Two Procedures for Analyzing the Reliability of Open Government Data

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

[1] VU University Amsterdam
{d.ceolin, w.j.fokkink, v.maccatrozzo, guus.schreiber}@vu.nl
[2] University of Southampton
{l.moreau, kmo, nrs}@ecs.soton.ac.uk
[3] Hampshire County Council
{alistair.sackley@hants.gov.uk}
[4] SynerScope B.V. {willem.van.hage@synerscope.com}

Abstract. Open Government Data often contain information that, in more or less detail, regard private citizens. For this reason, before publishing them, public authorities manipulate data to remove any sensitive information while trying to preserve their reliability. This paper addresses the lack of tools aimed at measuring the reliability of these data. We present two procedures for the assessment of the Open Government Data reliability, one based on a comparison between open and closed data, and the other based on analysis of open data only. We evaluate the procedures over data from the data.police.uk website and from the Hampshire Police Constabulary in the United Kingdom. The procedures effectively allow estimating the reliability of open data and, actually, their reliability is high even though they are aggregated and smoothed.

1 Introduction

Open Government Data are often sensitive and hence need to be properly processed in order to reduce the amount of personal information exposed. This process consists of aggregation and so-called "smoothing" procedures which introduce some imprecision in the data, to avoid the reconstruction of the identity of a citizen from a piece of data. The value of this data might be affected by such procedures, as they limit the extent to which we can rely on them. Throughout the paper, we will refer to the published Open Government Data as "open data" and to the original data as "closed data".

Open data are exposed in different modalities by different sources. For instance, Crime Reports [5] and `data.police.uk` [15] both publish data about crimes occurring in the UK, but in different format (maps versus CSV files), level of aggregation, smoothing and timeliness (daily versus monthly update). The smoothing process unavoidably introduces some error in the data. There might be other reasons as well for possible reliability differences among these

datasets, like the fact that a given dataset is not based on timely data (or, in general, it is generated from questionable data sources) or the fact that an erroneous aggregation process inadvertently introduced some mistakes. For the police, as well as for citizens, it is important to understand how different two sources are, in order to understand how much they can rely on the data they expose. The police, who can access the original, raw data, is interested in measuring the reliability of the open data in order to know how much they can rely on them, e.g., when establishing projects involving citizens. For citizens, it is important to understand the reliability of the different datasets, since that represents a reason why data exposed by authoritative sources may present discrepancies.

Our goal is to cope with the lack of tools and methodologies to actually measure and compare these data. We address this problem by means of a twofold contribution: first, we propose a procedure for computing the reliability of an open dataset, when having at our disposal both the open and the closed data. We apply this procedure on a set of UK police data. We show that the reliability of these data is not highly affected by the smoothing and aggregation procedures applied to them and that this procedure, once properly instantiated, allows guiding the analyzer to the discovery of points of policy changes with regard to open data creation and reliability variations. Second, we show how it is possible to estimate variations in the reliability of the open data by comparing them to each other, when the closed data are not at our disposal. Both procedures aim to measure and compare these datasets from the reliability point of view, and to guide a human analyzer to the discovery of possible critical points (e.g., policy changes, relevant errors) in these datasets. In both cases, *the reliability of an open dataset is measured as the percentage of non-significantly different entries from the corresponding closed dataset.* In the case of the procedure for analyzing open data only, we can only estimate a reliability variation, but not measure it.

The rest of this paper is structured as follows: Section 2 describes related work; Section 3 describes a procedure for determining the reliability of open data given closed data and Section 4 presents a procedure for analyzing open data. In Section 5 we put forward a case study implementation of both procedures. Section 6 provides a final discussion.

2 Related Work

This work completes a previous work from the same authors [2], by improving the open data procedure and extending the validation of both procedures. The analysis of open data is increasingly being spread, for instance, by the leading Open Data Institute [14]. Koch-Weser [11] presents a work on the analysis of the reliability of China's Economic Data which, although focused on a different domain, shares with this work the goal to understand the reliability of open data. Tools for the quality estimation of Open Data are being developed (see for instance Talend Open Studio for Data Quality [13] and Data Cleaner [7]). These tools are designed to understand the adherence of data to particular standards, similar to our goals, but they aim at constituting a proper middleware component

of the entire business process of data management and curation. These tools are not limited to monitoring data quality, but they aim also at quantifying the risk and the financial impact of these data, as well as how to intervene in the business process in case of any problem discovered. Our goal is less business-oriented and more targeted, as we aim at developing procedures for measuring and estimating open data reliability. However, this can be seen as a step towards the development of a more comprehensive tool.

A paper of Ceolin et al. [3] shares with the work here presented the statistical approach in modeling categorical Web data and the use of the Wilcoxon signed-rank test (which is a non-parametric hypothesis test that determines whether two probability distributions are significantly different [16]) to measure the reliability of these data. We do not have at our disposal information about the impact of the different processes on the reliability of the resulting data, but in the future we plan to adopt an approach similar to the one of Ebden et al. [6] to understand this a posteriori. Closer to the topic of the case studies analyzed, i.e., the reliability of Police Open Data, this work can be seen as complementary to the one of Cornelli [4], who researches on the reasons citizens have to trust police.

3 Procedure for Comparing Closed and Open Data

Closed data are aggregated and smoothed in order to not expose sensitive information when publishing them. Aggregation, that is to present the data at a coarser, higher level than available, preserves the data correctness and reduces their granularity. It is not intended to introduce imprecisions, but a faulty aggregation process or the wrong use of heterogeneous data sources might unexpectedly affect the data reliability. "Smoothing" is an anonymization procedure used especially when aggregation is not sufficient to guarantee anonymity (e.g., in case of data about low-populated areas). By smoothing, authorities voluntarily introduce some small errors in the data so that they remain reliable at a coarse level, but it is not possible (or at least, hard) to reconstruct the details of the single items. We describe a procedure to evaluate the reliability gap existing between open and closed data, if any. The procedure is generic and in Section 5 we propose some possible implementations.

Select the relevant data This selection might involve temporal aspects (i.e., only data referring to the relevant period are considered), or their geographical location (select only the data regarding the area of interest). Other constraints and their combination are possible as well.

Roll up categorical data The categories used to classify the categorical data are ordered in hierarchies. Hierarchies are created to define different categories for different refinement levels when presenting categorical data. We cannot increase the refinement of the data categorized in a coarser manner, so we decrease the granularity level of the closed data.

Roll up smoothed categorical data This step is similar to the previous one, besides the fact that the expected result is not necessarily coincident with the original one since smoothing may affect the data precision.

Compare the corresponding counts by using, for instance, the ratio of the correct items over the total amount or the Wilcoxon signed-rank test [16].

4 Procedure for Analyzing Open Data

Open data counts may differ from each other with respect to different points of view (absolute differences, data distribution, etc.). We do not know a priori what is the best manner to compare the data counts, so we aggregate several similarity tests performed on pairs of datasets. When analyzing open datasets, we can compare only data about related facts: for instance, the typology of facts can be the same (e.g., crimes in a given area), but the time they refer to differs. The results that we can expect from this kind of analyses are much less detailed and definite than before: since we do not have at our disposal a gold standard, we can not test properly our hypotheses. We estimate points of reliability change using the following procedure, based on the idea that by analyzing the similarity of the datasets using different similarity measures, these changes can emerge.

Choose one or more dataset similarity scores. We compute the similarity of two dataset d_1 and d_2 as: $sim(d_1, d_2) = avg(t_1(d_1, d_2), \ldots, t_n(d_1, d_2))$ where avg computes the average of the results of the similarity scores resulting from the tests t_i on the items (i.e., values) in d_1 and d_2. We use the Wilcoxon signed-rank test to check if the data counts differ significantly. Other tests are possible as well. The similarity between two dataset is obtained by aggregating these tests using, for instance, a (weighted) arithmetic average. In subjective logic [8], we can treat the tests as "subjective opinions" about the similarity of the two datasets, and merge them using the "fusion" [9] operator to obtain a beta probability distribution describing the probability for each value in $[0, 1]$ to be the correct similarity value.

Compute the similarity, with one or more scores. Measure the pairwise similarity between each dataset and the one of the following month (crime counts are aggregated on monthly bases).

Identify change points in the similarity sequence. Change points in the similarity sequence are likely to indicate policy changes in the data creation and hence reliability changes resulting from these policy modifications.

Aggregate all the evidence of policy changes. Each change point identified in the previous step represents evidence of a policy change. We run more similarity analyses to reduce the risk of error. Since we are dealing with uncertain observations, we also adopt subjective opinions here and we compute a binomial opinion for each dataset.

There can be natural reasons that explain a variation in the data (e.g., a new law) that do not imply a lack of reliability in one of the two datasets. Moreover, a similarity value taken alone may be difficult to interpret: what does it mean that the similarity between two datasets is, e.g., 0.8? So, we focus on similarity trends and not on single values, and we pinpoint variations in such trends, since such variations have a higher chance to indicate a change in the data reliability.

5 Case Study - Police Open Data Analyses

We evaluate the procedures that we propose over police data for the Hampshire Constabulary. As open data we adopt the corresponding entries from the `data.police.uk` website, in particular in the interval from April 2011 until December 2012. `data.police.uk` data are released monthly, and aggregated in time within a month and in space to the level of police neighborhoods, that comprise at least eight postal addresses. We focus on the datasets presenting the counts aggregated per police neighborhood because this kind of classification, although not as detailed as the classification per address, allows an easy comparison between entries and reduces the burden of having to geolocate and disambiguate addresses. As closed data, we have at our disposal a series of datasets from the Hampshire Police Constabulary covering the interval from October 2010 until September 2012 and reporting distinct information for each single crime in that period. The two datasets do not perfectly overlap, but we focus mainly on the intersection between the two intervals covered, which still is the largest part of both datasets. We show that the two procedures allow providing similar findings, even though the first procedure is clearly less uncertain than the second one.

5.1 Analyzing the Reliability of Police Open Data

We focus on the intersection between the open and the closed data at our disposal (that is, the period from April 2011 until September 2012).The data at our disposal contain: category, date and geographic Cartesian coordinates of the crime. Following the procedure described in Section 3, we compare the distribution of the crime counts among the different categories for each neighborhood using a statistical test and we aggregate the results in a subjective opinion, because we consider the outcomes of the tests as pieces of evidence about the reliability of the open data, and we treat them as error-prone observations.

Data preprocessing. First, we convert the coordinates from the Cartesian system to latitude and longitude using the RGDAL library [1]. Then we look up the postal code that is closest to the point where the crime happened. This step potentially introduces some error in the analyses, because of the approximation in the coordinates conversion and because although looking up the closest postal code to the point that we are analyzing is the best approximation we can make, but it is not always correct. We manually checked some sample items to confirm the robustness of this procedure. Our results show that the impact of these imperfections is limited. It was not possible to compute the postal code of all the points that we had at our disposal, because some data entries were incomplete and some incorrect.

Select the relevant data. First, we query the MapIt API [12] in order to retrieve the police constabulary each postal code belongs to and discard the crime items not belonging to the Hampshire Constabulary in the closed datasets. Second, we select the open data for the months for which closed data are available. Lastly, we exclude crime counts of categories not shared

between closed and open data. For instance, the "Anti-social behaviour" category is present in the open data but not in the closed data.

Aggregate the data. Also data aggregation is performed in three steps. **Temporal aggregation** is made to group together data about crimes occurring in the same month. **Geographical aggregation** is made to aggregate the data at police neighborhood level. To aggregate the data at neighborhood level, we match zip code and neighborhood by querying the MapIt API [12]. **Categorical aggregation** is performed by aligning the classifications of the crimes in the open and closed datasets, that is, by bringing the closed data at the same, coarse, level as the open data. The categories of open and closed data belong to the same hierarchy, but closed data are classified using fine grained categories, open data using coarser ones.

Compare the aggregated data. Once the items are brought to the same level of aggregation, the reliability of the open data is measured. The comparison is made at neighborhood level. We **apply a Wilcoxon signed-rank test** to these counts to check (at 95% confidence level) if the two crime counts are significantly different (negative observation) or not (positive observation) and also we **measure the differences between the crime counts**. The results of the comparisons are aggregated using **binomial subjective opinions**, or the equivalent beta distributions. Of these, we use the expected value, $E = \frac{p+1}{p+n+2}$ where p and n are the amounts of non-significantly and significantly different entries respectively. Alternatively, we make use of **arithmetic average**. Given the high number of observations, the difference between the results obtained with the two methods is negligible.

Results We analyze the datasets from April 2011 to September 2012, that is the interval for which we have closed data. We start by comparing the distribution of crime counts per category on the intersection of neighborhoods in the closed and open datasets. We want to check if the distribution of the crime counts in the matching neighborhoods is affected by data manipulation procedures. We present a series of graphs resulting from a series of analyses in R. The closed data at our disposal are quite complete, but they do not match perfectly the open data, as we can see from Fig. 1(a). We apply the Wilcoxon-signed rank test on the crime counts of each matched neighborhood to check if the rank of the crime categories in terms of crime occurrences is preserved. We can see in Fig. 1(c) that the open datasets score high, as to confirm their high reliability, in the overlapping neighborhoods. The error introduced by smoothing may move the geolocation of one crime item to a wrong neighborhood. We extend the open and the closed datasets to have them covering the same neighborhoods: when a neighborhood is not present in one of the two datasets, we assume that it presents zero crimes there. In this way we take the point of view of laymen people, who can deal only with the open data without knowing which are the overlapping neighborhoods. Fig. 1(d) addresses the issue of how the open data are representative of the actual crime distribution in that area. There are at least two trends which, we suspect, correspond to policy changes. One change possibly regards

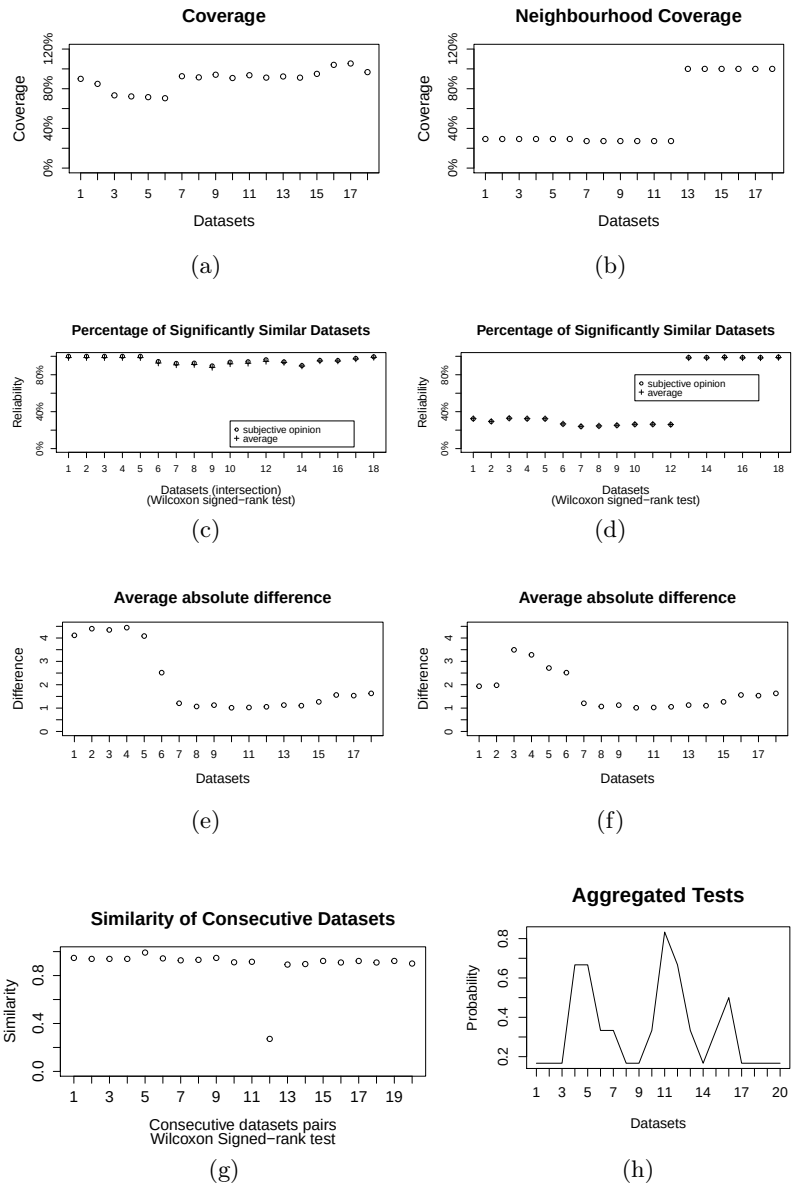

Fig. 1. Plots of the open data coverage in terms of counts (Fig. 1(a)) and neighborhoods (Fig. 1(b)). Follow the plots of the closed data analyses using the Wilcoxon signed-rank test (Fig. 1(c) and 1(d)) and count differences (Fig. 1(e) and 1(f)). Finally, we present a plot of the similarity of consequent datasets using again the Wilcoxon signed-rank test (Fig. 1(g)), and plot of the aggregated tests of the similarity of consecutive datasets (Fig. 1(h)).

the smoothing technique adopted, which determines the neighborhood a crime belongs to. Fig. 1(b) shows that initially only about 30% of the neighborhoods were present in both the open and closed datasets, and then this percentage suddenly rose to 100%. This is due to a change in the smoothing algorithm that makes the more recent open data more reliable and explains the "step" shown in Fig. 1(d). Starting from the sixth dataset, the reliability of the extended datasets corresponds to the percentage of matching neighborhoods. In Fig. 1(d) there is also another change point, between the fifth and the sixth dataset. We identify it by checking the absolute errors in the data, averaged per neighborhood and per crime category (see Fig. 1(e)). Here also there are two trends. The first one breaks were approximatively we expected (at the sixth month instead of at the fifth), and these two trends match the trends shown in Fig. 1(a). So we focused on the fifth and sixth datasets (August and September 2011), and three facts emerged.

First, the closed datasets at our disposal do not contain crime counts for the "Drugs" category for July, August and September 2011. We suppose that the closed data at our disposal lack counts for that category in that time interval because that category presents relevant figures in the other months, although it is possible that no drug crime occurred in that period. Still, the procedure correctly identifies a possible anomaly in the data.

Second, in the September 2011 open dataset, the entry relative to the neighborhood "2LW02" presents two very high figures for "Drugs" and "Other.theft", 205 and 319 respectively (the average counts for these categories in the other neighborhoods are 8.44 and 1.68). A similar pattern recurs only in July and September 2012. We suspect that those high counts share the same explanation, and, although we can not verify this with the information at our disposal, the procedure identifies another possible critical point in the data.

Third, a policy change occurred. From September 2011, the set of crime categories was extended to include also {Criminal.damage.and.arson, Shoplifting, Other.theft, Drugs, Public.disorder.and.weapons}. Before, the corresponding crimes were generically classified as "Other.crime". We reclassified the crimes in the first trend belonging to those categories as "Other.crime", and we recomputed the average differences. The error decreases (on average, of 1.56 counts per month, see Fig. 1(f)). In this part, the error is still high because the correct crime classification contains fewer categories than the rest of the datasets, so here the same error weighs more.

Thanks to the procedure proposed, we: (1) discovered two changes in the open data policies that affect data reliability (one about crime classification, one about smoothing); and (2) measured the reliability of these datasets.

5.2 Estimating the Reliability of Police Open Data

We analyze open data by applying the following procedure.

Compute the similarity of the neighborhoods of consecutive datasets. If more than one similarity measure has been chosen, then aggregate the scores for each neighborhood.

Aggregate the similarity scores of the neighborhoods to obtain an overall similarity value. We aggregate using subjective opinions, to take into account also the uncertainty in the sample, that is quite small.

Look for variations in the series of similarities that may signal a policy variation, automatically, by means of the *changepoint* package in R [10].

Aggregate all the evidence per dataset couple telling whether that couple is a change point or not.

Results We apply the procedure introduced above using different similarity tests (Wilcoxon signed-rank test, absolute differences between counts, etc.). From each test we extrapolate a series of change points, by analyzing the variations in the mean of the cumulative sums (*multiple.mean.cusum* function of the *changepoint* package) and we aggregate them, again by means of an R script and the results are shown in Fig. 1(h). We start from an analysis which is similar to one performed before. We compare, on neighborhood basis, the distribution of the crime counts among the crime categories, and we represent the similarity between two datasets as the percentage of neighborhoods that are statistically similar (using a Wilcoxon signed-rank test). The results of the comparison are reported in Figure 1(g). The datasets are indicated by means of a sequential number (the first circle corresponds to the similarity between the first and the second dataset, and so on). The plot highlights that the twelfth comparison constitutes a change point: before that, the files are highly similar to each other, and likewise after it. But at that point, the similarity trend breaks and starts a new one: that is likely to be a point where the reliability of the datasets diverges. We have found one of the discontinuity points we discovered in Section 5.1 (see Fig. 1(h)). There is also a third peak, less pronounced, but it is not connected to any policy change we are aware of. Also, despite the previous case, we can not say whether a change point indicates the start of an increase or decrease in reliability. Still, these results are useful to facilitate a human analyzer to understand the eventual magnitude of the reliability variation.

6 Conclusions

We present 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 the analysis of open data. Both procedures are evaluated using data from the `data.police.uk` website and from the Hampshire Police Constabulary in the United Kingdom. The first procedure effectively allows estimating the reliability of open data, showing also that smoothing procedures preserve a high data reliability, while allowing anonymizing them. Still, we can see an impact of the procedures adopted to produce these data on their reliability, and the most recent policies adopted show a higher ability to preserve data reliability. The second procedure is useful to grasp indications about data reliability and to identify the same critical points detected using the first procedure. The quality of the results achieved with this method is lower than the one achieved with

the first method, because it does not allow directly and explicitly estimating the reliability of the data analyzed. However, the results obtained are useful and retrieved in a semi-automated manner. These two procedures provide additional value to the open data, as they allow enriching the data with information about their reliability: even though these data are already provided by authoritative institutions, these procedures can increase the confidence both of insider specialists (the first procedure, which relies on closed data) and of common citizens (the second procedure, which relies only on open data) who deal with them.

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, T. Keitt, B. Rowlingson, E. Pebesma, M. Sumner, and R. Hijmans. *RGDAL: Bindings for the Geospatial Data Abstraction Library*, 2010. `https://r-forge.r-project.org/projects/rgdal/`.
2. D. Ceolin, L. Moreau, K. O'Hara, G. Schreiber, A. Sackley, W. Fokkink, W. R. van Hage, and N. Shadbolt. Reliability Analyses of Open Government Data. In *URSW*, pages 34–39. CEUR-ws.org, 2013.
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. R. Cornelli. *Why people trust the police. An empirical study.* PhD thesis, Università degli Studi di Trento, International Ph.D. in Criminology, 2003-02-13.
5. CrimeReports. Crimereports. `https://www.crimereports.co.uk/`, 2013.
6. M. Ebden, T. D. Huynh, L. Moreau, S. Ramchurn, and S. Roberts. Network analysis on provenance graphs from a crowdsourcing application. In *IPAW'12*, pages 168–182. Springer-Verlag, 2012.
7. Human Inference. DataCleaner. `http://datacleaner.org`, 2013.
8. A. Jøsang. A logic for uncertain probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(3):279–311, 2001.
9. A. Jøsang. The consensus operator for combining beliefs. *Artificial Intelligence Journal*, 142:157–170, 2002.
10. R. Killick and I. A. Eckley. *changepoint: An R Package for Changepoint Analysis*, 2013. `http://www.lancs.ac.uk/~killick/Pub/KillickEckley2011.pdf`.
11. 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`, 2013.
12. Mapit. Mapit. `http://mapit.mysociety.orgs`, 2013.
13. Talend. Talend Open Studio for Data Quality. `http://www.talend.com/products/data-quality`, 2013.
14. The Open Data Institute. The Open Data Institute. `www.theodi.org`, 2013.
15. United Kingdom Police Home Office. data.police.uk. `data.police.uk`, 2013.
16. F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.