Proceedings of Statistics Canada Symposium 2018
Combine to Conquer: Innovations in the Use of Multiple Sources of Data

Data integration method: A consolidation of semantic heterogeneity and data sources with the England and Wales custodial policy evaluation project

Marie-Ève Bédard[1][2]

Abstract

This research will highlight the methodological issues that arose from the use of multiple data source for the research project done on the 2014 England and Wales evaluation of safety policies performances in custody. The project used several data sources, such as administrative data, survey data, key performance indicators, and Prison Quality Model data gathered by researchers. A method was developed to consolidate these data sources and their semantic heterogeneity for evaluation purpose, by means of residual change score analysis, principle component factor analysis, robust standard error regression and scales. With this type of data integration method, although successful, the data analysis remained with several limitations, which leaves room for further research into finding a way to close the gap between these data sources.

Key words:

1. Introduction

This research will highlight the methodological issues that arose from the use of multiple data source in a research project covering 51 prisons in England and Wales. This particular research is an evaluation of safety policies and performances in custody. The project used several data sources, such as administrative data, survey data, key performance indicators, and Prison Quality Model data gathered by researchers. Since the concept measurements surrounding the data were taken from different sources and some data were not collected on both time points, there is a lack of certitude on cause and effect of the policy. A method was developed to consolidate these data sources and their semantic heterogeneity for evaluation purpose, by means of residual and raw change score analysis, principle component factor analysis, robust standard error regression and scales. These methods have been combined with other methods, such as Cook's distance, y computation and variance inflation factor test in order to unify these data and make them comparable for this kind of evaluation. With this type of data integration method, although successful, the data analysis remained with several limitations, which will be explored in this paper.

2. Framework of Analysis and Data

Concepts and framework of analysis for this method is based on the evaluation of an administrative implementation of a policy concerned with an increase of safety of the prisoners in England and Wales, following by the implementation on site of this policy. The concept of Safety has been defined by the ministry of Justice of the United Kingdom and is composed of the rates of death, self-harm behaviour and assaults. The policy chosen is the PSI 64/2011 (safer custody) and it was effective on April 1st 2012 until January 31st 2016. For the “on site” level of the evaluation of the Safety policy, the answers from the prisoners extracted from the Inspectorate reports by the HM Chief Inspector of Prisons has been used to determine if prisoners have received services mandated in the above mentioned policy. The Inspectorate reports are distributed every year to prisoners by the Her Majesty’s Inspectorate office, therefore the 2007-2011 and the 2012-2013 were used to measure changes before and after implementation.

2.1 Phase 1

[1]Marie-Ève Bédard, Statistics Canada, 100 Tunney’s Pasture, Ottawa, Ontario, Canada, K1A0T6
[2] The research was conducted in independently at the Freie Universitaet Berlin, Germany in 2013-2014.

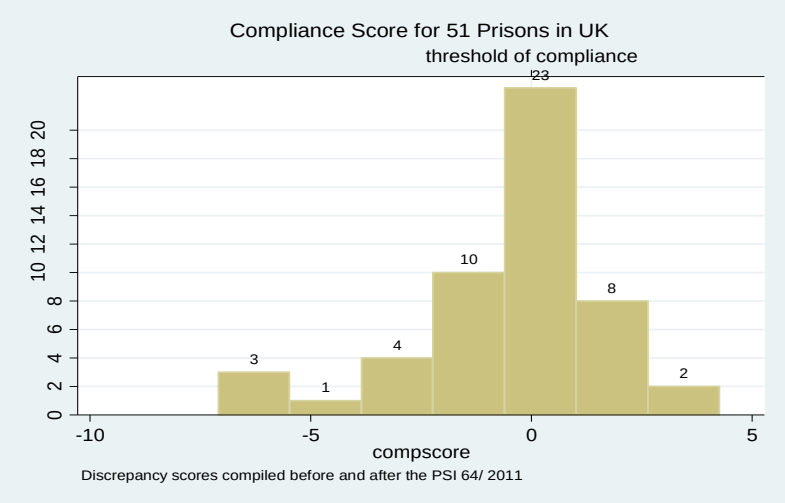
Before measuring any policy changes, each establishment had to be measured in a compliance index, which mean that the first phase had to reflect if the policy had been implemented or not. It was achieved by extracting questions from the Inspectorate, for the compliance index that were reflecting three dimensions of safety represented in the Prison Quality Model of Liebling, 2011. This particular model is aiming at measuring the moral dimension of the quality of prisons based on the concept of Performance and the links between managerialism and moral values. Each dimension has a psychological and physical level that represents Order, Safety and Well-being.[3] A matching process with the help of Principal Component Analysis was made to determine which question from the Inspectorate was matching the Safety Policy. With this matching process, it was possible to determine if the policy has been implemented or not. Below, one can see an example of this matching process.

PSI 64/2011
Prisoners have timely access to Listeners wherever located
MOJ,2013,p.23

HM Inspectorate
Can you speak to a Listener at any time if you want to? Yes or No
HM,Blundeston,2013,p.89,Q-4.7

In order to establish the Goodness of fit, several tests were conducted; such as Principal component analysis, Scatter plot, R-squared and F test, Two-tail P value, 95% Confidence Intervals, Breusch-Pagan/ Cook-Weisberg test of heteroscedasticity, test of multicollinearity, Y hat computation, Cook's distance, Kernel density and Linear regression and Average Marginal Effects (95% Cls). Afterwards, dimensions were built by generating a discrepancy times the factor loading, for each of the 6 dimensions.[4] The results of the 51 selected prisons can be seen in **Figure 2.1-1**.[5]

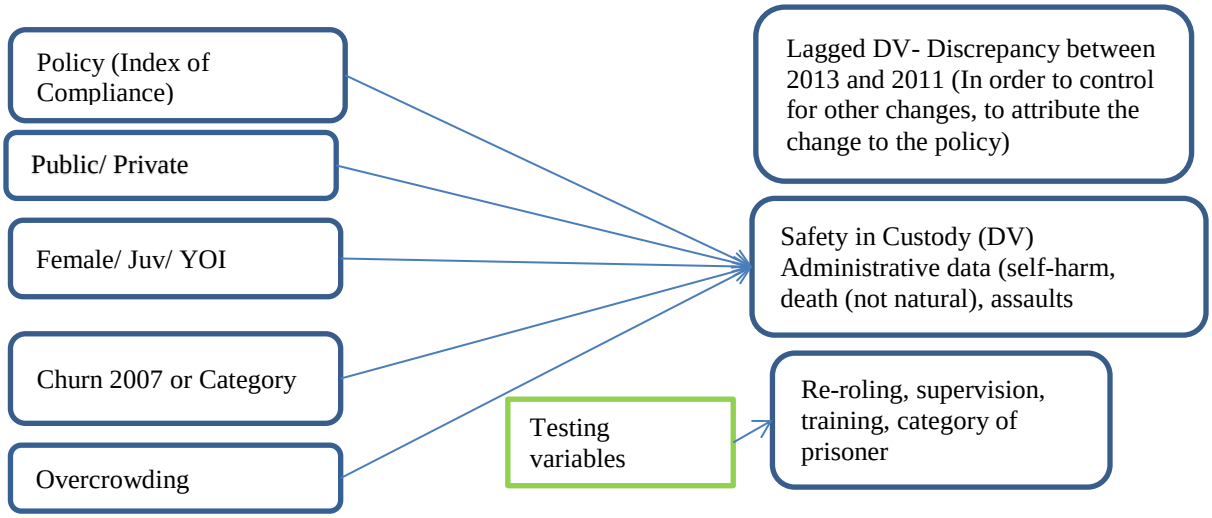
Figure 2.1-1
Compliance score index of prisons



[3] These three dimensions have been found to be composing what Safety is in a prison. The concept is presented in details in Professor A. Liebling's extensive research. Liebling, A. (2010) Identifying, Measuring and Establishing the Significance of Prison Moral Climates.
[4] The dimensions based on the Prison Quality Model of A. Liebling were as follow: Psychological; health, wellbeing, trust. Physical; health, adaptation, stability
[5] 28 establishments have done worse from the inspectorate wave 1 to inspectorate wave 2. The meaning of these results is that a plus sign is an improvement and a minus one is a worsening. It can be resume that a negative number indicates a non-real compliance to the policy and that a positive one to a real compliance to the policy. It simply points to services provided to the inmates that each establishment was compelled to offer or to implement.

2.2 Phase 2

In the second phase the evaluation was limited to the services provided to detainees and the fluctuation of the rates of self-harm behavior, suicide and assaults. In this phase the safety scores will be evaluated based on a Raw Change Score Analysis and a Residualized Change Score Analysis. In this section the compliance score from phase 1 became an independent variable for phase 2, along with other independent variables, such as churn, public/private, female and others.



Two models were chosen; raw change score analysis ($\boldsymbol{D\tau = Y_1 - Y_2}$) and residualized change score analysis ($\boldsymbol{R = (Y_j - \bar{Y}) - B_{x,y}(X_j - \bar{X})}$). This choice of method is based on the qualities and the limitations of both formula. The model of the change score formula is built on the subtraction of Τ1 onto Τ2 producing a difference score that will be used as the dependent variable and will represent the change between safety in 2011 and safety in 2013. The model allows a vector of change.

The residualized score method from Borhnsted is representing the simple difference between the Τ2 score and the estimated Τ2 score, which is by definition uncorrelated with Τ1. The residual will be obtained by removing from Τ2 scores the portion that can be linearly predicted by Τ1 (Dalecki and Willits, 1991). The method used for obtaining such score is a simple linear regression with Τ1 as an explanatory variable, the end result will be a residualized score that has already accounted for Τ1 (Dalecki and Willits, 1991). The same method is reproduced for each other explanatory variables, as per recommended by Bohrnstedt (Bohrnstedt, 1969 cited in Dalecki and Willits, 1991).

In the second model a raw change score was generated by subtracting Τ2 from Τ1. It is known and widely practiced that Τ1 gets subtracted from Τ2, but in this particular case, because of the nature of the data it had to be generated in reverse. Since a high number represents low safety, it was more convenient for the analysis and the understanding of the change that this formula is reversed, this way it is now possible to read the coefficient in the same logical direction as depicted. It means that a positive coefficient indicates a better of the score, in other words that the rate of death, self-harm and assaults have reduced from Τ1 to Τ2. In order to control for Τ1 the covariate s1[6] was introduced, thus eliminating one of the controversial point against raw change score's models. The second model was mainly introduced in order to compensate for the abstract meaning of the residual coefficients.

3. Methodology limitations

3.1 The policy

[6] S1 refers to safety scores in 2011.

Since no methodology is perfect, neither are data; there are always concerns about fallacy, data errors and impreciseness of results. The first limitation that needs to be addressed is the policy itself. The PSI64/2011 is a policy that affects safety on a general term and not on a specific level, such as clear guidelines to deal with women issues or youth issues, therefore the factor variables of female and youth stays predominant on the influence of safety. What it means is that if some services had been clearly defined by the policy to target only female or young offenders (and juvenile) it would have been possible to extract some specific questions in the questionnaire of the Inspectorate and target these specific policies.

3.2 The Inspectorate

The second limitation this research is accounting for is related to the questionnaires themselves (Inspectorate). The questions tackling the policy were taken from another concept measurement, mainly the general evaluation of the prison, which was largely based on the quality of prison model from Liebling. Even if the factor analysis was congruent, maybe other policy guidelines (from the same policy) were better implemented on the service level such as family relations; these guidelines were not represented in the concept due to the lack of data. For example, in the policy they do point to try to better relationships between the inmate and its family, as this could bring more information on the prisoner's background and state of mind as well as having positive impact on the inmate (MOJ, 2013). This precise guideline could have been represented by the question: "Have staff supported you and helped you to maintain contact with your family/friends while in this prison? Yes or No" (HM Holloway, 2013). The problem that occurred is that this question was not present for the first Inspection (T1), thus had to be immediately rejected as a T1 score and would not have been possible to gather.

3.3 Internal Management

Another problem that arises with the policy is that broad sections of the policy deal with the internal management of safety, and those sections could not be represented on the service level; therefore if the prison spent more time and energy rearranging management, the service level will not be concluding. In order to represent this issue, one can refer to the policy mandatory actions that are emitted in the last 30 pages of the policy. In the last sections of the SC policy, one can find guidelines on how to deal with food refusal, staff ending contact with the family, returning property to family after death, funeral, prisoner or staff witness, inquest or 93 palliative care for prisoners with a terminal or serious illness and many more. The problem is that all of these guidelines cannot be quantified as well as cannot be public, so if the staff and the administration have concentrated their efforts on these guidelines, the service level will ultimately suffer in the questionnaire.

3.4 Data Collection

The fourth limitation of this study is concerned with the data collection. In fact, some data were collected in a very short period of time after the implementation ("on site"), it is possible that the effect was not felt yet, and because of the small N of the study, it has reduced the effect considerably. Since the whole sample consists of the total amount of prisons in England and Wales and that comparisons between countries is impossible based on the very wide diversity of mode of management, characteristics and the uniqueness of the SC policy (mandatory actions targeted only to England and Wales), it is important to mention that no comparison group was introduced in this research in order to help establishing causal effect.

The next methodology concern is also related to the data. The compliance score generated in phase one could also be biased, since the data chosen are subjective and the population from the first inspectorate and the one of the second inspectorate is not the same, as prisons have a high turnover of prisoners especially when it comes to local institutions, because they usually detain convicts as well as people waiting for trials for a very short period of time. Additionally the population sample chosen was done by the prison rather than the researcher, which could imply, despite their great effort to have a representative sample of the prison, some fallacies along the process. As a result of these possible weaknesses of the data, the compliance score index that was generated might not be an absolute mirror of the factual services provided to the inmates in the selected prisons.

3.4 Sample Size

Another concern that should be mentioned is that the number of variables that was included in this research could also have been too high for such a small N sample. The great amount of explanatory variables, which are trying to explain one phenomena (compliance in phase one and safety score in phase two in only 51 establishments (N)) can lead to less reliable results. The removal or the addition of one of these variables changes the results one will obtain tremendously (Zucknick and Richardson, 2014). This is a problem that is occurring when a small number of the sample is linked to a particular relation (Zucknick and Richardson, 2014). At last, one problem that made the gathering of the data very difficult was to find the accurate orientation of the prison. The orientation is meant by finding out if a prison was female or male or youth offender institutions, training, local, the category of prisoners in it, and so on. The problem is that many sources are unreliable due to a lack of care (the sites were not updated) or simply the lack of information.

4. Policy evaluation and future model modifications

As it was shown, many questions were raised or could not be answered in a perfect way (some not at all), but there are modifications that could be brought to this model in order to account for its flaws, or attempt to get better or more reliable results. One of the first possible changes would lie in the questionnaires themselves. It would be a great improvement if the questionnaires could be built especially with the aim of answering this research question. In other words, it would be much more accurate if questionnaires were generated with special questions targeting the exact outline of this policy[7] and the services rendered to the inmates, while simultaneously the same questionnaire would be distributed to the staff asking if they have rendered these services; this way it would be possible to assess the exact gap between the rendering and the receiving of these mandatory services. This process would surely involve costs and would have to be reproduced on a regular basis (once every two years for example) in order to be able to analyze the fluctuation as well as providing real complete information about the "on site" effectiveness of these policies. The second change that could be brought to the model would be to find a way to really deepen the analysis of the churn factor,[8] as this variable was only represented through the local variable, based on reverse causality issues, and could not even be included in the second phase due to very high correlation between female and youth with local prisons.[9] The last modification is based on the possible replicability of this research. It would be good to point out that the replicability of this research would generate much more reliable results. There are several reasons for this, the first being that the questionnaires have been standardized from 2012 on. The standardization of the questionnaires would allow retaining many more questions that would better represent the same policy or another one. A second aspect of replicating the study is that because the questionnaires are standardized, many more prisons will be retained for a new analysis, blowing the sample size, resulting in more reliable results. The only limit that imposes itself is the end of the PSI for safer custody. This policy was effective until 2016, which mean is no longer mandatory for the prisons to physically implement these actions after this period of time. Nonetheless, a second review of this model would give a better perspective on the accuracy of these findings.

At last, this methodology, although applied to policy in prisons, and even with its limitations, is a good framework for a more fair representation of implementation of policies. The amalgamation and use of survey data and administrative data in a complex environment, such as prisons, gives quantifiable results that is taking into account several behavioral and changeable variables and is providing a bigger picture than simply administrative data. This methodology also provides policy makers better evaluation tools that is considering more than just one facet of an issue. It would be worth it to keep refining such tool for future use and also test its robustness in another kind of policy implementation.

---

[7] Of course this could be done for any other policy.

[8] The churn of an establishment, as defined by the MOJ, has the risk factor existing in a prison to establish the likelihood of suicide, self-harm and assault in an establishment during the early stages of custody (MOJ, 2013). This churn score can be calculated on the individual and the establishment level. On the individual level, it is calculated by the length of time they have spent in custody, and for the establishment churn it follows as such: Establishment churn= (First Receptions+ Transfers in from other prisons / Populations) (MOJ, 2013).

[9] The coefficients became irrelevant when this variable was included as the correlation between female youth, local and churn was too high, so local and churn had to be removed in order to have readable results.

References

Dalecki, M., and F. K. Willits (1991), "Examining Change Using Regression Analysis: Three Approaches Compared", *Sociological Spectrum*, 11:2, pp. 127-145.

HM Chief Inspector of Prisons (2013), *Report on an announced inspection of: HMP Blundeston*, London: Her Majesty's Inspectorate of Prisons.

HM Chief Inspector of Prisons (2008), *Report on an unannounced full follow-up inspection of: HMP/YOI Holloway*, London: Her Majesty's Inspectorate of Prisons.

Liebling, A., and H. Arnold (2004), *Prisons and their Moral Performance: A Study of Values, Quality, and Prison Life*, Oxford: Oxford University Press.

Liebling, A. (2010), *Identifying, Measuring and Establishing the Significance of Prison Moral Climates*.

Ministry of Justice (2013), *Safety in Custody Statistics England and Wales Update to March 2013.*

Ministry of Justice National Offender Management Service (2013), *Management of Prisoners at Risk of Harm to Self, to Others and from Others (Safer Custody).*

Zucknick, M., and A. Richardson (2014), *Algorithms for Bayesian variable selection in the logistic regression model for large-scale genomic applications*, Ithaca: Cornell University Library.