Combining Statistical Data for Machine Learning Analysis

Evangelos Kalampokis, Areti Karamanou, and Konstantinos Tarabanis

University of Macedonia, Thessaloniki, Greece
{ekal,akarm,kat}@uom.edu.gr

Abstract. Machine learning represents a pragmatic breakthrough in making predictions by finding complex structures and patterns in large volumes of data. Open Statistical Data (OSD), which are highly structured and generally of high quality, can be used in advanced decision making scenarios that involve machine learning analysis. Linked data technologies facilitate the discovery, retrieval, and combination of data on the Web. They enable this way the wide exploitation of OSD in machine learning. A challenge in such analyses is to specify the criteria for selecting the proper datasets to combine and construct a predictive model. This paper presents a case study that aims at creating a model to predict house sales prices in fine grained geographical areas in Scotland using a large variety of Linked Open Statistical Data (LOSD) from the Scottish official statistics portal. To this end, we present the machine learning analysis steps that can be enhanced using LOSD and we define a set of compatibility criteria. A software tool is also presented as a proof of concept for facilitating the exploitation of LOSD in machine learning. The case study proves the importance of discovering and combining compatible datasets when implementing machine learning scenarios for decision-making.

Keywords: statistical data · machine learning · compatibility.

1 Introduction

Opening up data for others to reuse is a priority in many countries around the globe. Although the global annual economic potential of open data is estimated to $3 trillion [14], this potential has been unrealized to a large extent. This is explained by a number of barriers that hamper the implementation of sophisticated solutions [20] at the institutional level (e.g. the task complexity of handling data, legislation, information quality) and technical level [8] .

A promising path to overcome open data barriers is to focus on numerical data and, more specifically, statistics [11]. Open Statistical Data (OSD) constitute a large part of open data [6]. Their added value is related to the fact that they are highly structured, hence they can be easily processed. Moreover, they describe financial, social, and political aspects of the world, thus playing crucial role for being a major element in economic and social decision-making [7].

However, OSD are barely used in advanced decision-making scenarios that involve machine learning analysis. Machine learning represents a pragmatic breakthrough in making predictions by finding complex structures and patterns in large volumes of data. Recent examples indicating the potential of applying machine learning in statistical data to support decision making include the identification of important factors related to bicycle crashes [15], analysis of consumption patterns [5], prediction of crime through both demographic and mobile data [1], definition of consumer profile using internal company and statistical data [2].

This difficulty of using statistical data in advanced machine learning scenarios can be explained, among others, by the fragmented environment of OSD [7]. OSD are usually provided by Web portals as downloadable files (e.g. CSV, JSON) or through specialized APIs. In the first case, data about an indicator are provided through hundreds, even thousands, of different files. For example, searching for "unemployment" in the UKs official open data portal results in more than 2.000 relevant files [13]. In the latter case, existing APIs do not address requirements regarding the combination of data from multiple datasets or sources [19]. As a result combining statistical datasets in order to involve them in advanced machine learning analysis remains a difficult task.

Linked data technologies facilitate discovering, retrieving and combining of data on the Web by semantically annotating data, creating links between them and enabling their access using the query language SPARQL. Linked data have been recently become a W3C standard [18]. Indeed, during the last years many National Statistics Institutes and governments have created Web portals providing Linked Open Statistical Data (LOSD). Examples include the UK's Office for National Statistics[1] and the Scottish Government[2]. Early research in this area contributed towards this direction (e.g. [12,9,10,16]). All LOSD portals use standard Web technologies (e.g. HTTP, RDF, URIs) and vocabularies (e.g. RDF data cube, SKOS, XKOS).

The large volume and variety of datasets provided by LOSD portals are necessary in sophisticated machine learning scenarios in order to create predictive models. A challenge in such scenarios is to specify the criteria that should be considered when selecting datasets to use in order to solve a specific problem.

The aim of this paper is to present a case study that combines LOSD in order to perform machine learning analysis and support advanced decision-making. Towards this end, we first specify the criteria that define which datasets can be used to solve a problem using machine learning. The datasets of our case study are selected based on these criteria. We also present the Compatible LOSD Selection tool, a proof of concept of the case study that facilitates the selection of datasets that will be combined for machine learning analysis.

The rest of the paper is organised as follows: Section 2 presents the method of this paper. Section 3 defines the compatibility criteria. Section 4 presents the case study and its results. Section 5 presents the Compatible LOSD Selection tool. Finally, Section 6 concludes and discusses the results.

[1] `http://statistics.data.gov.uk`
[2] `http://statistics.gov.scot`

2 Method

The method used in the case study includes four steps:

1. *Problem definition.* The problem definition step enables users to define the problem they are interested to solve using machine learning analysis. To this end, the response variable of the predictive model is defined (including geographical boundaries, time constraints, units of measure etc.). This requires exploring the metadata of available datasets. Moreover, the type of the problem is specified (e.g. regression, classification etc.). For example, a problem could be to predict the 2012 house prices in the 2001 data zones of Scotland.
2. *Data selection.* The data selection step selects the datasets that will be combined with the response variable and contribute towards solving the problem defined in the previous step. The selection of the datasets uses five structural criteria based on the granularity of the geographical dimension, the temporal dimension, the unit of the measure, the type of the measure and additional dimensions.
3. *Feature extraction.* This step extracts from the datasets selected in the previous step numerous features aka predictors. Features are extracted from the combination of different dimensions and measures in one or more datasets. Dimensions determine and explain a feature. For example, an unemployment dataset with four dimensions i.e. age group (15-25, 25-54, 55-64), type of unemployment (cyclical, frictional, structural), measure type (count, ratio), reference period (2001-Q1, ..., 2016-Q4) could result in 3 x 3 x 2 x 64 =1152 features.
4. *Feature selection and model creation.* The feature selection step selects among all extracted features the ones that will be used to construct the predictive model. Those are features that are significantly correlated to the response variable. Features considered as redundant or irrelevant are ignored. Machine learning methods to select features include (Least Absolute Shrinkage and Selection Operator) Lasso[17], stepwise selection and tree boosting. For our case study we use the Lasso method to select features. In addition, in order to assess the result of the machine learning method used to select features, criteria such as Mean Squared Error (MSE) which measures the average of the squares of the errors (i.e. the difference between the actual and the predicted value) and the misclassification error are commonly used. In our case study we use Root Mean Squared Error (RMSE) to assess the result of Lasso.

LOSD contribute in the second step of the methodology by facilitating the selection of datasets that can be combined with the response variable in order to construct the predictive model. The next Section specifies the criteria to consider in order to select compatible LOSD that can contribute in a predictive model as a response variable or as a feature.

3 Combining statistical datasets for machine learning analysis

In general, statistical data are aggregated data that describe a measured fact (e.g. house prices) in specific geographical points (e.g. a country, city or building) and in a specific period of time (e.g. a year, month, week). In this case, statistical data are compared to a data cube, where each cell contains a measure or a set of measures, and thus we can refer to statistical data as data cubes or just cubes [4]. The geographical point and the period of time that describe a measure are called dimensions (geographical and temporal respectively). A statistical dataset can be described by additional dimensions as well such as age, gender etc. It is frequently useful to create a subset of a statistical dataset. This subset fixes all but one (or a small subset) of the initial datasets' dimensions and is called a slice through the dataset [3].

The second step of our methodology requires selecting the slices of statistical datasets that will contribute as the response variable (also called Y) and also as the features (also called Xs) of the defined problem based on:

1. The granularity of the geographical dimension.
2. The temporal dimension.
3. The unit of the measure.
4. The type of the measure.
5. Additional dimensions.

We specify the above criteria separately for the response variable and the features. In particular, the selection of the slice that will be used for the response variable is based on:

1. The granularity of the geographical dimension. Commonly the defined problem focuses on geographical points with a specific granularity level (e.g. to predict the house prices in the 2001 data zones of Scotland). As a result the slice selected for the response variable should use this specific granularity level. This will be the open dimension of the slice.
2. The temporal dimension. The defined problem focuses on a specific period of time (e.g. to predict the 2012 house prices in the 2001 data zones of Scotland). As a result the slice selected for the response variable should have the time dimension fixed to the selected period of time.
3. The unit of the measure. Datasets usually use a unit to describe their measure. Common units of measures are ratio and count. Depending on the problem slices using ratio or count should be selected. If the selected dataset includes more than one units of measure the unit dimension should be fixed to the preferred unit of measure.
4. The type of the measure. The measure of a statistical dataset may be categorical or continuous. Continuous measures contain numbers with infinite number of values between any two values. Categorical measures contain a finite number of categories or distinct groups. The nature of the defined problem will specify the type of the measure to be selected for the slice of the response variable.

5. Additional dimensions. Additional dimensions in the selected slice are desirable (but also optional) as they increase the number of extracted features that could be used in the construction of more reliable predictive models. A common additional dimension is, for example, the gender. Additional dimensions should be also fixed to a specific value.

In addition the selection of the slices for the features is based on:

1. The granularity of the geographical dimension. The slices selected for the X variables of the predictive model should have the same granularity level with the slice of Y. As a result, only datasets that have the same granularity level in the geographical dimension with the Y variable should be selected.
2. The temporal dimension. Machine learning usually aims to predict a specific phenomenon based on historical data. As a result slices selected for the X variables should refer to the same or past years related to the Y variable.
3. The unit of the measure. Slices using ratio are preferably selected over count because ratio values are normalized. However, slices with count measures can be also selected provided that they will be combined with other count measures in the next step of the methodology (namely feature extraction) in order to construct new ratio variables. For example, one could select a slice counting the number of births and also a slice counting the number of deaths from the data portal of Scotland in order to create in the feature selection step the ratio 'number of births/number of deaths'.
4. The type of the measure. In a predictive model it is not mandatory for the Y and X variables to have the same type. As a result, when the Y variable is categorical the Data selection step can select slices with either categorical or continuous measures for the features and vice versa.
5. Additional dimensions. The selected slice can also have additional dimensions (e.g. the gender) with same or different values related to the respective dimension of Y.

4 Case Study: Predicting the House Prices in Scotland

The case study presented in this paper uses datasets from the official statistics data portal of Scotland i.e. `http://statistics.gov.scot` that was launched in August 2016. At the time of writing the portal provides access to 220 statistical datasets about Scotland. The datasets can be viewed in variable formats including tables, maps and charts or downloaded formats like CSV or N-triples formats. The datasets can be browsed by theme (e.g. Labour Force, Environment, Transport etc.) or by the organisation that published the dataset (e.g. Scottish government, SEPA or Transport Scotland).

Scottish official statistics are also provided in Linked Data format using the W3C's RDF Data Cube Vocabulary[3] which allows modelling statistical data as data cubes. In particular, each dataset in the portal is modelled as a data

[3] `https://www.w3.org/TR/vocab-data-cube/`

cube. Each data cube provides multiple ancillary dimensions in complement of the indicator which is the measure of the data cube. The two most common dimensions used to describe the datasets are the geographical dimension called *Reference area* and the temporal dimension called *Reference Period*. The geographical dimension of the datasets is based on a hierarchy of administrative or consensus-based areas covering from Scottish data zones to electoral awards and countries. Granularity refers to the levels of depth of the reference area dimension of each dataset. Some examples of these levels include country, council areas, electoral wards, and data zones. For example, the house sales dataset[4] describes the number of Residential property transactions recorded in different geographical levels of Scotland (e.g. Countries, Electoral Wards, 2001 Data zones and others) in different reference periods (1993-2017). Other commonly used dimensions include the gender and age group of the population.

Problem definition The objective of the case study presented in this paper is to predict the 2012 mean house prices in the 2001 data zones of Scotland. This is a description of the response variable of our problem.

2001 Data zones were introduced in 2004 and are the smallest geographical granularity level in Scotland. They have populations between 500 and 1,000 household residents. Selecting 2001 data zones for our response variable results in a great number of observations that help to avoid the curse of high-dimensionality (i.e. the state of having less observations than features), create a more robust model, and predict prices for a specific district or neighbourhood.

Regarding the machine learning method used, regression analysis Lasso method is selected to solve the above described problem. Lasso yields sparse models i.e. models that involve only a subset of the variables.

Data selection After the definition of our problem we first search for datasets in the Scottish data portal that can contribute as the response variable of our problem. The slice selected for the response variable comes from the dataset House prices of the Scottish portal[5] with the temporal dimension fixed to 2012, the measure type fixed to mean, and the values of the reference area dimension coming from the 2001 data zones in Scotland.

We then search for datasets that can contribute as features. The selected datasets should be compatible with the response variable. For this reason we search in the Scottish data portal for datasets based on the compatibility criteria described in the previous section. In particular, we search for datasets that:

1. The granularity level in their geographical dimension is Scottish 2001 data zones
2. Their temporal dimension refers to a year in the range 2009-2012

[4] `http://statistics.gov.scot/resource?uri=http%3A%2F%2Fstatistics.gov.scot%2Fdata%2Fhouse-sales`

[5] `http://statistics.gov.scot/resource?uri=http%3A%2F%2Fstatistics.gov.scot%2Fdata%2Fhouse-sales-prices`

3. Their unit of measure is ratio
4. Have a continuous or categorical measure
5. (Optionally) have additional dimensions

It should be noted that in this case study we only searched for datasets with ratio unit of measure. However, as also described in section 3, count datasets could also be selected provided that they will be transformed in the next step to ratio values.

In addition, some datasets may be truly correlated with the response variable of our case study i.e. the house sales prices and should be excluded. For example, the Council Tax Bands dataset provides the rate of houses that belong to a specific council tax band in each Scottish data zone. This measure is however actually derived from the price of houses and for this reason we shouldn't include it in our case study. In reality Council tax bands is a discrete measure, which means it aggregates the number of houses according to their value.

The exploration of the Scottish data portal for datasets that satisfy the above criteria results in the selection of 21 compatible datasets.

Feature extraction In this step we extract multiple features from each selected compatible dataset. In our case study each feature is extracted by only one dataset. For instance, the dimensions of the "Age of First Time Mothers" dataset which describes the rate of first time mothers include reference period and age. For the age dimension three values are used: (1) 19 and under, (2) item 35 and over, and (3) All. If we also consider that we have selected 4 reference periods for our datasets (i.e. 2009, 2010, 2011 and 2012), the final number of features that can be extracted from this dataset is calculated as:

2 *(the two values of the age dimension - "All" is not included)* x 1 *(the number of the different unit types)* x 4 *(number of reference periods)* = 8 features.

The same applies to the rest of the selected datasets in order to extract all features. The feature extraction step results in 450 features.

Feature selection and model creation In order to eliminate insignificant features we use the regression analysis method called Lasso. The Lasso implementation was made using the glmnet library[6]. Lasso keeps only the important features (i.e. features which add value to our estimations) and removes the rest of them. A reduced number of features facilitates the interpretation of the results.

In our case study Lasso results in 34 features coming from 10 datasets. The initial number of features (i.e. 450) is hence significantly reduced (by more than 92%). The 10 datasets selected are:

1. Age of First Time Mothers
2. Ante-Natal Smoking

[6] `https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html`

3. Breastfeeding
4. Disability Living Allowance
5. Dwellings by Number of Rooms
6. Employment and Support Allowance
7. Hospital Admissions
8. Household Estimates
9. Income And Poverty Modelled Estimates
10. Job Seeker's Allowance Claimants

Table 1 presents the detailed results of the application of the Lasso analysis method. We can see that there is no significant change between the Lasso lowest RMSE and the Lasso one standard error.

Table 1: The results of the Lasso analysis method

Number of used Observations	5380
Number of Predictors	450
Type of Predictors	Ratio
Year of Predictors	2009-2012
Type of Response	Mean
Year of Response	2012
Lasso lowest RMSE	0.2664764
With number of selected variables	47
Lasso RMSE 1SE	0.2737672
With number of selected variables	34
Percentage of reduction	92%

Lasso uses the cross-validation method to separate the data in training and test data and make the prediction. Cross-validation divides the initial dataset into a number of roughly equal parts (aka folds). In each round of the cross-validation, each fold in turn is used as test data and the rest of the folds as training data. We use the Root Mean Squared Error (RMSE) of the log error to assess the result of Lasso. Log error is the log of the predicted value minus the log of the actual value.

In our case study we randomly select the folds used as test and trained data (using a seed). This means that each time someone repeats the same procedure with the same datasets, he/she will result in different train and test data and, hence, in different RMSE. We repeat the same Lasso analysis using the same datasets 100 times (which is also the default value for the gmlnet library) in order to see the variance of RMSEs during the multiple repetitions. Fig 1 presents two boxplots. The left boxplot illustrates the variance of the RMSE calculated in all 100 repeated Lasso experiments. We can see that the median of the RMSEs is close to 0.273 and that the distance between the median and the lower and upper quartiles is limited. The right boxplot presents the variation of the total number of the selected features based on one SE. We can see that the median number of features is 28 which is also the lower quartile.

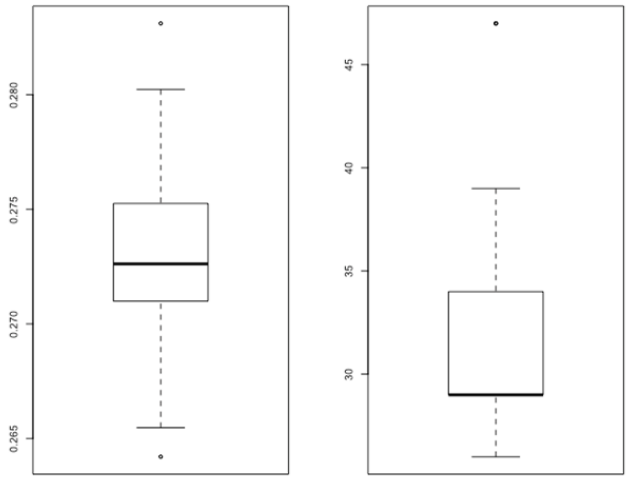

Fig. 1: Variance of RMSEs and total number of selected features

Cross-validation allows selecting the best value for the tuning parameter (lambda), or equivalently, the value of the constraints. To this end, we compute the lambda parameter. Lambda parameter controls the amount of regularization, so choosing a good value for it is crucial. In cases with very large number of features, lasso allows to efficiently find the model that involves a small subset of the features. The value selected for lambda is the one that corresponds to the smallest error or the value with one standard error. The plot in Fig. 2 shows how the RMSE fluctuates for different number of lambda (or features). Higher values of lambda produce less flexible functions and, hence, higher errors while lower values of lambda produce more flexible functions and, hence, lower errors. We select the optimal value for the lambda that corresponds to the minimum RMSE i.e. -4. Following this rule the final number of selected features is 34.

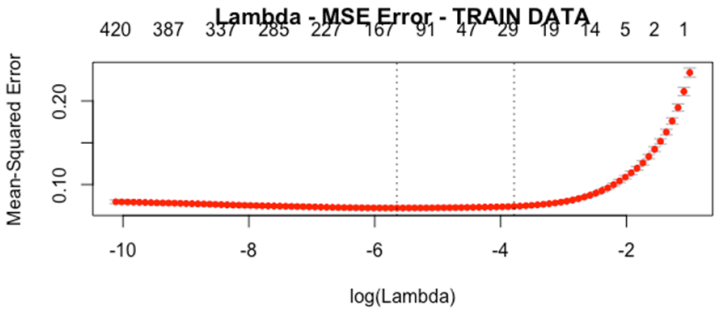

Fig. 2: Lambda - RMSE

5 The Compatible LOSD Selection tool

We develop an open source tool as a proof of concept of the case study. The tool offers an interface that facilitates the selection of compatible statistical datasets that can be used for machine learning analysis. The tool is based on R Shiny[7] and obtains statistical datasets from the Scottish data portal. It allows selecting a dataset from the Scottish portal, and searches and presents compatible datasets based on the defined compatibility criteria. The tool is available on GitHub [8].

Fig 3 presents a screen-shot of the Compatible LOSD Selection tool. On the left panel 2012 house prices has been selected as the first dataset. On the right panel the 20 compatible datasets are presented. The selected compatible datasets can be extracted to contribute in the creation of a predictive model.

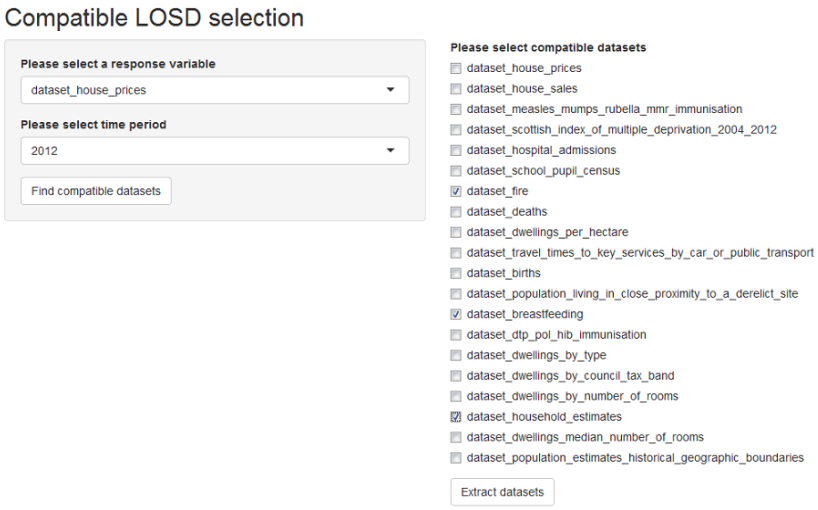

Fig. 3: The Compatible LOSD Selection tool

6 Conclusions

Although governments and other organisations are continuously opening up their statistical data, the potential of open data has been unrealized to a large extent due to institutional and technical barriers. In machine learning analyses, linked data facilitate the discovery, retrieval, and combination of data on the Web. However, a challenge in such analyses is to specify the criteria to be considered in order to select the proper datasets to construct the predictive model.

[7] `https://shiny.rstudio.com/`
[8] `https://github.com/akaramanou/compatible-LOSD-selection-tool`

In this paper we presented a case study that applied machine learning methods to compatible statistical datasets from the Scottish data portal in order to support advanced decision-making scenarios. The case study aimed to predict the house prices in Scotland. To facilitate the discovery of compatible datasets we defined five compatibility criteria. Based on the criteria we discovered 21 datasets compatible with the response variable. From these datasets we extracted 450 features and applied the Lasso method in order to select the most important features. We resulted in 34 features coming from only 10 datasets (over 92% less features than the ones initially identified). This means that there is a strong relationship between the house prices in Scotland and these 10 datasets. We also developed the Compatible LOSD Selection tool that facilitates discovering compatible LOSD datasets to perform machine learning analysis.

This case study is indicative of the importance of using machine learning to analyse statistical datasets and support decision making. Starting from a problem that needed to be solved we resulted in identifying relationships between datasets, some of them previously unknown. For example, our case study revealed a strong relationship between the breastfeeding percentage and the mean house prices in Scotland. Other relationships were more obvious such as the one between Income and Poverty estimates and mean house prices. Eliminating in an easy way all irrelevant datasets can be really beneficial for decision makers as it saves them time from dealing with unessential data and help them understand which variables matter most and which can be ignored. More importantly, this case study proves that decision makers can yet easily exploit historical statistical data using machine learning in order to take evidence-based decisions.

The case study also proved that, when it comes to statistical data, effectively discovering compatible datasets is crucial to be able to create successful predictive models. The compatibility criteria we defined is only a first attempt to define the compatibility between statistical datasets. However, this first attempt proved that discovering compatible datasets forms the basis to extract meaningful results using machine learning.

Acknowledgments. This work is funded in the context of the project "Integrating open statistical data using semantic technologies (MIS code 5007306), EDBM34: Supporting researchers with emphasis on new researchers". The project is co-financed by Greece and the European Union (European Social Fund) by the Operational Programme Human Resources Development, Education and Lifelong Learning 2014-2020.

References

1. Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., Pentland, A.: Once upon a crime: towards crime prediction from demographics and mobile data. In: Proceedings of the 16th international conference on multimodal interaction, pp. 427–434. ACM (2014)
2. Coleman, S.Y.: Data-mining opportunities for small and medium enterprises with official statistics in the UK. Journal of Official Statistics **32**(4), 849–865 (2016)

3. Cyganiak, R., Reynolds, D., Tennison, J.: The rdf data cube vocabulary. W3C Recommendation, W3C (2014)
4. Datta, A., Thomas, H.: The cube data model: a conceptual model and algebra for on-line analytical processing in data warehouses. Decision Support Systems **27**(3), 289–301 (1999)
5. Değirmenci, T., Özbakır, L.: Differentiating households to analyze consumption patterns: a data mining study on official household budget data. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **8**(1) (2018)
6. European Commission: Guidelines on recommended standard licences, datasets and charging for the reuse of documents (2014). C240/1
7. Hassani, H., Saporta, G., Silva, E.S.: Data mining and official statistics: the past, the present and the future. Big Data **2**(1), 34–43 (2014)
8. Janssen, M., Charalabidis, Y., Zuiderwijk, A.: Benefits, adoption barriers and myths of open data and open government. Information systems management **29**(4), 258–268 (2012)
9. Kalampokis, E., Nikolov, A., Haase, P., Cyganiak, R., Stasiewicz, A., Karamanou, A., Zotou, M., Zeginis, D., Tambouris, E., Tarabanis, K.A.: Exploiting linked data cubes with opencube toolkit. In: International Semantic Web Conference (Posters & Demos), vol. 1272, pp. 137–140 (2014)
10. Kalampokis, E., Roberts, B., Karamanou, A., Tambouris, E., Tarabanis, K.A.: Challenges on developing tools for exploiting linked open data cubes. In: SemStats@ ISWC (2015)
11. Kalampokis, E., Tambouris, E., Karamanou, A., Tarabanis, K.: Open statistics: The rise of a new era for open data? In: International Conference on Electronic Government and the Information Systems Perspective, pp. 31–43. Springer (2016)
12. Kalampokis, E., Tambouris, E., Tarabanis, K.: Linked open government data analytics. In: International Conference on Electronic Government, pp. 99–110. Springer (2013)
13. Kalampokis, E., Tambouris, E., Tarabanis, K.: Linked open cube analytics systems: Potential and challenges. IEEE Intelligent Systems **31**(5), 89–92 (2016)
14. Manyika, J., Chui, M., Groves, P., Farrell, D., Van Kuiken, S., Doshi, E.A.: Open data: Unlocking innovation and performance with liquid information. McKinsey Global Institute **21** (2013)
15. Prati, G., Pietrantoni, L., Fraboni, F.: Using data mining techniques to predict the severity of bicycle crashes. Accident Analysis & Prevention **101**, 44–54 (2017)
16. Tambouris, E., Kalampokis, E., Tarabanis, K.: Processing linked open data cubes. In: International Conference on Electronic Government, pp. 130–143. Springer (2015)
17. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) pp. 267–288 (1996)
18. W3C: Data on the web best practices (2017). URL `https://www.w3.org/TR/dwbp/`. W3C Recommendation
19. Zeginis, D., Kalampokis, E., Roberts, B., Moynihan, R., Tambouris, E., Tarabanis, K.: Facilitating the exploitation of linked open statistical data: JSON-QB API requirements and design criteria. In: 5th International Workshop on Semantic Statistics (SemStats2017) co-located with the 16th International Semantic Web Conference (ISWC2017), vol. 1923 (2017)
20. Zhu, Y.Q., Kindarto, A.: A garbage can model of government it project failures in developing countries: The effects of leadership, decision structure and team competence. Government Information Quarterly **33**(4), 629–637 (2016)