**Administrative data**

*A dangerous gold-mine in official statistics*

Anton Färnström
The National Crime Prevention Council of Sweden
antonfa@gmail.com

**Abstract**

Fast technological developments have enabled state authorities around the world to generate, store and communicate huge amounts of information every day. Such data is often timely, comprehensive and highly cost-effective, making it a potential gold mine for the production of official statistics. Unfortunately, research investigating the shortcomings associated to these sources has not been at par with the utilization of the data. In particular, little information about the errors in the raw data makes it difficult to understand the quality of any downstream statistics. This is especially challenging to investigate since the data registration is often handled by a different institution than the statistical authority. To address this, the Crime Prevention Council of Sweden has undertaken an in-depth study regarding the quality of the statistics of cleared offences. We used a combination of automatic text classification, interviews and literature studies to get a holistic understanding of the statistical production process. Our results revealed large measurement errors in the raw data, making the statistical product flawed. Interestingly, the multifaceted approach applied in the study proved invaluable in reaching these results. Further, we generalize this approach into a model on how to investigate measurement errors in statistics based on administrative sources.

 **1. Introduction**

A number of factors underpin the increasing relevance of administrative data for statistical purposes. The continuously improved IT solutions for efficient and safe storage of large amounts of data, as well as a higher IT literacy in the general population makes these sources increasingly attractive. At the same time there are demands on the statistical authorities to produce more and better statistics and to decrease the response burden while not increasing expenses [1].

As the use of administrative registers for statistical purposes increases, so does the need to develop suitable theories and methodologies to assess the quality of these statistics. However, this research field is largely undeveloped by comparison to survey methodology [2]. Nonetheless, a number of authors have carried out significant work outlining how the error sources for register-based statistics could be dealt with and conceptualized. For example; Zhang has proposed a framework where potential errors sources for registry-based statistics are mapped to the production processes of single source and integrated micro-data[2]. Wallgren and Wallgren have recently published the second edition of their book concerning register-based statistics, where they aim at describing and explaining the methods that should be used for register surveys [3]. In collaboration with Statistics Netherlands, Bakker and Daas have done much to propose tools for assessing the quality of administrative data, while pushing the theoretical work forward [4,5].

Many of the studies in this field take on a system-wide perspective, where all the potential error sources and the full production process is mapped and elaborated upon. This allows establishing frameworks that enable complete assessments of the quality of the statistics. While this serves as a very good starting point for any producer of official statistics, the relative importance of some error sources may get lost.

In this paper, the argument is made that more focus should be directed towards the data generating process, which is one error source where administrative sources differ starkly from statistical surveys. Further, based on the experience from in-depth quality studies done by the Swedish National Council for Crime Prevention, a simple model on how to approach and detect measurement errors is proposed.

**2. Measurement errors in the Swedish crime statistics**

The Swedish National Council for Crime Prevention is responsible for producing, administering and developing Sweden’s official crime statistics. The raw-data for the statistics are collected via automated data transfers from the authorities in the legal sector such as the Police, the Prosecution Authority and the Courts. As a way to increase the understanding of which quality the raw data holds, the Council have commenced a series of quality studies aimed to bring greater clarity into this.

In April 2014 the Council published its second study in this series, which focus on the so-called decision codes used by the Police. When a police officer decides to terminate or dismiss an investigation, this decision should be registered in the police’s case management system. When registering these decisions, police officers have to choose between different decision codes. The choice of decision code reflects the reason for the decision, and constitutes part of the information that is communicated to the injured party. These decision codes are then used to produce statistics regarding cleared offences, persons suspected of offences and offence-participations. Thus, a correct and uniform registration of the codes is essential for producing accurate statistics.

The study on the decision codes was motivated by an observation that the use of one of the codes had increased a lot over the years. This code, with the name *13 (Other)*, differs from the other codes in the system, since it doesn’t have a pre-defined motivation. Instead, the user can choose to add a free-text motivation to why the investigation is discontinued. It can therefore also be seen as a residual category to use when other pre-defined codes doesn’t apply.

The main purpose of the study was to get an understanding of how the decision-code system works in practice and how this influences the statistics and also to evaluate the possibility to produce statistics with a better quality based on a new decision-code classification.

Different types of analyses were made to get good understanding of how the system with decision-codes works in practice. A review of the most important reference literature was conducted to fully understand the intended field of application of the decision codes. Interviews with police officers were conducted in order to ascertain how they interpret the different decision codes and whether there is any disagreement among these interpretations. The Police's decision code number 13 (Other)was reviewed to establish the content of the free-text formulations. All-together 120 000 free texts were categorized with the aid of automatic coding and a random sample. Furthermore, a range of statistical data from the Council’s database has been examined in order to illustrate how often the different decision codes are used.

The results of the study revealed several deficiencies in the Police’s decision code system, as well as in the application of these decision codes. There was a general lack of instructions and the names of the codes turned out to be counter-intuitive and unclear. The interviews with police officers revealed competing views on how to utilize the decision-codes and the re-classification of free-text motivations connected to decision code 13 (Other) showed a large over-utilization of that code. It turned out that a majority of the free-text motivations could have been substituted by other codes.

Given the above-mentioned results, a number of issues were identified in the presentation of the statistics on cleared offences. To begin with, it turned out that there are problems distinguishing between offences that have been “cleared” and offences that have not. In part, this problem relates to the overutilization of free texts associated with Decision Code 13 (Other), which counts as cleared, as many of those could have been substituted with codes that doesn´t count as cleared. Another problem in the presentation of statistics relating to delimitation problems between the different codes was the level of detail. In the official tables, the various types of clearance are presented in separate columns, often one for each decision code. Given the extensive problems that exist with regard to distinguishing between the different decisions codes, this is inappropriate.

The final assessment of the quality of the statistics was that cleared offences have an insufficient quality for the intended field of application. Following this, we are now in the process of reviewing the structure of the statistics, with the aim of creating a more relevant categorization with higher quality. Further studies concerning the quality of the statistics on suspect persons and participation in criminal activities are also planned.

Apart from providing an assessment of the quality of some of the products in the Swedish crime statistics, based on the experiences from the quality studies that the Council has carried out, a couple of conclusions can be made concerning how to investigate quality of statistics based on data from administrative authorities, such as the Police. (For a description of the first quality study made by the Council, see [6] or [7]).

*\** Through in-depth studies, errors have been disclosed that were not possible to detect through the regular controls that the Council carries out in the production of the official crime statistics. By combining different types of analyses, issues on how, where and why errors occur could be explored. Particularly, interviews with registering personnel proved useful for the purpose on understanding why errors occur in the coding. \* Instructions and close collaboration with the data-providing authority does not seem to be sufficient for getting high quality in the raw data. As is shown in the studies by the Council, even when detailed instructions and long established collaborations exist, errors in the coding of key variables were found. This shows that it is not sufficient to rely only on formal arrangements, or to monitor the inflow of data. It could even be stated that a particular danger lays in the false sense of control that neat looking instructions and formal arrangements on a leadership level will instill.

Based on this, in the following chapter a simple model is presented on how issues concerning quality in the raw data of registry based statistics can be approached.

**3. Assessing quality in the raw data of registry-based statistics.**

For the statistical producer, as data collection for registry based-statistics often takes place at another institution, it is tempting to treat that institution as one unit. However, administrative authorities, such as the Police, are often large organizations with offices spread all over the country and divided in different hierarchical levels. Unsurprisingly, the perspectives on what significance different variables and registration procedures holds can vary between different parts of the organization.

Therefore, in order to assess the quality if the input data from administrative authorities, a model is needed that recognizes this. The simplest way to do this is to look at the administrative authority as consisting of two levels; management and operations.

Figure 1. The statistical producer and the administrative authority.

Management related indicators

Statistical authority

Administrative authority

Operations related indicators

On the management level the strategic planning, follow-up revisions and audits, it-strategy, formal contacts with other authorities as well as politicians and media, are taken care of. At the operations level, case officers are registering and handling applications, police reports, tax return and so on. This is where the requests and instructions from the management level are carried out.

After dividing the administrative authority in management and operational layer, quality indicators for the different levels can be assigned (Figure 1). The term quality indicator in this context denotes anything that can reveal potential errors in the raw data.

Quality indicators that belong to the management level are:
- The quality of the classifications that are utilized.
- The quality of instructions and other types of support to assist the registration procedure.
- The existence and quality of educations concerning registration procedures.
- The existence and depth of routines to inform and involve the statistical producer in changes that are planned in the registration process, i.e. change of classifications, change data collection mode, change of computer systems etc.
- The existence and quality of common work groups with the statistical producer to ensure that concepts and definitions are relevant and useful for both the administrative and the statistical purpose.

Quality indicators, and assessment methods, that belong to the operations level are:

- Variations between different units. An easy test is to divide the authority into different organizational units and study the relative frequency of different registrations. If a large variation exists, this will indicate that the registering personnel also interpret the definitions, codes or meaning of the registration in different manners.

- Variation over time. Large changes over time can indicate that the perception of what the registration signifies has changed. This is especially important to control for, when there is no logical explanation for the variation.

- Recoding of cases. If possible, select a random amount of cases, and redo the coding of central variables with the help of auxiliary information, such as free-text descriptions of the cases. In many instances, the free-text descriptions will hold more relevance for the registering personnel than the selection of a code in a classification.

- Interviews with registering personnel. If even a small selection of for example personnel working in different regions within the same authority holds different views on the meaning of codes and central variables, it is likely that differences will be wide-spread, thus making the statistics that are based on those variables hard to interpret.

The indicators at the management level are the easiest to control, but it is important to remember that they cannot by themselves reveal the quality of the data. For example, if there are no instructions on how to register some variables, it is likely that the registering personnel will make errors. However, this doesn’t mean per se that the registering personnel aren’t doing the registrations in a standardized manner. A consensus within the organization can still exist due to other reasons such as work-networks, a long time established practices and so on. On the contrary, instructions are by themselves not sufficient to ensure that those instructions actually are followed at all moments. High workload, unawareness of the utility of the registration, change of software and so on, can all explain deviation from compelling instructions and otherwise clear nomenclatures.

The second group of indicators relate to the level in an organization where the data are generated. The main point to keep in mind for this level is that for a variable to be suitable for statistical purposes it needs to fulfill one of the following two requirements; either it needs to hold significance by itself or every person registering the information need to understand it in the same manner. A variable holds significance in itself if its registration has direct legal or practical implications. For example, an attorney might make a decision to prosecute a suspect person. This decision could be right or wrong, but the fact that it has been taken is undisputable. If no such clear significance can be attributed to a variable, everyone that makes the registrations needs to interpret it in the same way for it to be useful in statistics. This might be easy when it comes to some basic information, such as gender and age. But when the level of abstraction increases, complications will follow.

These two sets of indicators and tests are ultimately about finding ways to assess whether the opinions and views of everyone involved in the production process is synchronized. It will not matter if the statistical authority is in agreement with the management of an administrative authority on the importance of a certain variable if the registering personnel don’t interpret it the same way. At the same time, if the statisticians misinterpret the meaning of the registration of some information, this can affect the final statistics equally bad.

It´s important to recognize that statisticians need to continuously make checks to ascertain that their understanding of the raw data is in line with the registering personnel. Otherwise discrepancies in the understanding of classifications and registered information can grow large, creating dangerous pitfalls for the interpretation of the statistics. Many statistical products are created at one point, with a particular set of technical and legal prerequisites, but over time, as the prerequisites change, the necessary updates in the statistical products are not always made. Therefore, as a last checkpoint, the statistical authorities should also plan for recurring in-depth studies to verify the basic assumptions regarding the significance of the input data.

**4. Discussion and conclusions**

Since the data collection process usually is not under the control of the statistical producer, when it comes to data generated in administrative processes, it is challenging to map and control for all the potential error sources. Often, the apparent solution for quality controls is to establish a close collaboration with the data-sending authority at a management level. However, this can give a false sense of security. Registering personnel can hold different views of the registration procedures than the management level. Furthermore, registering personnel in different parts of the country can differ in their interpretations. These types of discrepancies may have crucial impact on the final statistics.

In this paper, a simple model has been proposed where the statistical producer sets up data quality indicators dedicated for both management and operations. The essence of the model is to get a good understanding of what the data from administrative authorities represent, by thoroughly examining the data generating process. This broadens the perspective on quality assessments for register-based statistics to some extent. Despite this there are still more angles that could be explored.

While for surveys there is usually a lot of research before changing the data collection mode, this is usually not the case for data-generation in administrative registers. As an example, when the Police in Sweden opened up the possibility to file a report online, the data collected in this mode is expected to differ considerably from cases reported at the Police stations. Ideally statistical institutions would allocate equally many resources as for surveys to understand the impact of these types of changes.

To sum up, the data generated by the regular administrative processes in state institutions such as the Police, the Tax agency and the health care, are potential gold mines for official statistics. In order to fully deliver on the promise that these data holds, it is imperative that the producers of statistics direct considerably more efforts to exploring and understanding the error sources associated with this data.

**5. References**

[1] Laux, R, Baigorri, A and Radermacher,W. (2009), Building confidence in the use of administrative data for statistical purposes, 57th session of the international statistical institute- world statistical congress, web-publication.

[2] Zhang, L.C (2012). Topics of statistical theory for register-based statistics and data integration, Statistica Neerlandica Vol 66, nr 1, pp 41-63.

[3] Wallgren, A. and Wallgren, B. (2014) Register-Based Statistics: Statistical Methods for Administrative Data, Second Edition, John Wiley & Sons, Ltd, Chichester, UK.

[4] Bakker, B.F.M and Daas, P.J.H, (2012), Methodological challenges of register-based statistics research, Statistica Neerlandica Vol 66, nr 1, pp 2-7

[5] Daas. P, Ossen. S, Vis-Visschers, R and Arendt-Toth, J. (2009). Checklist for quality evalutation of administrative data sources, Discussion paper, The Hague, Statistics Netherlands.

[6] Färnström, A. (2013) The policeman and the statistician. Paper presented to the
59th ISI World Statistics Congress, Hong Kong, Session IPS071.

[7] The Swedish National Council for Crime Prevention (2012), Användningen av brottskoder: En kvalitetsstudie inom kriminalstatistiken (Translation: The utilization of crime codes: A quality study within the crime statistics), Stockholm.