the information about the electricity consumption from the system. This query is received by the distributor/integrator:

SELECT ?event ?state ?year ?consumption GROUP BY (State, Year).

The variables in this initial query are not bound, since the purpose of the initial query is to be divided into several SPARQL queries, one for each node involved. Therefore, the distributor/integrator will send the following queries for the DW and the IE nodes:

Node: DW
SELECT ?state ?year ?consumption WHERE{
?state a ***universalSchema****:State .*
?year a ***universalSchema****:Year .*
?consumption a ***universalSchema****:Consumption .*
}

Node: IE
SELECT ?event ?state ?year WHERE{
?event a ***IESchema****:Event .*
?state a ***universalSchema****:State .*
?year a ***universalSchema****:Year .*
}

These individual queries are received by each particular node. Our architecture takes into consideration the fact that not every technology uses SPARQL as the querying language. Therefore, we include a driver on each node which receives the query and translates it into the appropriate language for that particular node; currently, the architecture supports translating SPARQL into MDX (and thus, indirectly SQL), as shown in [53]. Other querying languages or APIs would need to be mapped from SPARQL in order to be supported. Consequently, at the moment the architecture supports any database (whether SQL or NoSQL such as Neo4j) or platform that is compliant with the standard specifications of SPARQL, MDX, and SQL querying languages. Other technologies would require a mapping before being supported.

After the information has been retrieved from each node the distributor/integrator integrates it using the “state” and “year” information as indicated in the initial query, adding the corresponding event (if any) and consumption to each tuple. The result of this sample query is all the information available regarding electricity consumption across the different states and years together with the information regarding any disasters that may have occurred.

Table 7 in Appendix shows a summary of the technologies involved in the architecture and it also includes a justification of the choices. This table summarizes the variety of technologies required by our proposal, which can be organized into two categories: firstly, those used for the structured information (the first four rows in Table 7); and Secondly, those used for the unstructured information (the remaining rows in Table 7). The benefits of the technologies selected and reported in this table will be proved in the experimentation section.

The iterative methodology for setting up the system is detailed in the following subsections given that adding new nodes to the system and generating predictions of energy consumption is a complex process.

3.2.1. Phase 1: Setup

In this phase, the setup of the different nodes that comprise our system takes place. In addition to the traditional configuration of the different nodes, during this phase, the acquisition, pre-processing and storage of large volumes of data (as Big Data [19] [80]) are performed. This will allow us to expedite the query and integration during the Integration Phase. For example, when the SN node proceeds to obtain information about a given topic (e.g. Political changes) from the opinion of users in a social network such as Twitter, it will need to download and process a large volume of historical data or tweets in a semi-structured format (JSON in the case of Twitter). In this case, the computational cost of data retrieval and extraction processes is high, and we believe that such processes must be carried out in a pre-integration phase. Below, configuration processes taking place in each of the nodes of our system are described.

(1) The DW Node. We start with an already defined data warehouse which has been loaded with historical data about energy consumption. The multidimensional data model of this node is defined using the model presented in Figure 1 [49] [63], considering all its elements (measures, dimensions and hierarchies of the dimensions) as internal elements.

During this setup phase, we can: i) update the information of the DW; or ii) modify the elements of the data model in case we need to add new measures, dimensions or hierarchies.

(2) The IE Node. Information Extraction (IE) is the task of automatically extracting specific structured information from unstructured and/or semi-structured machine-readable documents. Each IE application is specifically designed for each particular extraction process(usually guided by domain ontologies)which fills the slots of a set of predefined templates that determine the information being searched in the collection of documents. However, part of the template can be dynamically generated, for example, when the IE template is defined to extract “natural disasters”, it can be dynamically refined in order to search “natural disasters in 2012” or “in Detroit”. In our case scenario, the IE Node will scan unstructured documents in order to extract previously defined data that can enrich the consumption prediction process (e.g. natural disasters, governmental or legislative decisions, wars, economic crisis, recession, population movements, etc.). This extracted RDF data will populate a NoSQL [35] Neo4j[3] database.

[3] http://neo4j.com/ (visited on 4th of June, 2015).

Neo4j supports the SPARQL query language natively, unlike what happened in the DW Node. As a result, it is not necessary to implement any type of translation between SPARQL query language and another on this node. Furthermore, the recovery, pre-processing and storage of data speed up the query and integration will be performed in Phase 2.

The unstructured documents are selected from specialized webs. Given that IE applications are computationally costly, IE applications are forced to be run on small datasets. Therefore, a previous Information Retrieval process must be run (e.g. through search engines such as Google) in order to filter the Big Data documents. For example, the information requirement of "natural disasters in Alabama State in 2012", is posed to Google, jointly with additional keywords selected by the IE Node (e.g. natural disasters, governmental or legislative decisions, wars, economic crisis, economic depression, population movements).

(3) The SN Node. The SN Node (Social Networks Node) is configured to extract opinion information on a topic from data generated in social networks like Twitter. To achieve an application that can exploit the available API's (Application Programming Interface) is required in order to retrieve the tweets that match a search filter (query). In this case, given the complexity of queries to be performed, due to both the characteristics of the API and the tweets themselves, the user must enter the query in a parameterized way (sets of words, words to exclude, hashtags, location, tweet language, etc.). Tweets recovered according to the posed query are stored in a NoSQL MongoDB documental database, whose native document format is JSON.

Once the tweets are stored in the MongoDB database, different applications can be used to extract the relevant information. For instance, a sentiment analysis algorithm [6] can be applied to the texts of the tweets in order to determine the degree of positivity (or negativity) of each tweet. The extracted opinion information is stored along with other relevant data from the tweets, such as country, state or year of publication, in another database created in Neo4j. In this way, we support natively the SPARQL language and expedite the process of querying and integration that will take place in Phase 2, as in the case of the IE node.

(4) The N Node. The N node represents the possibility of adding new nodes to the system. As previously described with other nodes, in this first phase the N node will be configured according to its characteristics. In the event that the node has to download and process large volumes of Big Data information, it will take place in this first phase, storing the preprocessed data in a repository (e.g. a NoSQL database). Thus, the query and integration of data will be expedited in Phase 2. However, if such a repository does not support the SPARQL language natively, the node will need to implement the translation between SPARQL and the query language supported by the repository.

Furthermore, it is necessary to amend the universal schema and establish the links between the new external elements and the N node in order to join the information using new union points. This is necessary so that during the Integration Phase, the Distributor/Integrator module is able to distribute to each node the corresponding sub SPARQL query and then it can integrate the results received from each node.

3.2.2. Phase 2: Query and Integration

After configuring the various nodes that are part of the system, in this second phase, the query posed by the user in the GUI module is processed and the data provided by the various nodes are integrated and returned in a format supported by the DM module.

For instance, the user (e.g. the analyst or data scientist) poses a query to the GUI in natural language, according to the information required by the DM process to generate energy consumption predictions for a set of years or countries. Then, the query is translated into SPARQL by the GUI and is transferred to the distributor/integrator module. This module, based on the defined universal schema, is capable of generating the corresponding sub SPARQL query to be sent to each node, depending on the information that each node is capable of handling. The nodes process the query and return the information as RDF triples to the distributor/integrator module, which is responsible for integrating data using the required algebraic operations (e.g. unions or intersections). Finally, the result is transferred to the GUI, which is responsible for generating the data received in the supported format by the DM module.

3.2.3. Phase 3: Generating the forecasting

At this time, the user already has the required data in the GUI and can use them as a source of data in the DM module. In our case, the DM module is implemented in Java, which makes use of the algorithms provided by Weka [33]. This application generates predictions and produces the most suitable visualizations for the analysis.

From the extracted knowledge of the generated predictions, we can identify the need for: i) new data sources; ii) changes in the universal scheme to include new elements that will enhance the integration and analysis of the available data sources; iii) storing the results of the predictions in the DW for integration with data from other nodes in the system in subsequent runs of the process. Therefore, we consider this 3-phase process as an iterative process, where at any time we can go back to perform the entire process or repeat any of the phases. Through the progressive refinement of the process, we will be able to improve the accuracy of the energy consumption predictions generated by our system.

Although for the sake of clarity we have described the system as being deployed in a particular server, both the architecture and its functionalities could be deployed in the cloud, turning the forecasting architecture into a service. This approach, often followed for Big Data solutions [17, 18], can improve the availability [17] and performance [18] of the platform, which is important for consumption prediction since it has to constantly provide information on power generation, and thus any downtime or, worse, a disaster, can have a severe impact.

4. Case scenario

4.1. Overview

Now that we have introduced the system architecture, in the following four subsections we explain the application of our framework to the scenario in which the objective is to carry out consumption predictions using the data described in subsection 4.2. We have followed the design science research methodology (DSRM), [69] [52], in order to design and evaluate our proposal (sections 4 and 5).

First of all, experiments with structured data (subsection 4.3) were performed in order to predict energy consumption. The purpose of the following subsection is to show the improvements obtained in forecasting through the inclusion of semi-structured information on natural disasters. Finally, subsection 4.5 presents the information enrichment with unstructured data from the information on opinions extracted from social networks.

4.2. Data description

The data used in this case study[4] have characteristics that allow it to be classified as big data:

- Volume. The data obtained for this study have initially been taken from US Energy Information Administration (EIA)[5] between 1990 and 2013. They already contain a huge volume of information although they can always be extended collecting more data through the website, in this example, extending the period of years.

- Variety. The greatest challenge is extracting useful knowledge from heterogeneous data. This complexity can be seen in the following three subsections of this case study.

In these, knowledge is extracted from structured (EIA data), semi-structured (meteorological phenomena from NOAA-NCDC, National Oceanic and Atmospheric Administration-National Climatic Data Center[6]) and, finally, unstructured (Social Networks, such as Twitter) data. Our objective in this section is not only to deal with different types of sources (structured, semi-structured and unstructured) but also to be able to demonstrate how our architecture performs given the diversity of information (in our case study, information about energy consumption, natural disasters, and the point of the economic cycle).

[4] Data are openly accessible in https://drive.google.com/open?id=0B5THYjdVsQyEdGxsakxJS1RHOWs (visited on 18th of March, 2016).

[5] http://www.eia.gov (visited on 4th of June, 2015).

[6] http://www.ncdc.noaa.gov/ (visited on 4th of June, 2015).

- Velocity. This is not a priority in our case study, although our architecture is able to run shorter-term predictions. The distributed architecture system allows us to process the different data sources in parallel in order to store and process the required information“in real time”.

- Veracity and value. Regarding the last two Vs ([48] [24]) while these are not explicitly contemplated in our data types, they are fulfilled implicitly.

1. Veracity depends on the type of data: (a) structured, the data comes from the website of the energy administration (EIA); (b) semi-structured, the textual data comes from NOAA-NCDC sources and therefore we need to apply Natural Language Processing techniques; (c) unstructured, our twitter filter processes will be applied in pre-processing. For the first, the data are correct for the source itself; for the latter two cases, as discussed in the following two sub-sections, filters allow us to eliminate inaccurate information.

2. Value is created through the relevant knowledge extracted that can be used to meet our objective of improving energy prediction. As shown in Figure 2, the IE and SN nodes will allow us to extract more information and enrich the initial value of the DW node. This will be corroborated in the following sections.

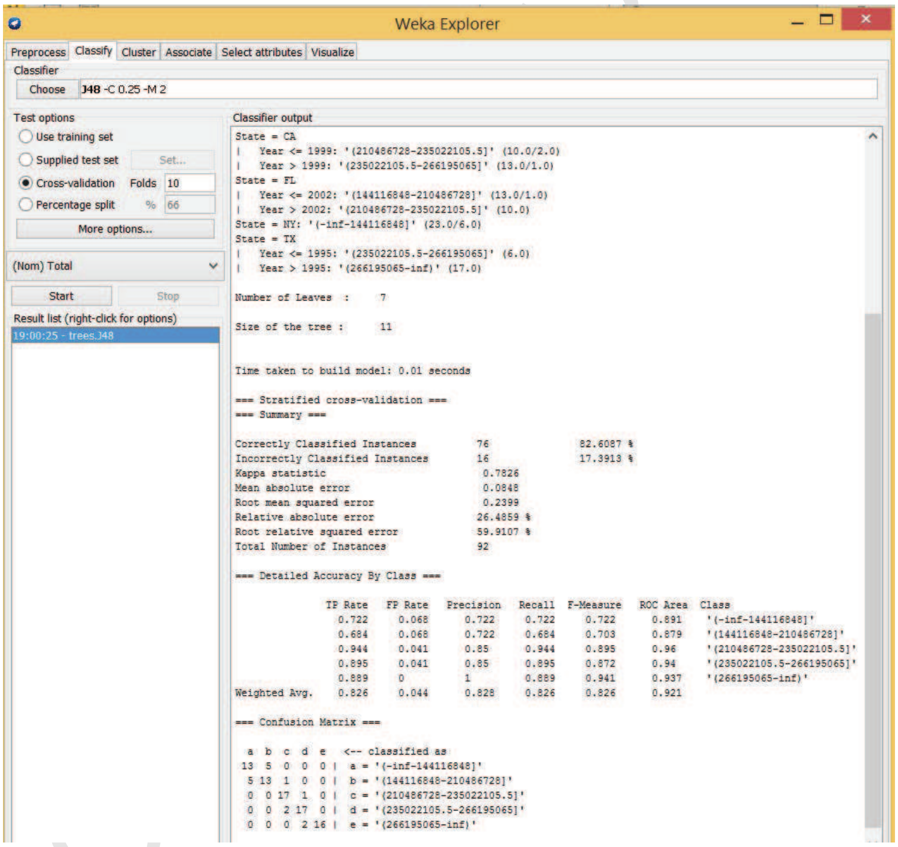
4.3. Structured data

First of all, we have carried out the experiments with the structured data from the EIA. We have used data that contain information on energy consumption for the different US states each year since 1990. These data have been loaded in a specific DW which contains historical data about energy consumption (DW is a node of the proposed architecture, Figure 2). The experiments were performed using Weka and referred to the DM module (Figure 2) only taking into account the internal data.

In order to predict the “consumption for electricity” variable, we start with a pre-processing step to discretize that numerical variable. Here it is necessary to work with DM prediction algorithms. In our case, we have experimented with several numbers of discretizations, specifically 5, 10, 15, 20 and 25 ranges or bins. Experience obviously tell us that the more we discretize the more imprecise the accuracy (i.e there are more errors in the instances classified because there are more output/classes and it is easier to misclassify). With these structured data we have carried out several tests with different DM algorithms, such as decision trees (DT), artificial neural networks (ANN) and support vector machines (SVM). The aim is to predict the electricity consumption for a year in a specific US State. The best performance was obtained with 5 ranges of discretizations[7]. The accuracy with the different AI techniques is similar and so we will describe only two of these which are the most representative in terms

[7] In a real case, and depending on the data, it is possible and recommendable to adjust, and especially increase, the number of output ranges to limit the prediction range. There is always a balance between accuracy and the selected number of ranges. In addition, in future work, the proposed method using classification techniques could be combined with regression algorithms to improve outcome.

of understanding. Both are DT methods and included in the Weka package: C4.5 [55] and RandomForest [79]. Figure 3 shows the Weka environment when Tree C4.5 has been selected. We can see several parameters and variables. For example, we can see the number of leaves (7), the size of the tree (11 nodes) and the energy consumption prediction for the different US states (consumption has been measured in Megawatt hours). In order to simplify the case study, we have taken the most representative states (California, Florida, New York and Texas) instead of examining them all. Furthermore, we can see information that is very significant, like the confusion matrix (with a clear defined diagonal and only a few errors) and we find a high accuracy of 82.6087% (correctly classified instances) after running the experiments on Tree C4.5.



Classifier
J48 -C 0.25 -M 2

Test options
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Total

Result list (right-click for options)
19:00:25 - trees.J48

Classifier output

```
State = CA
|   Year <= 1999: '(210486728-235022105.5]' (10.0/2.0)
|   Year > 1999: '(235022105.5-266195065]' (13.0/1.0)
State = FL
|   Year <= 2002: '(144116848-210486728]' (13.0/1.0)
|   Year > 2002: '(210486728-235022105.5]' (10.0)
State = NY: '(-inf-144116848]' (23.0/6.0)
State = TX
|   Year <= 1995: '(235022105.5-266195065]' (6.0)
|   Year > 1995: '(266195065-inf)' (17.0)

Number of Leaves  :     7

Size of the tree :      11


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          76               82.6087 %
Incorrectly Classified Instances        16               17.3913 %
Kappa statistic                          0.7826
Mean absolute error                      0.0848
Root mean squared error                  0.2399
Relative absolute error                 26.4859 %
Root relative squared error             59.9107 %
Total Number of Instances               92

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.722     0.068     0.722     0.722     0.722      0.891    '(-inf-144116848]'
                 0.684     0.068     0.722     0.684     0.703      0.879    '(144116848-210486728]'
                 0.944     0.041     0.85      0.944     0.895      0.96     '(210486728-235022105.5]'
                 0.895     0.041     0.85      0.895     0.872      0.94     '(235022105.5-266195065]'
                 0.889     0         1         0.889     0.941      0.937    '(266195065-inf)'
Weighted Avg.    0.826     0.044     0.828     0.826     0.826      0.921

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 13  5  0  0  0 |  a = '(-inf-144116848]'
  5 13  1  0  0 |  b = '(144116848-210486728]'
  0  0 17  1  0 |  c = '(210486728-235022105.5]'
  0  0  2 17  0 |  d = '(235022105.5-266195065]'
  0  0  0  2 16 |  e = '(266195065-inf)'
```

Figure 3: Weka Environment: parameters and variables for Tree C4.5 using structured data.

Figure 4 shows all the variables, and displays the correlation between each of these and the output with the five discretizations of electricity consumption, represented by five different colors. Therefore, we can graphically identify the influence of each individual input variable (year and state) and the energy consumption

Tree C4.5, Figure 5, shows, in the form of a tree, the conditions for predicting

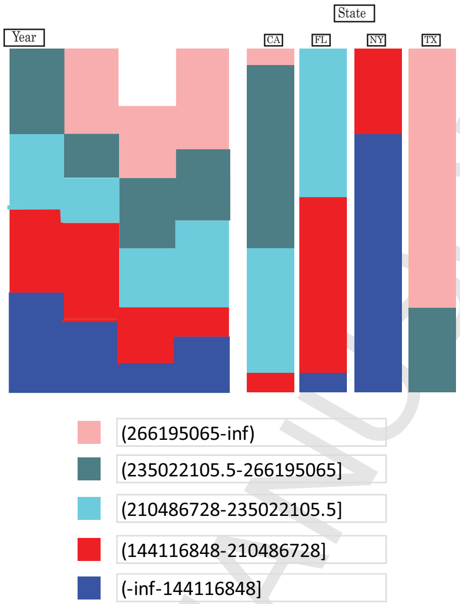


Figure 4: Correlation between the variables with five discretizations of electricity consumption using structured data.

energy consumption. The branches can also be seen as rules or conditions. For example, for the state of Florida (branch marked “=FL”) depending on the specific year (“<= 2002” or “> 2002”) the consumption will be “(144116848-210486728]” or “(210486728-235022105.5]”.

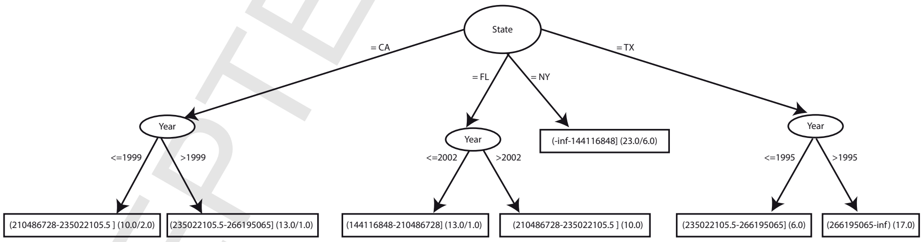


Figure 5: Tree C4.5 output with structured data.

The RandomForest Tree, Figure 6, shows a much denser tree as RadomForest is an advanced algorithm that operates by constructing a multitude of decision trees at training time and outputting the class. This density in the tree allows us to obtain a wider set of rules in much more detail, along with conditions or rules and eventually more knowledge. For example, for the abovementioned Florida consumption, more specific rules have been defined: if the “year” is “<= 2002” and “<= 1990” the consumption will be “(-inf-144116848]” whereas if the year is “<= 2002” and “> 1990” the consumption will be “(144116848-210486728]”, etc.

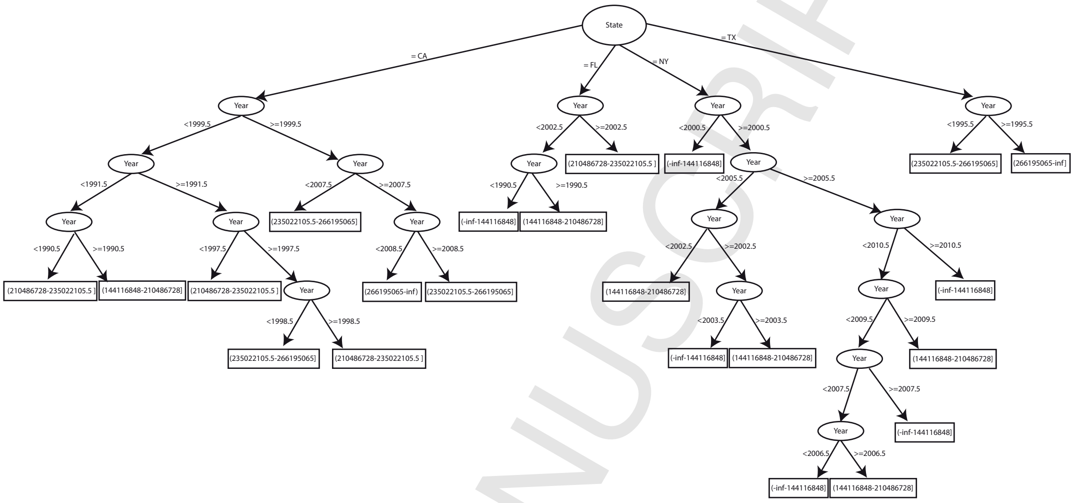


Figure 6: RandomForest Tree output with structured data.

4.4. Semi-structured data: Natural disasters

Once the basic experiments in the previous subsection have been carried out using structured data, the main challenge is to improve not only on the accuracy of the DM module, but also on the forecasting of knowledge, by enriching the decision tree with new variables obtained from the IE node (right side of Figure 2). This complements the initial DM experiments shown in figures 3, 4, 5 and 6.

In this case study, this will be accomplished by using additional information on natural disasters in several data formats in order to find out how these meteorological phenomena can affect energy consumption. There is a huge amount of information recorded on natural disasters, but to limit the problem and simplify the example we have used the essential annual information that consists of location (US state), year, type of natural disaster and description. All of this information on meteorological phenomena in the US since 1990 has been extracted from NOAA-NCDC.

Subsequently, in the textual description field of this database, we have run an IE system to extract more detailed information about these natural disasters (this process has been accomplished in the IE node of our architecture, Figure 2). We have extracted a classification of the different phenomena (e.g. Dense Fog, Heavy Snow, High Wind or Tornado) according to their strength/intensity (category 1, 2 and 3).

The process of deriving numerical values for natural disasters is based on IE and Sentiment Analysis techniques in order to grade the natural disaster according to a previously defined taxonomy (in this case, three strength/intensity categories have been defined). These techniques work on the textual description field of the abovementioned database. This description gives the general nature and overall activity of the episode by reporting additional and specific details of the individual event, as can be seen from the following two examples:

(Example 1) “Vigorous energy cycling around a seasonally strong upper level disturbance crashing into Texas brought a final round of potent thunderstorms through much of the Rio Grande Valley during the late night hours of May 10th into daybreak on May 11th. Initial severe thunderstorms and isolated tornadoes pummeled the Coastal Bend and South Texas Plains from Corpus Christi to near Laredo during the evening of the 10th. As night progressed, pockets of colder air behind the initial line pushed its west end southward; just before midnight, this end would link up with individual cells moving off the Sierra Madre west of Zapata County and form a separate, smaller line from Jim Hogg to Zapata County. This line quickly intensified and dumped heavy rain, high winds some large hail, and one confirmed tornado. As the line marched from the Ranchlands into the Rio Grande Valley, it weakened a bit; as the remnants neared the coast before daybreak, individual small rotating storms developed and redeveloped in Cameron County, with several reports of funnel clouds and at least one confirmed tornado in San Benito (just after 3 AM) from a few hours prior to until just after daybreak on the 11th. While minor flooding was reported across some of the hilly terrain of the Rio Grande Plains, rainfall was more welcome in the Lower and Mid Valley, with many locations between San Benito, Harlingen, and McAllen receiving 1 to 2 inches of soaking. Frequent to excessive lightning strikes knocked out power to portions of the Rio Grande Valley as the intense storms passed; the lightning caused one structure fire in Pharr”

(Example 2) “A significant tornado outbreak occurred across portions of Kingman and Harper counties. In all, nine tornadoes touched down on this day, the strongest being rated as and EF-3. Many other counties experienced large hail and strong winds.”

The equation is quite complex given the complexity of an IE or Sentiment Analysis system (e.g. those quoted in Section 4.4: OpeNER and GPLSI system). It is based on the extraction of textual features that are used in a final formula, whose results are segmented into three intervals: 0-0.32 for category 1, 0.33-0.65 for category 2 and 0.66-1 for category 3. In equation 1, the formula is shown in a simplified form, where IV is the Intensity Value to be calculated through TD (i.e. the natural disaster textual description), number of deaths, injuries, damages, magnitude and PI (i.e. polarity expression of intensity):

$$IV(TD) = \frac{(deaths + injuries + damages + magnitude + PI)}{maximum\_intensity\_value} \tag{1}$$

In equation 1, the formula field *deaths* means the number of deaths due to the natural disaster, which is automatically extracted from the description field of the database. Similarly, the formula field *injuries* refers to the number of injuries caused by the natural disaster. The *damages* and *magnitude* fields are also extracted. For the “polarity expressions of intensity” field, Sentiment Analysis techniques have been applied, in which textual expressions such as “record event” or “impressive disaster” are valued. All these formula fields are normalized to a range between 0 and 1 (e.g. the absolute value of *deaths* is divided by the maximum number of deaths caused; while the *magnitude* field is graded according to ontology information about natural disaster magnitudes).

This information has been included in the DM module for the different states. In the experiments, this extracted information was added to the system as the numerical variable “disasters”, derived from a weighted equation for the various natural disasters taking place each year, as shown in equation 2 (where S is the state for which the disasters variable is being calculated; Y is the year of calculation; n_ds is the number of disasters extracted for the state, year and, in addition, for $C1$, when only category 1 disasters are considered, $C2$ for category 2 and $C3$ for category 3).

$$Disasters(S, Y) = n\_ds(S, Y, C1) \cdot 0.33 + n\_ds(S, Y, C2) \cdot 0.66 + n\_ds(S, Y, C3) \cdot 1 \tag{2}$$

Figure 11 in Appendix shows this new variable, named “disasters”, which makes the tree slightly bigger in terms of including a new variable and this also means new rules and conditions. This is reflected in the new tree that includes this additional variable. In particular, we can note that both the “CA” and “NY” states include this variable in their branches. For instance, the impact of “disasters” on consumption in California can be observed: if the year is “$<=$ 1999” and “disasters” is “$<$ 1410.5” or “$>=$ 1410.5” the consumption will be different.

In order to show the decisive influence of the new variables introduced in the DM module, we have analyzed an example in which the prediction made (without the inclusion of the new variables with external information) failed according to figures 5 and 6: the prediction for NY state in 2013 is “(-inf-144116848]” while the real consumption was 147,895,127. However, if we take into account the new variable then it no longer fails since the information is now more precise.

With the inclusion of the “disasters” variable (in the case of NY this has a value of 1,575 in 2012) we can see in the tree, Figure 11, that for years after 2000 there are only two ways depending on the value of this variable (“disasters $<$ 2915” and “disasters $>=$ 2915”). In the “disasters $>=$ 2915” branch there is only one prediction and it is incorrect. In our example, we would select the “disasters $<$ 2915” branch and different consumption forecasts that match the true NY energy consumption in 2013 (there are several nodes with forecasts “(144116848-210486728]”) in the descendants nodes can be observed. However we do not reach the leaf node with the correct prediction, mainly due to having only introduced information about one external variable related to natural disasters. The enrichment of the system with the inclusion of new variables (economic crises, wars, political changes, etc.) will generate a tree with more specific conditions, leading to more accurate predictions.

4.5. Unstructured data: Economic crisis information extracted from social networks

As presented in [28], the Web 2.0 has become one of the most important sources of data from which to extract useful and heterogeneous knowledge. Texts can provide factual information, such as descriptions and lists of features, and opinion-based information, which would include reviews, emotions, or feelings. This subjective information can be expressed through different textual genres, such as blogs, forums, social networks and microblogs.

One example of a microblogging social network is Twitter, which has gained much popularity over the last few years. This website enables its users to send and read text-based messages of up to 140 characters, known as tweets. This site can be a vast source of subjective information in real time; millions of users share opinions on different aspects of their everyday life. Extracting this subjective information is of great value for both general and expert users. However, it is difficult to exploit, mainly because of the short length of the tweets, their informality, and the lack of context. Sentiment Analysis systems try to deal with the challenges arising from this new textual genre, by extracting the polarity of the opinions expressed in these tweets (positive, negative or neutral).

In our proposal, in order to extract and include relevant information from social networks (a task performed in the SN node, Figure 2), it is necessary to first select the texts which may influence the predictions performed (in our example, energy consumption). To do this, we have defined a set of keywords to search: war, a global/world stock market/exchange (Stock Exchange), Wall Street, bankrupt (bankruptcy), financial crisis, recession, political decisions (legislation, payment suspension, etc .), natural disaster (tropical storm, hurricane, typhoon, etc .), and oil/petroleum price. From all these factors, if we focus exclusively on the topic of economic crises, the following list of search terms has been selected: stock exchange recession, stock exchange collapse, stock market recession, stock market collapse, Wall Street recession, Wall Street collapse, bankrupt, bankruptcy, financial crisis, financial shock, economic crisis, economic shock, banking crisis, banking shock, subprime crisis and suppressed shock.

In our case scenario, we have experimented with the Twitter API in order to extract information on the topic of economic crises. To do this, tweets from 1990-2012 were extracted with the aim of detecting the global economic crisis that began in August 2007 and analyzing its potential impact on energy consumption. The abovementioned list of terms relating to economic crises was used; in the search, hashtags (#) and Twitter accounts names (@) could also be included. The selected language is English and the chosen location is the US. Given these conditions, about four million (6 GB) tweets per year and state have been extracted. To process these tweets, only the plain text field containing the tweet as it was written has been taken into account.

Once the tweets corresponding to a period of time have been extracted, we need to incorporate this external information into our system. It is important to emphasize that our main objective is to show how this external information, extracted from peoples comments about the economic crisis, will modify the initial forecast made for energy consumption.

In our approach, the information from tweets is incorporated into the system using Sentiment Analysis and Opinion Mining tools on the texts to be able, for example, to extract the sentiment and opinion about a certain topic (e.g. economic crises, natural disasters, political decisions, etc.) in tweets. These systems determine whether a message (or a fragment of it) expresses a positive, negative, or neutral sentiment.

We have used two systems: (1) OpeNER, Open Polarity Enhanced Name Entity Recognition [47] [2], is a project funded by the European Commission under the 7th Framework Program (FP7-ICT-2011-SME-DCL-296451. 2012-2014). OpeNER's main goal is to provide a set of ready to use tools to perform some natural language processing tasks (tokenizer, POS tagger, polarity tagger, opinion detector, etc.) that are free and easily adapted so that academics, researchers and small and medium-sized enterprises can integrate them into their workflow. (2) GPLSI system: supervised sentiment analysis in Twitter [28] submitted for the SemEval 2014 Task 9 (Sentiment Analysis in Twitter). This consists of a supervised approach using machine learning techniques, without employing any external knowledge and resources.

In Table 1, two examples of tweets related to the economic crisis can be seen. On the left hand side, the output generated is through: (1a) OpeNER using Opinion Detector Basic (rule based) that detects and extracts fine-grained opinions. In particular, the opinion expressed (the actual opinion), the target (what the opinion is about) and the holder (who is expressing the opinion) will be detected for each opinion; (1b) OpeNER using Opinion Opener Deluxe Detector (machine learning based) that extracts fine-grained opinions as the basic version. It is based on Machine Learning, using two Artificial Intelligence algorithms (Conditional Random Fields and Support Vector Machines) to induce models from annotated data. On the right hand side, the output generated shown is through (2) GPLSI: sentiment analysis in Twitter (machine learning based), which uses the terms in the dataset as features. These terms are combined to create skipgrams (not-adjacent ngrams), used as features for a supervised machine learning algorithm.

We can observe in the table how OpeNER identifies partial opinions in the text ("economic crisis" or "crisis worse worse") or an opinion that matches the entire text ("USA's economic crisis is getting worse and worse"). A polarity and strength or intensity is assigned to each opinion. In the case of GPLSI: Sentiment Analysis, the polarity and intensity of all the text is identified; moreover, it is possible to introduce a topic to obtain the polarity and intensity about this.

Here we are working on the new variable TEC (i.e. Twitter economic crisis for a specific state, year, positive, negative and neutral opinion, and intensity of opinion) to be introduced into the DM module as equation 3 (in which n_tw is the number of tweets that satisfy the argument conditions, and W_Pos, W_Neg and W_Neutr is the set of weights that will be trained conveniently). This formula will use the polarity ($PosOpin$, positive; $NegOpin$, negative; or $NeutrOpin$, neutral opinion) and intensity (Int, numerical value) of each tweet.

Table 1: Examples of tweets processed with OpeNER Opinion Detector and GPLSI Sentiment Analysis.

Tweet: USA’s economic crisis is getting worse and worse	

OpeNER: Opinion Detector Basic
<opinions><opinion oid=”o1”>
<opinion_target><!–crisis–>
</opinion_target>
<expr polarity=”positive” strength=”1”>
<!–economic–></expr>
</opinion>
<opinion oid=”o2”>
<opinion_target><!–crisis–>
</opinion_target>
<expr polarity=”negative” strength=”-2”>
<!–worse worse–></expr>
</opinion></opinions>
	GPLSI: Sentiment Anal.
	{
	”subject”: ”OVERALL”,
	”intensity”: 0.98102534,
	”emotionLabels”: null,
	”sentimentCat”: ”negative”
	},
	{”subject”: ”economic crisis”,
	”intensity”: 1,
	”emotionLabels”: null,
	”sentimentCat”: ”neutral”
	}

OpeNER: Opinion Detector Deluxe <opinions><opinion oid=”o1”> <expr polarity=”negative” strength=”1”> <!–USA ’s economic crisis is getting worse and worse–></expr> </opinion></opinions>	
Tweet: The economic crisis was caused by bankers and the 1% - we shouldn’t allow the rhetoric to blame migrants #leadersdebate	

OpeNER: Opinion Detector Basic
<opinions><opinion oid=”o1”>
<opinion_target><!–crisis–>
</opinion_target>
<expr polarity=”positive” strength=”1”>
<!–economic–></expr>
</opinion>
<opinion oid=”o2”>
<opinion_target><!–migrant–>
</opinion_target>
<expr polarity=”negative” strength=”-1”>
<!–blame–></expr>
</opinion></opinions>
	GPLSI: Sentiment Anal.
	{
	”subject”: ”OVERALL”,
	”intensity”: 0.9168395,
	”emotionLabels”: null,
	”sentimentCat”: ”negative”
	},
	{”subject”: ”economic crisis”,
	”intensity”: 1,
	”emotionLabels”: null,
	”sentimentCat”: ”neutral”
	}

OpeNER: Opinion Detector Deluxe <opinions><opinion oid=”o1”> <expr polarity=”negative” strength=”1”> <!–we shouldn ’t allow the rhetoric to blame migrants></expr></opinion></opinions>	

$$\begin{gathered} TEC(S,Y) = n\_tw(S,Y,PosOpin,Int) \cdot W\_Pos + n\_tw(S,Y,NegOpin,Int) \cdot \\ \cdot W\_Neg + n\_tw(S,Y,NeutrOpin,Int) \cdot W\_Neutr \end{gathered} \tag{3}$$

5. Analysis of the results in the case scenario experimentation

In order to conduct a thorough analysis of the impact of combining unstructured and structured data, we have carried out diverse experiments using different DM techniques to perform a comparison between these different approaches and the one we have taken. Specifically, we have used Decision Trees (DT, and in particular the C4.5 algorithm), Artificial Neural Networks (Multi-Layer Perceptron, MLP, was selected in the evaluation), Support Vector Machines (SVM) and Linear Regression (LR). The data used for the experiment are described in sections 4.3 and 4.4 for the years 1990-2012 for all US states and they correspond to structured and semi-structured data respectively. There are too many artificial intelligence methods that can be used in order to predict energy consumption. In our case, these four techniques were chosen in order to carry out a broad set of experiments with diverse methods. They are quite different and all of them offer their own advantages. Furthermore, they have all been used traditionally in predictions problems. ARIMA approaches are linear in their predictions of future values [11] [39]. LR, with some similarities, was therefore chosen. ANN has probably been the most referenced and used method in history. It appears in some highly cited works [44], in others that are very well referenced, such as [78], and in others that include social network data [10]. SVM provides similar accuracy to ANN, while DT provides excellent results as well as great visualizations.

In this experiment, we use cross-validation [45] [30] techniques in order to avoid the classifier overfitting the training data. Usually, this is the best strategy when the experiment is carried out with a large amount of data extracted from real data. In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (testing dataset). The aim of cross-validation is twofold: (1) to define a dataset to test the model in the training phase (i.e., the validation dataset), in order to limit problems like overfitting, and (2) to give an insight into how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem). Specifically, we have used k-cross-validation with a k value of ten.

Table 2: Comparison between different DM techniques with structured data. Accuracy results.

	DT	MLP	SVM	LR
5	95.83%	95.91%	95.39%	95.39%
10	91.04%	89.74%	89.91%	88.70%
15	85.65%	80.43%	76.26%	74.61%
20	82.43%	76.43%	76.43%	66.70%
25	74.09%	69.22%	64.96%	66.17%

Table 2 shows the results of applying various DM techniques using exclusively structured data on energy consumption for all US states. For each technique (DT, MLP, SVM, and LR) the Accuracy obtained (calculated by Correctly Classified instances divided by Total Instances) is shown. The tests were conducted by gradually increasing the number of ranges of discretization (5, 10, 15, 20 and 25) to observe the evolution of the Accuracy.

As can be seen, all of the methods lose accuracy when the number of ranges is increased, since more discretization produces more classes and therefore more classification errors. Regression methods, such as LR, produce good results with small ranges but when the number of ranges increases the accuracy reduces sharply (unlike other methods); for this reason, currently, although they have not been entirely replaced, they are complemented by ANNs, SVM and others ([32]). DT, MLP, and SVM maintain better results as discretization is increased (unlike LR, which is a regression method). It is noteworthy that DT, in addition to offering very good results, provides a visual advantage (compared to other methods) when explaining the energy prediction and this is very useful for users of this kind of application.

The following experiments were conducted using the semi-structured data on meteorological phenomena in US, automatically classified by type of phenomena and degree of intensity. All the weather information obtained has been conveniently combined to obtain the numerical "disasters" variable which was included in the DM module (the data used and the equation formulated for calculating the "disasters" variable are explained in detail in Section 4.4). Table 3 shows the different DM techniques previously applied to the structured data but now with the inclusion of the "disasters" variable.

Table 3: Comparison between DM techniques with structured + semi-structured data (inclusion of "disasters" variable). Accuracy results.

	DT	MLP	SVM	LR
5	95.91%	96.09%	95.57%	95.48%
10	90.87%	90.35%	90.00%	88.96%
15	85.22%	81.04%	80.43%	78.96%
20	82.52%	77.74%	78.70%	71.04%
25	74.35%	69.91%	69.65%	65.65%

As can be seen from the table, almost all of the accuracy results are improved after the inclusion of the information extracted from semi-structured data, especially LR. Although the improvement in DT is not very significant, this does not imply a disadvantage since this approach offers the best results, as well as the benefit of the easy and efficient visualization of the data. For all the abovementioned reasons, the combination of the classification methods suggested (DT, MLP, and SVM) and regression methods offers very interesting prospects for hybrid systems ([41] [40]) to obtain more accurate results.

The comparison between DM techniques using only structured data and those using structured and semi-structured data is shown in figures 7 and 8. In these figures, two examples of the accuracies obtained are represented graphically. Specifically, we have selected the experiments with 15 and 20 ranges. We can see the improvements achieved with the inclusion of the "disasters" variable.

In Table 4 we can see the percentage improvement in accuracy achieved

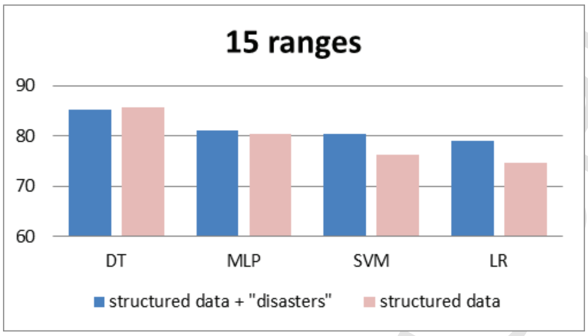


Figure 7: Comparison between DM techniques with structured and semi-structured data. Accuracy results with 15 ranges.

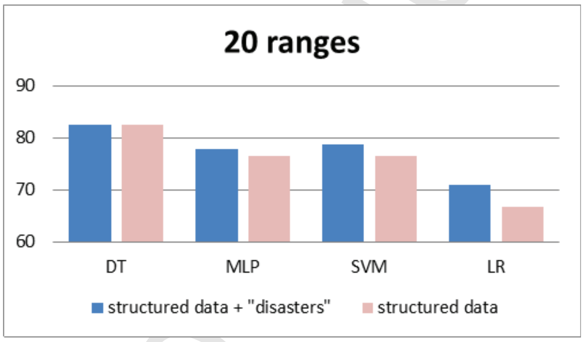


Figure 8: Comparison between DM techniques with structured and semi-structured data. Accuracy results with 20 ranges.

through the inclusion of the “disasters” variable.

Table 4: Percentage improvement in accuracy obtained in DM techniques with the inclusion of “disasters” variable.

	DT	MLP	SVM	LR
5	+0.08%	+0.19%	+0.19%	+0.09%
10	-0.19%	+0.68%	+0.10%	+0.29%
15	-0.50%	+0.75%	+5.18%	+5.51%
20	+0.11%	+1.69%	+2.88%	+6.11%
25	+0.35%	+0.99%	+6.73%	-0.79%

These results are very significant because considerable improvement percentages have been achieved by simply applying a single variable obtained from semi-structured data. The application of the most advanced Natural Language Processing (NLP) techniques will allow for the inclusion of new variables. The objective is for the results obtained with DM techniques exclusively to be significantly improved. As we saw in Section 4.5, we are currently working on the inclusion of the TEC variable, which has been extracted from Twitter on the topic of economic crises. In the future, we propose to include new variables from other information sources and on different topics (i.e. Stock Exchange, diplomatic conflicts, etc.).

In addition, a comparison of the error criteria between the approximations tested is shown. In order to calculate the forecasting errors, we have selected the following three standard criteria:

Kappa statistic (k). Cohen's kappa measures the agreement between two raters who each classify N items into C mutually exclusive categories. The equation for k is:

$$\kappa = 1 - \frac{1 - p_o}{1 - p_e} \tag{4}$$

where p_o is the relative agreement observed among raters, and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly choosing each category.

Mean Absolute Error (MAE) is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. The mean absolute error is given by:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |f_i - y_i| \tag{5}$$

where f_i is the prediction and y_i the true value.

Root Mean Square Error (RMSE). The square root of the mean/average of the square of all the error. The use of RMSE is very common and it offers an excellent all-purpose error metric for numerical predictions. Compared to MAE, RMSE amplifies and severely punishes large errors.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (x_{1,t} - x_{2,t})^2}{n}} \tag{6}$$

where $x_{1,t}$ and $x_{2,t}$ are measuring the average difference between two time series.

Further information regarding all these measures can be found at [58] [15] [26] [73] [74] [16].

In tables 5 and 6 we can see that the coefficient k gradually decreases (which is logical as there are more classifications options for a particular instance) while MAE and RMSE (represented by AE and SE respectively in the tables) are generally increasing slightly with increasing discretization ranges. No significant differences were observed between the results with structured data and those that include semi-structured data.

In figures 9 and 10 we can see the ROC curves of some of the experiments with the different machine learning methods as well as with different numbers of ranges.

Table 5: Comparison of the error criteria between the 4 models tested using structured data (without "disasters" variable).

	DT			MLP			SVM			LR		
	K	AE	SE	K	AE	SE	K	AE	SE	K	AE	SE
5	0.95	0.02	0.12	0.95	0.02	0.12	0.94	0.24	0.32	0.94	0.03	0.12
10	0.90	0.02	0.13	0.89	0.03	0.13	0.89	0.16	0.27	0.87	0.04	0.14
15	0.85	0.02	0.13	0.79	0.03	0.15	0.75	0.12	0.24	0.73	0.05	0.16
20	0.82	0.02	0.12	0.75	0.03	0.14	0.75	0.09	0.21	0.65	0.05	0.15
25	0.73	0.02	0.13	0.68	0.03	0.14	0.64	0.07	0.19	0.65	0.04	0.13

Table 6: Comparison of the error criteria between the 4 models tested using structured and semi-structured data (with "disasters" variable).

	DT			MLP			SVM			LR		
	K	AE	SE	K	AE	SE	K	AE	SE	K	AE	SE
5	0.95	0.02	0.12	0.95	0.02	0.11	0.95	0.24	0.32	0.94	0.03	0.12
10	0.90	0.02	0.13	0.89	0.02	0.13	0.89	0.16	0.27	0.88	0.04	0.13
15	0.84	0.02	0.13	0.80	0.03	0.14	0.79	0.12	0.24	0.78	0.04	0.14
20	0.82	0.02	0.13	0.77	0.03	0.14	0.78	0.09	0.21	0.70	0.04	0.14
25	0.73	0.02	0.13	0.69	0.03	0.14	0.68	0.07	0.19	0.64	0.04	0.14

The area under the ROC curve has been used in many works to measure and compare several algorithms. In particular, we can find more detailed and further information on the evaluation of machine learning algorithms [13].

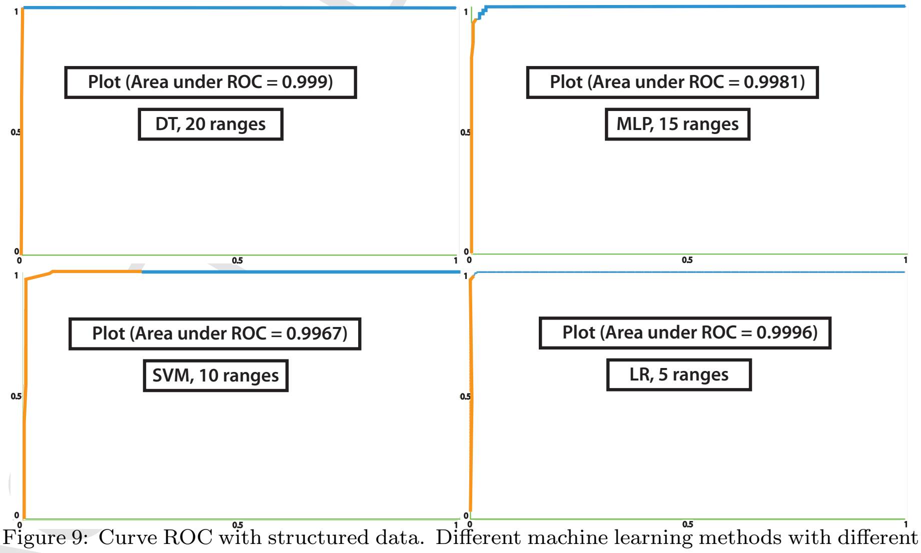


Figure 9: Curve ROC with structured data. Different machine learning methods with different numbers of ranges.

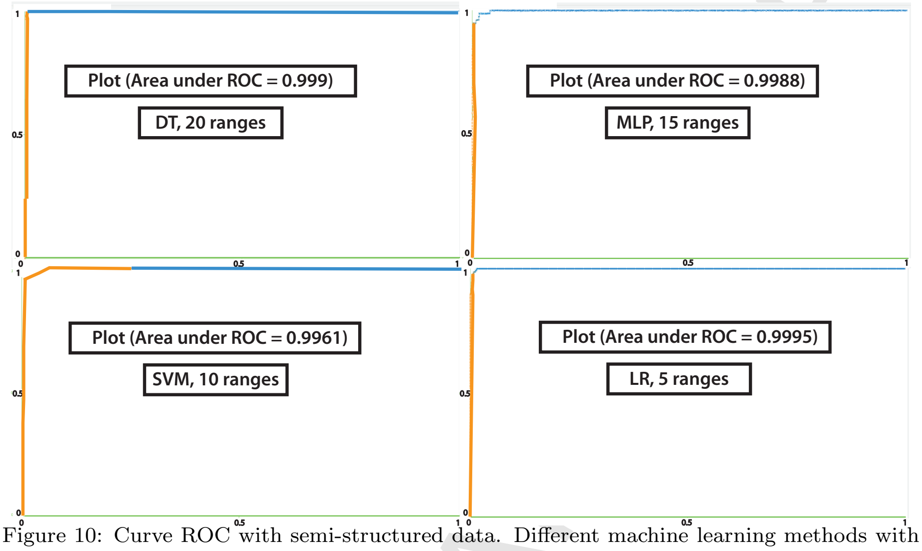


Figure 10: Curve ROC with semi-structured data. Different machine learning methods with different numbers of ranges.

6. Conclusions and future work

In this paper we have presented an integrated approach to allow data mining predictions with heterogeneous Big Data. In the original model, the objective is to carry out consumption prediction by using the input data of US Energy Information Administration (EIA). Our main goal is to enhance the predictive model by enriching it with the new variables obtained from an Information Extraction system and Social Networks.

We have created an extended multidimensional model that allows us to integrate on the fly information extracted from several specialized nodes, including Information Extraction and opinions from Social Networks. It is important to emphasize the modularity, flexibility, dynamism, and scalability of our proposal. Several DW, Database, Big Data, Information Extraction or Question Answering (QA[8]) source nodes can be added and connected to the system, each one with its own implementation, model and domain (e.g. we can connect an IE node specialized in political domains as well as a QA node specialized in financial domains). This allows us to enrich the original DW data in order to obtain a dynamic data mining model.

We have presented a real case scenario in which energy consumption predictions have been modified with the inclusion of external variables obtained from semi-structured data of natural disasters. In the near future, and as part of an

[8] Question Answering systems represent the potential future of Web search engines because QA returns specific answers as well as documents. It supposes the combination of Information Retrieval, which obtains information resources relevant to an information need from a collection of information resources, and IE techniques.

ongoing project, the information of the opinions concerning a topic in a specific period of time will be introduced as a new variable derived from the polarity and intensity of the opinions conveniently weighted.

The experiments carried out in this work are twofold: (i) using and comparing diverse AI methods, and (ii) validating our approach with data sources integration.

In order to improve the accuracy, when the number of ranges increases we will propose an improvement in the architecture, that is, data mining that consists of a hybrid system with two stages: (i) classification to determine a range and (ii) regression to refine it. In this way, the prediction will be more meaningful once the new hybrid approach has been established.

It is important to mention that in this work we only included a new variable obtained from the semi-structured data. In the future, we propose the inclusion of new variables from other information sources as well as different topics (i.e. Stock Exchange, diplomatic conflicts, etc.) as it has been incorporated in work by [51].

The main future work is to perform a thorough evaluation of the implementation of the model. Some of the difficulties are related to the nature of the data, as it is composed by very heterogeneous information. Once the integration is complete we will compare how the each enriched model performs compared to the others in order to obtain further insights about what information is more valuable for making decisions.

Acknowledgements

This work has been funded by the Spanish Ministry of Economy and Competitiveness under the project Grant TIN2015-63502-C3-3-R. Alejandro Maté is funded by the Generalitat Valenciana (APOSTD/2014/064).

Appendix

See Table 7 and Figure 11.

Table 7: Technology selection in our architecture.

Technology	Provider	Reason	Benefits
SPARQL	W3C	Standard for querying semantic resources in RDF	Interoperability with the semantic information widely available in the web
Pentaho BI	Pentaho	Open-source platform for MDX querying and data warehousing	Easy to adapt where necessary thanks to being open-source
MySQL	Oracle	Widely used SQL database, required for the DW	Easy to install, tested compatibility with Pentaho
Neo4j (NoSQL)	Neo Technology	NoSQL database for external information and entity storage used in information retrieval and social networks	Graph-based database with native support for RDF and SPARQL
Tree-tagger	Institute for Computational Linguistics	A language independent POS-tagger for Information Extraction. It is a tool for annotating text with part-of-speech and lemma information	Segments the text into words, and obtains part-of-speech and lemma information of each word.
SUPAR	in-house	Partial parser for running partial and full parsing. Solves some linguistic problems such as anaphora, ellipsis, clause and sentence segmentation, name entity tagging and classification	Segments the text into phrases, clauses and sentences, and obtains syntactic relations between words in each phrase (e.g. noun, verbal or prepositional phrases)
Freeling	TALP Research Center	Recognises dates, numbers, ratios, currency, physical magnitudes (speed, weight, temperature, density, etc.), enterprise, geographical names, etc.	Open source language analysis tool suite that processes different languages. It is portable between different platforms
WordNet	Princeton	Lexical database of English containing nouns, verbs, adjectives and adverbs, grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept and interlinked	Widely tested, used by numerous tools for different Natural Language Processing applications in many languages
OpeNER	OpeNER European Project	Named entity recognition for information extraction and social networks	Easy to adapt and integrate with other technologies
GPLSI	in-house	Sentiment analysis processing	Supervised machine learning without requiring external knowledge and resources

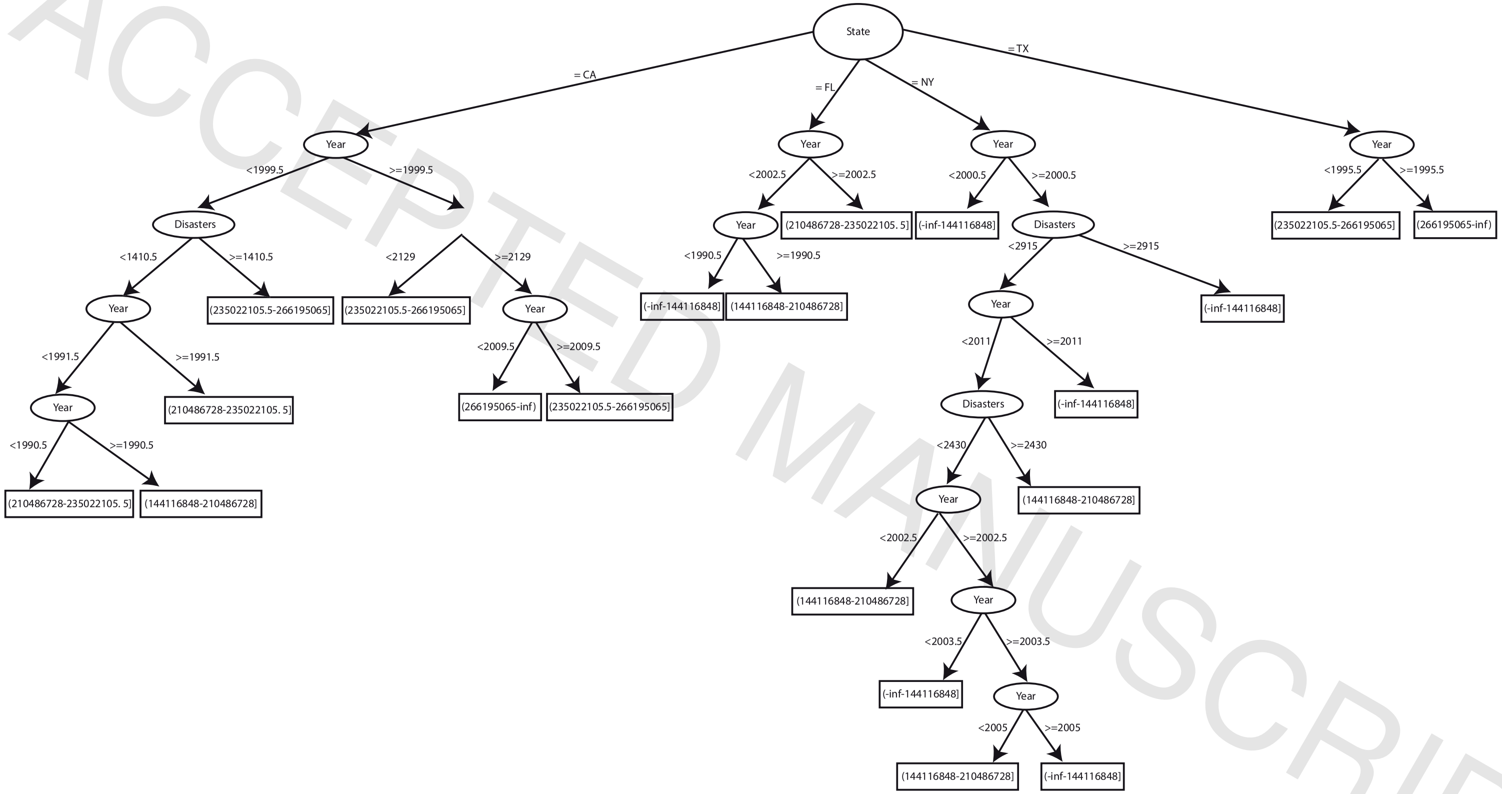


Figure 11: RandomForest Tree output with semi-structured data (presenting the “disasters” variable).

References

[1] E. Abdelaziz, R. Saidur, and S. Mekhilef. A review on energy saving strategies in industrial sector. *Renewable and Sustainable Energy Reviews*, 15(1): 150–168, 2011.

[2] R. Agerri, M. Cuadros, S. Gaines, and G. Rigau. Opener: Open polarity enhanced named entity recognition. *Procesamiento del Lenguaje Natural*, 51:215–218, 2013.

[3] D. Akay and M. Atak. Grey prediction with rolling mechanism for electricity demand forecasting of Turkey. *Energy*, 32(9):1670–1675, 2007.

[4] H. K. Alfares and M. Nazeeruddin. Electric load forecasting: literature survey and classification of methods. *International Journal of Systems Science*, 33(1):23–34, 2002.

[5] A. Azadeh, S. Ghaderi, S. Tarverdian, and M. Saberi. Integration of artificial neural networks and genetic algorithm to predict electrical energy consumption. *Applied Mathematics and Computation*, 186(2):1731–1741, 2007.

[6] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

[7] S. Banfi, M. Farsi, M. Filippini, and M. Jakob. Willingness to pay for energy-saving measures in residential buildings. *Energy economics*, 30(2): 503–516, 2008.

[8] F. Bauer, D. Recheis, and M. Kaltenböck. A new key portal for open energy data. Environmental software systems. *Environmental Software Systems. Frameworks of eEnvironment*, 359:189–194, 2011.

[9] F. Benzi, N. Anglani, E. Bassi, and L. Frosini. Electricity smart meters interfacing the households. *Industrial Electronics, IEEE Transactions on*, 58(10):4487–4494, 2011.

[10] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[11] T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327, 1986.

[12] D. Bourgeois, C. Reinhart, and I. Macdonald. Adding advanced behavioural models in whole building energy simulation: a study on the total energy impact of manual and automated lighting control. *Energy and Buildings*, 38(7):814–823, 2006.

[13] A. P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[14] D. Brodt-Giles. OpenEI an open energy data and information exchange for international audiences. In *2012 World Renewable Energy Forum. International Solar Energy Society*, 2012.

[15] J. Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254, 1996.

[16] V. Chang. The business intelligence as a service in the cloud. *Future Generation Computer Systems*, 37:512–534, 2014.

[17] V. Chang. Towards a big data system disaster recovery in a private cloud. *Ad Hoc Networks*, 35:65–82, 2015.

[18] V. Chang and G. Wills. A model to compare cloud and non-cloud storage of big data. *Future Generation Computer Systems*, 57:56–76, 2016.

[19] M. Chen, S. Mao, and Y. Liu. Big data: A survey. *Mobile Networks and Applications*, 19(2):171–209, 2014.

[20] M. Cotterell, J. Zheng, Q. Sun, Z. Wu, C. Champlin, and A. Beach. Facilitating knowledge sharing and analysis in energy informatics with the ontology for energy investigations (OEI). *Sprouts: Working Papers on Information Systems*, 12(4):1–18, 2012.

[21] A. Daouadji, K. K. Nguyen, M. Lemay, and M. Cheriet. Ontology-based resource description and discovery framework for low carbon grid networks. In *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, pages 477–482. IEEE, 2010.

[22] A. T. de Almeida, P. Fonseca, and P. Bertoldi. Energy-efficient motor systems in the industrial and in the services sectors in the European Union: characterisation, potentials, barriers and policies. *Energy*, 28(7):673–690, 2003.

[23] J. Dean and S. Ghemawat. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[24] Y. Demchenko, P. Grosso, C. De Laat, and P. Membrey. Addressing big data issues in scientific data infrastructure. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on*, pages 48–55. IEEE, 2013.

[25] C. Dobre and F. Xhafa. Intelligent services for big data science. *Future Generation Computer Systems*, 37:267–281, 2014.

[26] B. Efron and R. Tibshirani. *Statistical data analysis in the computer age.* University of Toronto, Department of Statistics, 1990.

[27] L. Ekonomou. Greek long-term energy consumption prediction using artificial neural networks. *Energy*, 35(2):512–517, 2014.

[28] J. Fernández, Y. Gutiérrez, J. M. Gómez, and P. Martínez-Barco. GPLSI: Supervised sentiment analysis in twitter using skipgrams. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 294–299, 2014.

[29] Global e-Sustainability Initiative et al. Smart 2020 report: Global ICT solution case studies. *The Climate Group, Tech. Rep*, 2008.

[30] G. H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

[31] E. González-Romera, M. A. Jaramillo-Morán, and D. Carmona-Fernández. Monthly electric energy demand forecasting based on trend extraction. *Power Systems, IEEE Transactions on*, 21(4):1946–1953, 2006.

[32] E. Hadavandi, H. Shavandi, and A. Ghanbari. Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting. *Knowledge-Based Systems*, 23(8):800–808, 2010.

[33] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[34] C. Hamzaebi. Forecasting of Turkey's net electricity energy consumption on sectoral bases. *Energy Policy*, 35(3):2009–2016, 2007.

[35] J. Han, E. Haihong, G. Le, and J. Du. Survey on NoSQL database. In *Pervasive computing and applications (ICPCA), 2011 6th international conference on*, pages 363–366. IEEE, 2011.

[36] J. Han, Y. K. Jeong, and I. Lee. Efficient building energy management system based on ontology, inference rules, and simulation. In *Proceedings of the 2011 International Conference on Intelligent Building and Management, Singapore*, volume 5, pages 295–299, 2011.

[37] L. Hilty, W. Lohmann, and E. Huang. Sustainability and ICT - An overview of the field. *POLITEIA*, 27(104):13–28, 2011.

[38] T. Hong. Energy forecasting: Past, present and future. *The International Journal of Applied Forecasting*, 32:43–48, 2014.

[39] D. A. Hsieh. Chaos and nonlinear dynamics: application to financial markets. *The Journal of Finance*, 46(5):1839–1877, 1991.

[40] C.-L. Huang and C.-Y. Tsai. A hybrid sofm-svr with a filter-based feature selection for stock market forecasting. *Expert Systems with Applications*, 36(2):1529–1539, 2009.

[41] S.-C. Huang and T.-K. Wu. Integrating ga-based time-scale feature extractions with svms for stock index forecasting. *Expert Systems with Applications*, 35(4):2080–2088, 2008.

[42] J. Keirstead and K. H. Van Dam. A comparison of two ontologies for agent-based modelling of energy systems. In *First International Workshop on Agent Technologies for Energy Systems*, pages 21–28, 2010.

[43] J. Keirstead, N. Samsatli, and N. Shah. SynCity: an integrated tool kit for urban energy systems modelling. *Energy Efficient Cities: Assessment Tools and Benchmarking Practices, World Bank*, pages 21–42, 2010.

[44] T. Kimoto, K. Asakawa, M. Yoda, and M. Takeoka. Stock market prediction system with modular neural networks. In *Neural Networks, 1990., 1990 IJCNN International Joint Conference on*, pages 1–6. IEEE, 1990.

[45] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, pages 1137–1145, 1995.

[46] M. Krötzsch, D. Vrandečić, and M. Völkel. Semantic mediawiki. In *2006 International Semantic Web Coference*, pages 935–942, 2006.

[47] E. Maks, R. Izquierdo, and P. Vossen. OpeNER and the automatic generation of sentiment lexicons in five languages. In *Proceedings of CLIN-2014*, 2014.

[48] B. Marr. *Big Data: using SMART big data, analytics and metrics to make better decisions and improve performance.* John Wiley & Sons, 2015.

[49] A. Maté, H. Llorens, and E. de Gregorio. An integrated multidimensional modeling approach to access big data in business intelligence platforms. In *Advances in Conceptual Modeling*, pages 111–120. Springer, 2012.

[50] W. J. Mitchell. *E-topia: "Urban life, Jim–But not as we know it".* MIT press, 2000.

[51] J. Moeyersoms and D. Martens. Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector. *Decision Support Systems*, 72:72–81, 2015.

[52] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee. A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3):45–77, 2007.

[53] J. Peral, A. Ferrández, E. De Gregorio, J. Trujillo, A. Maté, and L. J. Ferrández. Enrichment of the phenotypic and genotypic data warehouse analysis using question answering systems to facilitate the decision making process in cereal breeding programs. *Ecological Informatics*, 26:203–216, 2015.

[54] J. Pérez, M. Arenas, and C. Gutierrez. Semantics and complexity of SPARQL. *ACM Transactions on Database Systems (TODS)*, 34(3):16, 2009.

[55] J. Quinlan. *C4. 5: programs for machine learning*. Morgan Kaufmann, 1993.

[56] A. Reinders, K. Vringer, and K. Blok. The direct and indirect energy requirement of households in the European Union. *Energy Policy*, 31(2): 139–153, 2003.

[57] W. Shi. Renewable energy: Finding solutions for a greener tomorrow. *Reviews in Environmental Science and Biotechnology*, 9:35–37, 2010.

[58] J. Sim and C. C. Wright. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, 85(3):257–268, 2005.

[59] G. J. Smit. Efficient ICT for efficient smart grids. In *IEEE PES Innovative Smart Grid Technologies, ISGT 2012*. IEEE Power & Energy Society, 2012.

[60] D. Srinivasan. Energy demand prediction using GMDH networks. *Neurocomputing*, 72(1):625–629, 2008.

[61] L. Suganthi and A. A. Samuel. Energy models for demand forecasting—A review. *Renewable and Sustainable Energy Reviews*, 16(2):1223–1240, 2012.

[62] L. G. Swan and V. I. Ugursal. Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and sustainable energy reviews*, 13(8):1819–1835, 2009.

[63] R. Tardío, E. De Gregorio, A. Maté, R. Muñoz-Terol, D. Gil, H. Llorens, and J. Trujillo. Modelado multidimensional para la visualización integrada de big data en plataformas de inteligencia de negocio. In *XIX Jornadas de Ingeniería del Software y Bases de Datos (JISBD)*, pages 33–38, 2014.

[64] A. Ünler. Improvement of energy demand forecasts using swarm intelligence: The case of Turkey with projections to 2025. *Energy Policy*, 36(6): 1937–1944, 2008.

[65] V. Utgikar and J. Scott. Energy forecasting: Predictions, reality and analysis of causes of error. *Energy Policy*, 34(17):3087–3092, 2006.

[66] M. C. van der Sanden and K. H. van Dam. Towards an ontology of consumer acceptance in socio-technical energy systems. In *Infrastructure Systems and Services: Next Generation Infrastructure Systems for Eco-Cities (INFRA), 2010 Third International Conference on*, pages 1–6. IEEE, 2010.

[67] E. Vine. An international survey of the energy service company (ESCO) industry. *Energy Policy*, 33(5):691–704, 2005.

[68] J. Virote and R. Neves-Silva. Stochastic models for building energy prediction based on occupant behavior assessment. *Energy and Buildings*, 53: 183–193, 2012.

[69] R. H. von Alan, S. T. March, J. Park, and S. Ram. Design science in information systems research. *MIS quarterly*, 28(1):75–105, 2004.

[70] S. V. Vrbsky, M. Galloway, R. Carr, R. Nori, and D. Grubic. Decreasing power consumption with energy efficient data aware strategies. *Future Generation Computer Systems*, 29(5):1152–1163, 2013.

[71] R. T. Watson and M. C. Boudreau. Energy Informatics. *Green ePress*, 2011.

[72] M. Webb et al. Smart 2020: Enabling the low carbon economy in the information age. *The Climate Group. London*, 1(1):1–1, 2008.

[73] C. J. Willmott and K. Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79, 2005.

[74] C. J. Willmott, S. G. Ackleson, R. E. Davis, J. J. Feddema, K. M. Klink, D. R. Legates, J. O'donnell, and C. M. Rowe. Statistics for the evaluation and comparison of models. *Journal of Geophysical Research: Oceans (1978–2012)*, 90(C5):8995–9005, 1985.

[75] C.-M. Wu, R.-S. Chang, and H.-Y. Chan. A green energy-efficient scheduling algorithm using the dvfs technique for cloud datacenters. *Future Generation Computer Systems*, 37:141–147, 2014.

[76] F. Xhafa and L. Barolli. Semantics, intelligent processing and services for big data. *Future Generation Computer Systems*, 37:201–202, 2014.

[77] K. Young, T. Reber, and K. Witherbee. Hydrothermal exploration best practices and geothermal knowledge exchange on OpenEI. In *Proceedings of the 37th Workshop on Geothermal Reservoir Engineering*, pages 1455–1469, 2012.

[78] Y. Zhang and L. Wu. Stock market prediction of s&p 500 via combination of improved bco approach and bp neural network. *Expert systems with applications*, 36(5):8849–8854, 2009.

[79] Y. Zhao and Y. Zhang. Comparison of decision tree methods for finding active objects. *Advances in Space Research*, 41(12):1955–1959, 2008.

[80] P. Zikopoulos, C. Eaton, et al. *Understanding big data: Analytics for enterprise class Hadoop and streaming data.* McGraw-Hill Osborne Media, 2011.

*Biographies (Text)

Alejandro Maté is a postdoc researcher at the University of Trento, Italy. He received a BS and MSc Degree in Computer Science from the University of Alicante. He earned his PhD degree in Computer Science from the University of Alicante in 2013. He has published several papers in international conferences and journals such as CAiSE, ER, RE, JSS, and IS. His research involves conceptual modeling, data warehouses, model driven development, and requirements engineering.

Jesús Peral is an assistant professor at the Department of Software and Computing Systems in the University of Alicante. He obtained his PhD in Computer Science from the University of Alicante (2001). His main research topics include: Natural Language Processing, Information Extraction, Information Retrieval, Question Answering, data warehouses and Business Intelligence applications. He has participated in numerous national and international projects, agreements with private companies and public organizations related to his research topics. He has published many papers (more than 40 papers) in Journals and Conferences related to his research interests.

Antonio Ferrández is an assistant professor at the Department of Software and Computing Systems in the University of Alicante. He obtained his PhD in Computer Science from the University of Alicante (1998). His research topics include Information Extraction, Information Retrieval, Question Answering, Natural Language Processing, Ellipsis and Anaphora Resolution. He has participated in numerous national and international projects, agreements with private companies and public organizations related to his research topics. He has participated in many conferences and most of his work has been published in international journals and conferences, with more than 70 published papers.

David Gil is an assistant professor at the Department of Computing Technology and Data Processing in the University of Alicante. His main research topics include artificial intelligence applications, data mining, open data, big data, decision support system in medical and cognitive sciences. He has participated in numerous national and international projects, agreements with private companies and public organizations related to his research topics. He has participated in many conferences and most of his work has been published in international journals and conferences, with more than 50 published papers.

Juan Trujillo is a Full-time Professor at the Department of Software and Computing Systems in the University of Alicante and the leader of the Lucentia Research Group. His main research topics include Business Intelligence applications, Business Intelligence 3.0, Big Data, data warehouses' development, OLAP, data mining, UML, MDA, data warehouses' security and quality. He has advised 11 PhD students and published more than 200 papers in different highly impact conferences. Currently, he is Senior Editor of the "Decision Support Systems" journal (Q1). He has participated in many national and international projects and he has been the lead investigator on several research projects and technology transfer related to Business Intelligence issues. Finally, he has the international credential Project Management Professional (PMP®) for project management awarded by the prestigious Project Management Institute (PMI).

*Biographies (Photograph)

Click here to download high resolution image

*Biographies (Photograph)

*Biographies (Photograph)

Click here to download high resolution image

*Biographies (Photograph)

*Biographies (Photograph)