Privacy protection, data validation and secure machine learning with new statistical methods while preserving transparency

Violeta Calian[1]

[1]Statistics Iceland, Iceland

Abstract

The goal of the present study is to show that the standard methodology of official statistics may be enriched and its quality may be improved by applying the most up-to-date developments in theoretical and applied statistics. The progress of computational tools and resources in recent years make it possible to apply, in a routinely and timely manner, advanced tests, models and algorithms which have been previously restricted to academic or industrial purposes.

We focus on several stages of the statistical production and illustrate these ideas with concrete applications at Statistics Iceland regarding:

(i) Bayesian rule extraction for data validation processes. Data application: comparing domain expert knowledge-based rules and results of rule mining.

(ii) protecting (statistical disclosure control of) tabular data using Bayesian modelling and cryptography inspired methods. Data application: publishing detailed census grid data.

(iii) quantifying uncertainty of machine learning classification algorithms while adding interpretability features for transparent and complete communication of results. Data application: random forests and neural networks for demography applications.

One of the main conditions for employing any of these methods is to be able to quantify their performance and to report on the uncertainty in results associated to: data variability, model fit/complexity, distributional differences between training and predicting data, measurement/register errors or interactions of all these types of errors. We show that these tasks can be achieved for the most complex models and exemplify with results for the cases described above.

Keywords: data validation, statistical disclosure control, machine learning

1. Introduction

In this paper, we argue that a critical measure of quality in official statistics has several key dimensions: employing accurate data, producing reliable and privacy preserving statistics, while using advanced, up-to-date methods *in a scientific way*. This requires thorough evaluation of the performance of these methods and of the uncertainty of their results, as a main step whose importance is increasing with their complexity. The more advanced the models or machine/deep learners are, the more rigorous the evaluation, testing, calibration and uncertainty measurements should become.

The general principles driving our choice and implementation of methods are outlined in section 2. These are exemplified by concrete problems and proposed new solutions concerning data validation (section 2.1.), statistical disclosure control (SDC), discussed in section 2.2. and detailed case studies (section 2.3) where we recommend types of reporting the output uncertainty. This last step is in our opinion essential. While adding useful information to output and maintaining transparency in a rigorous way, it is often overlooked in literature.

Interpretability and assumption testing regarding data or models play important roles and deserve to be addressed systematically as well. We use/produce open-source implementations of these steps.

2. Methods and illustrative examples

The methodology we propose follows in fact the standard mathematical statistics approach to model fitting, model selection, out of sample prediction and uncertainty evaluation: we show that this general, scientific framework can be adapted and applied for the special cases of new data validation, disclosure control and inference based on complex models/algorithms.
Remarkably, all solutions have several standard stages:

(i) exploring and describing the data (analysing distributions, correlations, clustering, outliers)

(ii) training a set of algorithms / fitting complex models, measuring their performance according to well defined metrics and identifying their optimum regimes based on the prediction goals (e.g. classification for census, survey optimization, or forecasting with hierarchical Bayesian models)

(iii) quantifying and reporting the uncertainty associated with the predictions, including:

- the uncertainty associated with the variability in the training data.
- the uncertainty due to the model fit and model complexity issues.
- errors due to distributional differences between training and predicting data sets.
- measurement errors
- errors due to interactions between model and data-dominated uncertainty.

(iv) describing the results in simple terms, by using interpretability tools which allow the user to understand the relations between the predictions and the features/variables involved in the AI/ML algorithm or model. This stage usually involves for instance measures of feature importance, surrogate models in the case of complex classifiers or conditional effects and posterior distribution checks as in the case of hierarchical Bayesian models.

2.1. New data validation methods

We have analysed in a recent project the performance of the data validation approach based on confronting a data set with a set of rules (constraints) usually defined according to general requirements concerning data structures and expert knowledge. This is usually followed by identifying the fields in the data which need to be modified/imputed. For this purpose, we have employed well-known R-packages (see (van der Loo, 2021) and (van der Loo, 2024)) with good performance and studied (Gislason, 2019) the dependence of running time of the error identification as a function of the number of errors per record, rule category, sub-categories of data.

This classical approach is very systematic, it has a clear interpretability although the two-step structure (which includes error location and error correction) is often difficult to solve in a general way and may requires additional transformations of the problem.

New solutions are emerging and show promising results, though. They are based on the premise that most of the records and fields of a data set are valid and only a minority is anomalous. Consequently, we can fit a model or train an algorithm using the whole body of data in order to identify association rules and/or errors and to infer needed field values accordingly.

This implies that two well-known whole classes of statistical methods may be employed for solving this very problem:

(i) ML based discovery of association rules. We are employing *Apriori* and *eclat* algorithms as typical examples (see e.g. (Agrawal, 1993) and (Hahsler, 2005)) with the goal of rule mining and selecting new and meaningful rule sets. Testing these sets against the traditional rule repositories is essential and ongoing.

(ii) Bayesian modelling (previously part of the imputation stage only) may be used for simultaneous detection and correction of errors, as shown in (Kim, 2015) for continuous microdata and in (Manrique-Vallier, 2017) for categorical data.

The features and advantages of a Bayesian approach suggest that one could even add privacy protection to this process, at the same time. This step has not yet been completed, according to our knowledge. At the same time, mining validation rules can exploit Bayesian methods (see e.g. (Tian, 2013)) and select the optimum set as defined by support, confidence and lift. Updating and testing the available (R, open source) software for this purpose is still ongoing. The prospects are optimistic since in recent years methods as the ones mentioned above have become a realistic choice in practice in terms of running time, in addition to their theoretical advantages, due to faster computational resources and easier to optimise software.

2.2. New SDC methods

We describe in this section a typical case study, i.e. the publication of 2021-Census data, for the purpose of simplifying the discussion. When regarded through the Census-grid(s), Iceland looks like a "virtual archipelago", i.e. a large set of disconnected populated cells, many of them with a very small number of inhabitants, separated by large unpopulated areas. This means that even aggregated data for such cell-systems pose a high disclosure risk when published.

The main condition we formulate for any valid data protection method is that it should preserve the relevant distributions over regional (and more, e.g. urban/rural) divisions. An additional condition has been imposed, namely producing "credible results". This has been translated into mathematical terms by the requirement to preserve the (approximate) location of certain outlier type of characteristics/groups which are public knowledge.

We tested several procedures for disclosure control of aggregated data and evaluated their best performance, for the case of our Census grid data. The advantage of the selected and newly proposed method, which is based on swapping of data between grid-cells (and not only individuals or households) is that it could be automatically and straightforwardly applied to any tabular dataset and that it preserves consistency at chosen levels of aggregation. The open code and results are reported/linked on the repository[1] of open code which already includes the information concerning our project on SDC for the small output area system and preliminary code for the new method evaluation.

As verified in a previous study (Calian, 2020), we can confirm that the most critical stages in applying and evaluating an SDC method are: the identification of risk variables and the risk-utility analysis. The former is a rather subjective process which is based on legal, cultural and information types of conditions (Hunderpool, 2012). The latter is the object of an interesting statistical problem, i.e. evaluating the effect of multivariate transformations (as implicitly defined by all standard methods) on multivariate data distributions. Measures for both risk and utility (standard, information based) are used to define the parameters of the optimum regime of the employed SDC method.

[1] https://github.com/violetacln/testingSDCtools

The results were compared with the ones obtained by using the SDC methods in the standard way. In the previous pilot-study we have already tested the recommended methods for the newly built system of small output areas albeit not for the grid-system, by using data of the 2011 Icelandic Census. Evaluation and testing of the newly proposed solution consist of the following standard steps: (i) identifying the cells with risk of attribute disclosure, identification, and differencing risk (ii) applying the SDC method based on cell-ID swapping, for multiple values of the critical parameters (iii) evaluating the residual risk and information loss on the output dataset (iv) comparing the results with the record swapping and cell-key methods of data protection. The implementation of the testing, evaluation and risk-utility measurement steps makes use of several R-packages from the ensemble of SDC-tools[2].

However, new directions for protecting data are currently under evaluation:

(i) using Bayesian estimates for generating synthetic data.

(ii) using deep-learning and/or cryptography inspired methods such as adversarial neural networks.

(iii) using differential privacy and its Bayesian variant which can guard against difficult scenarios built on deep learning.

The probability of an intruder correctly identifying an individual or individual attributes by using a released data set is at the core of risk definition. A Bayesian approach may be used to calculate predictive probabilities and disclosure risk under model uncertainty (with e.g. model averaging) while using joint data distributions (as initiated for instance in (Forster, 2005) since particularly suitable to this kind of reasoning.

We emphasize here the need of making decisions for official statistics regarding the possibility of future use of differential privacy which (citing from (Dwork, 2014).): “describes a promise, made by a data holder, or curator, to a data subject: ‘You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources, are available.’ At their best, differentially private database mechanisms can make confidential data widely available for accurate data analysis, without resorting to data clean rooms, data usage agreements, data protection plans, or restricted views.” One should also keep in mind the corollary to a well-known information law which states that too many accurate answers to too many questions will destroy privacy protection (as shown in the same reference) and proceed accordingly. Limitation, as opposed to elimination of disclosure, thus preserving utility and accuracy, by using differential privacy is

[2] https://github.com/sdcTools

the pragmatic goal for official, health, technological institutions which use or even stream data and open source tools and libraries (such as Google, Apple, US Census Bureau[3]).

2.3. Uncertainty and performance evaluation for complex models/algorithms

Using novel and complex models or even publishing classical statistical estimates should fulfil the conditions listed at the beginning of this section: verifying data and model assumptions, evaluating the results' uncertainty due to the data variability and data errors, as well as to the uncertainty in model fit and choice and interaction of all these factors.

Publishing this type of results routinely, as opposed to point estimates and unique values of performance measures is a best practice which becomes even more important when using advanced tools and models/learners.

We exemplify this best practice in what follows. Statistics Iceland decided to estimate the true resident population at any point in time, based on combining multiple register data and formulating the overcounting due to lack of de-registration issue as a binary standard classification problem (Calian, 2023 and associated R-code repository[4]). For this purpose, multiple machine learning algorithms were trained on data with known outcome and their performance was evaluated according to many standard measures as shown in detail in the paper cited above. The preliminary exploratory and data mining stages were used for validating the set of attributes included in the models.

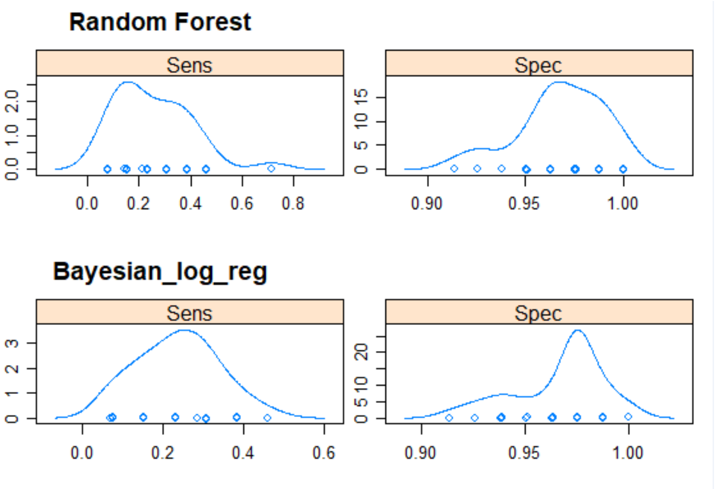
The results of the evaluation of algorithm performance, tuning and optimisation (i.e. choosing critical parameters as a function of the analysis goal) as well as the results of the predictions were communicated *together with their uncertainty and variability measures*.

The distributions of all standard performance metrics were obtained by resampling (k-fold cross-validation). We show in Figure 1 an example of two classifiers and distributions of two such metrics. More complete results are included in the paper. To the usual set of metrics, i.e. Sensitivity (true positive rate, TPR), Specificity (true negative rate, TNR), accuracy (proportion of cases correctly classified, out of the total number of cases), the harmonic mean F1 of sensitivity and specificity, the Youden's J statistics or the Kappa statistics (especially for comparing classifiers and using random chance as a baseline), we added a constraint reflecting the maximum admissible total population error.

[3] https://neptune.ai/blog/using-differential-privacy-to-build-secure-models-tools-methods-best-practices

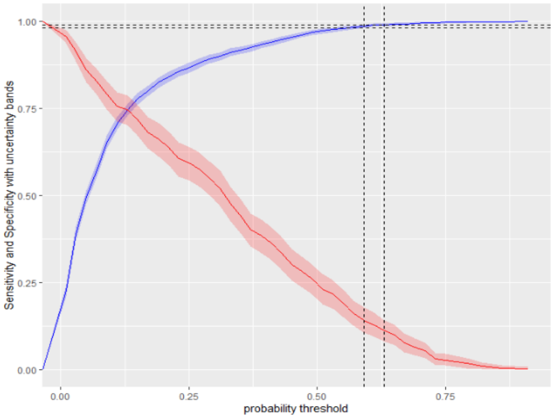
[4] https://github.com/violetacln/SLOPA

Figure 1: Uncertainty of typical performance measures for two classifiers (random forest and Bayesian logistic regression) reflecting training data and model variability



This inherent variability is manifest in the tuning and optimisation stages for any of the ML classifiers. For instance, the one we chose, i.e. random forest, has a classification probability threshold (critical value) to be selected for providing best results. Figure 2 shows the performance uncertainty and how much it varies when tunning the threshold values.

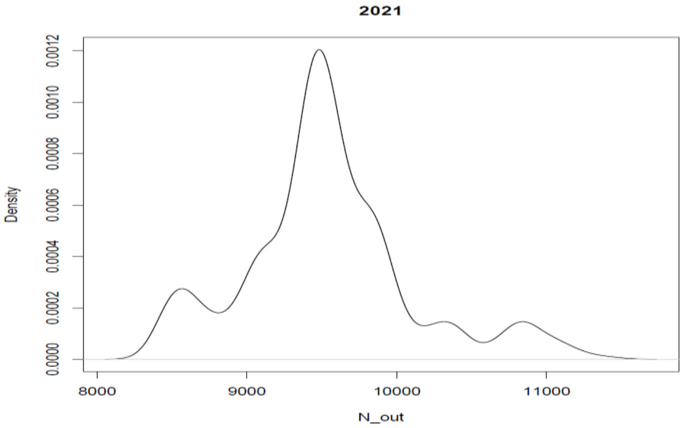
Figure 2: Confidence bands of performance measures of a random forest classifier reflecting training data and model variability, as functions of critical threshold values



The consequence of correctly measuring the uncertainty associated with the model/algorithm is reflected in the confidence/credible intervals of the predicted outcome, i.e. the predicted number of individuals with zero/one residence status. A cautionary remark is in order: the model uncertainty is not reflected at this stage but only its stochastic variability, which is captured by the resampling techniques. Model averaging could reflect it though.

We exemplify the uncertainty in the predicted outcome for the case of applying another classifier to the same problem as above, namely one based on expert-knowledge decision-trees, as proposed by the demography experts at Statistics Iceland. They identified the set of optimum critical values for the rules included into their algorithm and Figure 3 shows the distribution of the predicted outcome due exclusively to data variability, conditional on this set of optimum values.

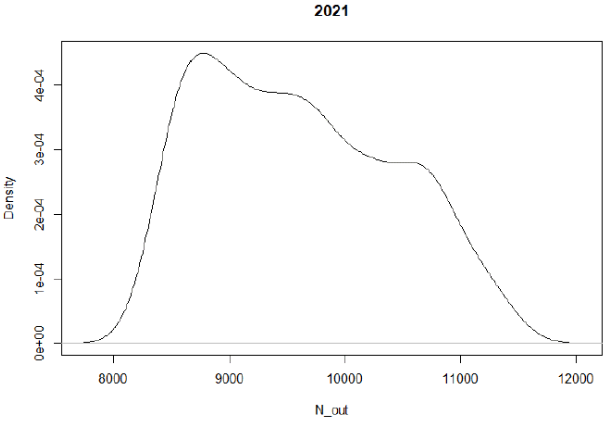
Figure 3: Distribution of classification results reflecting data variability for optimum regime



-

However, if the critical values included in their proposed decision tree were not fixed, the full variability of the model outcome is shown by a distribution like the one in Figure 4, i.e. much broader confidence interval and a much more uncertain result. This type of result can be used for selecting on the best regime of their algorithm.

Figure 4: Distribution of classification results reflecting both data and model variability



Yet an additional option for solving the same problem is offered by hierarchical models and we are testing this approach at the present time. It has the great advantage that it can make use of the whole population register data for training instead of the rather noisy and not very big survey data used for the ML algorithms. It can also capture well the data correlations and interactions. In the case of Iceland, where small numbers are frequently recorded, a good choice for solving such problems consists of using hierarchical Bayesian models with Gaussian process priors (for unknown nonlinear, smooth functions) which can be learned from data as we have already tested for forecasting purposes in (Calian 2023b and open-code repository[5]).

3. Conclusions

We showed in this paper that statistical products based on new data science technologies can be treated according to robust and transparent methods for measuring, controlling and reporting uncertainty while optimising for performance. The only limitations to such a process may arise from insufficient computational resources, input data or incomplete domain/interpretation knowledge.

References

MPJ van der Loo and E de Jonge (2021). Data Validation Infrastructure for R. *Journal of Statistical Software*, 97(10).
MPJ van der Loo (2024) *The Data Validation Cookbook* version 1.1.5. https://data-cleaning.github.io/validate

Agrawal, R., Imielinski, T. and Swami, A.. (1993). "Mining Association Rules Between Sets of Items in Large Databases." In *Proceedings of the 1993 Acm Sigmod International Conference on Management of Data*, 207–16. Washington, D.C., United States: ACM Press.

Hahsler, M., Grün, B. and Hornik, K. (2005). "Arules – A Computational Environment for Mining Association Rules and Frequent Item Sets." *Journal of Statistical Software* 14 (15): 1–25.

H. J. Kim, L. H. Cox, A. F. Karr, J. P. Reiter and Q.Wang. (2015). Simultaneous Edit-Imputation for Continuous Microdata. Journal of the American Statistical Association 110:511, 987-999.

Manrique-Vallier, D., & Reiter, J. P. (2017). Bayesian Simultaneous Edit and Imputation for Multivariate Categorical Data. *Journal of the American Statistical Association*, *112*(520), 1708–1719. https://doi.org/10.1080/01621459.2016.1231612

Gislason etall. (2019). Error detection for the statistics of external trade in goods. Nordic Statistical meeting, Helsinki, https://stat.fi/media/uploads/ajk_en/Events/nsm2019/gislason_-_error_detection_for_the_statistics_of_external_trade_in_goods.docx

Tian, D. et al. (2013). A Bayesian Association Rule Mining Algorithm,. 2013 IEEE International Conference on Systems, Man, and Cybernetics, Manchester, UK, 2013, pp. 3258-3264, doi: 10.1109/SMC.2013.555.

Calian, V. (2020). Methods of statistical disclosure control for aggregate data. With a case study on the new Icelandic geospatial system of statistical output areas. Statistical series: Working papers, 105(6), 2 September 2020, https://hagstofan.s3.amazonaws.com/media/public/2020/e9ea7160-5032-4580-9297-7b3b3cb634da.pdf

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, S., and de Wolf, P. (2012). Statistical Disclosure Control. Wiley.

Forster, J.J. (2005). Bayesian methods for disclosure risk assessment. Joint UNECE/Eurostat work session on statistical data confidentiality.
https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2005/wp.12.e.pdf

[5] https://github.com/violetacln/SIPP

Dwork, C., Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends R in Theoretical Computer Science Vol. 9, Nos. 3–4, 2014. https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

Calian, V., Harðarsson, Ó. and Zuppardo, M. (2023) Machine learning *estimation* of the resident population. Statistical Journal of the IAOS, vol. 39, no. 4, pp. 947-960. https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji230090

Calian, V. (2023b). Methodology of *population projections* based on hierarchical Bayesian models. Statistical series: Working paper, 108(4). https://hagstofas3bucket.hagstofa.is/hagstofan/media/public/2023/79a217c5-f567-4ddb-bed7-45329a32d531.pdf