Reliability Analyses of Open Government Data

Davide Ceolin[1], Luc Moreau[2] Kieron O'Hara[2], Guus Schreiber[1], Alistair Sackley[3], Wan Fokkink[1], Willem Robert van Hage[4], and Nigel Shadbolt[2]

[1] VU University Amsterdam, The Netherlands
{d.ceolin, guus.schreiber, w.j.fokkink}@vu.nl
[2] University of Southampton, United Kingdom
{l.moreau, kmo}@ecs.soton.ac.uk
[3] Hampshire County Council, United Kingdom
[4] SynerScope B.V., The Netherlands
willem.van.hage@synerscope.com

Abstract. Public authorities are increasingly sharing sets of open data. These data are often preprocessed (e.g. smoothened, aggregated) to avoid to expose sensible data, while trying to preserve their reliability. We present two procedures for tackling the lack of methods for measuring the open data reliability. The first procedure is based on a comparison between open and closed data, and the second derives reliability estimates from the analysis of open data only. We evaluate these two procedures over data from the `data.police.uk` website and from the Hampshire Police Constabulary in the UK. With the first procedure we show that the open data reliability is high despite preprocessing, while with the second one we show how it is possible to achieve interesting results concerning the open data reliability estimation when analyzing open data alone.

1 Introduction

Open Government Data are valuable for boosting the economy, enhancing the transparency of public administration and empowering the citizens. These data are often sensitive and so need to be preprocessed for privacy reasons. In the paper, we refer to the public Open Government Data as "open data" and to the original data as "closed data".

Different sources expose open data in different manners. For example, Crime Reports [4] and `data.police.uk` [15] both publish UK crime data, but in different format (maps vs. CSV files), level of aggregation, smoothing and timeliness (daily vs. monthly update), which all represent possible reasons for reliability variations. For different stakeholders it is important to understand how reliable different sources are. The police, who can access the closed data, needs to know if open data are reliable enough e.g. to be used in projects involving the citizens. The citizens wish to know the reliability of the different datasets to understand the reasons for differences between authoritative sources. We present two procedures to cope with the lack of methods to analyze these data: one for computing the reliability of open data by comparing them with the closed data, and one to estimate variations in the reliability of the open data by relying only on these.

The analysis of open data is spreading, led by the Open Data Institute (`http://www.theodi.org`) and others. For instance, Koch-Weser [10] presents an interesting analysis of the reliability of China's Economic Data, thus analyzing the same aspect as we are interested in, on a different typology of dataset. Tools for the quality estimation of open data are being developed (e.g. Talend Open Studio for Data Quality [14], Data Cleaner [8]), but their goal is less targeted than ours, since they aim at quantifying the quality of open data in general as to provide a substrate for a more comprehensive open data analysis infrastructure. Relevant for this work is also a paper from Ceolin et al. that uses a statistical approach to model categorical Web data [3] and one that uses provenance to estimate reliability [2]. We plan to adopt the approach proposed by Ebden et al. [6] to measure the impact of different processes on the data.

The rest of this paper is structured as follows: Section 2 describes a procedure for measuring the reliability of open data given closed data and a case study implementation; Section 3 presents a procedure for analyzing open data and a case study; lastly, Section 4 provides final discussion.

2 Procedure for Comparing Closed and Open Data

The UK Police Home Office aggregates (i.e., presents coarsely) and smoothens (introduces some small error) the open data for privacy reasons. We represent the open data provenance with the PROV Ontology [17] as in Fig. 1. In general, a faulty aggregation process or aggregating data coming from heterogeneous sources not properly manipulated might unexpectedly affect the resulting data reliability, while smoothing should affect it explicitly but in a limited and controlled manner. The following procedure aims at capturing such variation:

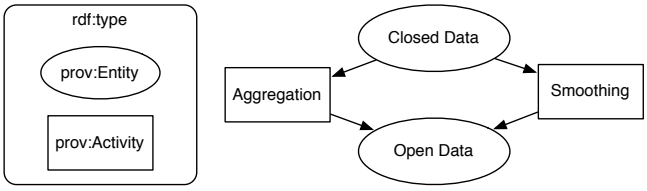

Fig. 1: Open Data Creation Provenance.

Select the relevant data Closed data might be spurious, so we select the data items that are relevant for our analyses. The selection of the data might involve the temporal aspect (i.e. only data referring to the relevant period are considered), their geographical location (select only the data regarding the area of interest), or other constraints and their combination;

Roll up categorical data There exists a hierarchy of categories because each level is available to a different audience: open data are presented coarsely to the citizens, while closed data are fine grained. We bring the categorization to the same level, hence bringing the closed data to the same level as the open data.

Compare the corresponding counts Different measures are possible, because the difference between datasets can be considered from different points of view: relative, absolute, etc.. For instance, the ratio of the correct items over the total amount or the Wilcoxon signed-rank test [18].

Case Study 1 We compare a set of crime counts per categories grouped per neighbourhood and month from `data.police.uk` with a limited set (30,436 relevant entries) of corresponding closed data from the Hampshire Constabulary by implementing the procedure above as follows:

Data Selection Select the data for the relevant months and geographical area. In this latter case, we load the KML file describing the Hampshire Constabulary area using the maptools library [1] in the R environment [13] and check if the crimes coordinates occur therein using the SDMTools library [16];

Data Aggregation We apply two kinds of aggregation: **temporal**, to group together data about the same month and **geographical**, to aggregate per neighbourhood. The closed data items report the address of occurrence of the crimes, while the open data are aggregated per police neighbourhood. We match the zip code of the addresses and the neighbourhoods using the MapIt API [11].

Data Comparison We average the result of the Wilcoxon signed-rank test applied per neighbourhood, to compare open and aggregated closed data.

For each neighbourhood we compute a Wilcoxon signed-rank test to check the significance of the difference between open and closed data and we average the outcomes (see Table 1a). We compute the test on the differences of the two counts (open and closed data) to check whether the estimated average of the distribution of the differences is zero (that is, the two distributions are statistically equivalent) or not.

The results at our disposal are limited, since we could analyze only two complete months. Still, we can say that smoothing, in these datasets, introduces a small but significant error. The highest error average (2.75) occurs with the entry with the highest error variance: this suggests that the higher error is due to a few, sparse elements, and not to the majority of the items. To prove this, we checked the error distribution among the entities and we reported the results in Table 1b. A χ^2 test [12] at 95% confidence level confirms that the two error distributions do not differ in a statistically significant manner.

3 Procedure for Analyzing Open Data

We propose here a procedure for analyzing open data alone, to be used when closed data are not available, which provides weaker but still useful results,

Table 1: Statistics about the errors in the comparison between open and closed data, and error distibution.

(a) Statistics about the comparison of open and closed data.

Months	Avg error	Var error	% Different Entries
month_1	2.75	12.28	79%
month_2	0.86	3.52	86%

(b) Percentage of items in each open dataset presenting a relative error of at most 0%, 25%, 50%, 75% and 100% with respect to the corresponding closed data item.

Month	% of Entries per Relative Error				
	0%	$\leq$ 25%	$\leq$ 50%	$\leq$ 75%	$\leq$ 100%
month_1	35%	44%	65%	74%	96%
month_2	34%	43%	57%	65%	91%

compared to the previous one. It compares each dataset with the consecutive one, measures their similarity and pinpoints the occurrence of possible reliability changes based on variations of similarity over time. We use a new similarity measure for comparing datasets, that aggregates different similarity "tests" performed on couples of datasets. Given two datasets d_1 and d_2, their similarity is computed as follows:

$$sim(d_1, d_2) = avg(t_1(d_1, d_2), \ldots, t_n(d_1, d_2))$$

where avg aggregates the results of n similarity tests t_i, with $i \in \{1 \ldots n\}$. We propose the following families of tests, although we are not restricted to them:

Statistical test Check with a statistical test (e.g. Wilcoxon signed-rank test) if the data are drawn from significantly different distributions.

Model Comparison test Build a model (e.g. linear regression [7] or Support Vector Machines [5]) on one of the two datasets and evaluate its performance (precision, recall) over the other dataset. These models represent an abstraction over the first dataset and by evaluating them over the other one, we check, according to such a model, how similar the two datasets are.

The tests can be aggregated, for instance, by averaging them or by merging them in a "subjective opinion" [9], which is a construct of a probabilistic logic that is equivalent to a Beta probability distribution about the correct value for the similarity. The expected value of the Beta is close to the arithmetical average, but the variance represents the uncertainty in our calculation, since it reduces as long as we consider more tests. The similarity measure alone does not stand for reliability: there can be many reasons for a similarity variation (e.g. a new law or a particular event that makes the crime rate rise) without implying a reliability change. Also, a similarity value alone might be difficult to interpret in terms of reliability, when a gold standard is not available. So we analyze the similarity of consecutive datasets to pinpoint items that possibly present reliability variations: if the similarity between datasets remains similar for a period of time, and then a variation occurs, one of the possible reasons for such a variation is a change in the data reliability. Unfortunately, we can not discriminate between this and other causes, unless we have additional information at our disposal.

Case Study 2 We analyze the police open data for the Hampshire Constabulary from `data.police.uk`, that consist of crime counts, aggregated per neighbourhood from April 2011 to December 2012. We know that in this period open data creation policy changes occurred. These might have affected the datasets reliability. We compare the distribution of the crime counts among the crime categories, and we represent the similarity between two datasets as the percentage of neighbourhoods that are statistically similar (according to a Wilcoxon signed-rank test). The results of the comparison are reported in Figure 2, where each point represents the similarity between two datasets, in sequence. At the twelfth comparison the similarity trend breaks and then starts a new one. That is likely to be a point where the reliability diverges as the similarity variation possibly hints, and it actually coincides with a policy change (the number of neighbourhoods varies from 248 to 232), and since the area divided by these neighbourhoods is the same, this possibly introduces a variation in the impact of the smoothing error, but we do not have at our disposal a confirmation of such impact. As we stressed earlier, the procedure allows us only to pinpoint possibly problematic data, but without additional information, our analysis cannot be precise, that is, we cannot be certain about the reason of the similarity change.

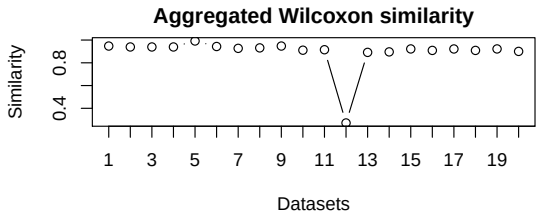

Fig. 2: Plot of the similarity of consequent datasets of crime counts for the Hampshire Constabulary from the `data.police.uk` website.

4 Conclusion and Future Work

We presented two procedures for the computation of the reliability of open data: one based on the comparison between open and closed data, the other one based on open data alone. Both procedures have been evaluated using data from the `data.police.uk` website and from the Hampshire Police Constabulary in the UK. The first procedure allows us to estimate the reliability of open data, and shows that smoothing procedures, although introducing some error, preserve a high data reliability. The second procedure is useful to grasp indications about the data reliability, although more weakly than the first one, since it allows only to pinpoint possible reliability variations in the data. Despite the fact that open data are exposed by authoritative institutions, these procedures allow us to enrich the open data with information about their reliability, to increase the

confidence of both the insider specialist and the common citizen who use them and to help in understanding possible discrepancies between data exposed by different authorities. We plan to extend the range of analyses applied and of datasets considered. Moreover, we intend to map the data with Linked Data entities to combine the statistical analyses with semantics.

Acknowledgements This work is supported in part under SOCIAM: The Theory and Practice of Social Machines; the SOCIAM Project is funded by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/J017728/1.

References

1. R. Bivand. *Tools for reading and handling spatial objects*, 2013.
2. D. Ceolin, P. Groth, and W. R. van Hage. Calculating the Trust of Event Descriptions using Provenance. In *SWPM*, pages 11–16. CEUR-WS.org, Nov. 2010.
3. D. Ceolin, W. R. van Hage, W. Fokkink, and G. Schreiber. Estimating Uncertainty of Categorical Web Data. In *URSW*, pages 15–26. CEUR-WS.org, 2011.
4. CrimeReports. Crimereports. `https://www.crimereports.co.uk/`, July 2013.
5. N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.
6. M. Ebden, T. D. Huynh, L. Moreau, S. Ramchurn, and S. Roberts. Network analysis on provenance graphs from a crowdsourcing application. In *IPAW*, pages 168–182. Springer-Verlag, 2012.
7. F. Galton. Regression Towards Mediocrity in Hereditary Stature. *Journal of the Anthropological Institute*, 15:246–263, 1886.
8. Human Inference. DataCleaner. `http://datacleaner.org`, 2013.
9. A. Jøsang. A logic for uncertain probabilities. *Int. Journal Uncertainty Fuzziness Knowledge-Based Systems*, 9(3):279–311, June 2001.
10. I. N. Koch-Weser. The Reliability of China's Economic Data: An Analysis of National Output. `http://www.uscc.gov/sites/default/files/Research/TheReliabilityofChina'sEconomicData.pdf`, Jan. 2013.
11. mySociety. MapIt. `http://mapit.mysociety.orgs`, July 2013.
12. K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–175, 1900.
13. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Sept. 2012. ISBN 3-900051-07-0.
14. Talend. Talend Open Studio for Data Quality. `http://www.talend.com/products/data-quality`, 2013.
15. United Kingdom Police Home Office. data.police.uk. `data.police.uk`, July 2013.
16. J. VanDerWal, L. Falconi, S. Januchowski, L. Shoo, and C. Storlie. *SDMTools: Species Distribution Modelling Tools: Tools for processing data associated with species distribution modelling exercises*, 2012.
17. W3C. PROV-O: The PROV Ontology. `http://www.w3.org/TR/prov-o/`, July 2013.
18. F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.