Deep learning for transparent production of official statistics

Violeta Calian, Statistics Iceland, violeta.calian@hagstofa.is

Abstract.

In this paper we show that employing machine/deep learning algorithms can be made practical, transparent and accountable by measuring, tuning and reporting the uncertainty of their results and of their performance metrics. To achieve this goal, one may use methods which are either analytical and therefore not computationally but only theoretically demanding, data-driven thus rather widely applicable or methods based on the Bayesian paradigm (such as the case of Bayesian Deep Learning, which quantifies uncertainty while integrating prior beliefs and new evidence) thus reporting on uncertainty by construction. We illustrate this principle with several concrete applications which are implemented in open code at Statistics Iceland.

Keywords: deep learning, validation, disclosure control, Bayesian forecasting methods

1. Introduction

For efficient and high-quality official statistics production, the adoption of novel computational tools and techniques, such as deep learning and machine learning (ML), has emerged as a valid strategy. National Statistical Institutes (NSIs) are increasingly exploring these algorithms to automate processes, handle larger datasets, and extract complex patterns. However, the integration of ML into official statistics is not without significant hurdles. Traditional black-box approaches are strictly not recommendable in an environment where trust, reproducibility, and accuracy are most important. Errors, biases, and algorithmic limitations must always be described and rigorously controlled.

While the interpretability of machine learning results has been more systematically reported and demanded in recent years, the uncertainty surrounding the performance of these algorithms, especially their output results, is far less frequently measured and declared. In the realm of official statistics, where decisions impacting economies and societies are made based on published data, practitioners must avoid overly confident predictions and pointwise evaluations.

In this paper, we demonstrate that employing deep and machine learning algorithms can be made practical, transparent, and accountable. This is achieved by systematically measuring, tuning, and reporting the uncertainty of their results and performance metrics.

2. Methodologies for Uncertainty Quantification

To transition from point predictions to accountable predictive distributions, statistical agencies can employ various methods to quantify uncertainty in machine learning pipelines. These methods generally fall into three categories:

2.1. Analytical Methods

Analytical methods are theoretically demanding but computationally efficient once established. A prime example is the mathematical equivalence between certain neural networks and Gaussian Processes (GPs). As demonstrated by Lee et al. [1], deep neural networks of infinite width converge to GPs, allowing researchers to perform

exact Bayesian inference. By mapping ML architectures to analytical statistical models, one can compute exact variance and covariance, yielding rigorous confidence intervals without the need for expensive simulation algorithms.

2.2. Data-Driven Methods

When analytical boundaries are too complex to derive—such as with massive, highly non-linear models—data-driven methods provide a widely applicable alternative. These include ensemble methods, bootstrapping, and exploratory metric evaluations. Recent studies on large-scale AI, such as the exploratory studies of uncertainty measurement in Large Language Models by Huang et al. [2], highlight how empirical observations of model outputs under varying inputs or conditions can be used to gauge confidence boundaries. For statistical agencies, adopting data-driven uncertainty metrics ensures that even off-the-shelf algorithms can be monitored for reliability.

2.3. Bayesian Methods

Bayesian methods address uncertainty by construction. Instead of learning fixed weights, Bayesian Deep Learning finds a probability distribution over the network's weights [3]. This paradigm is exceptionally valuable in the age of large-scale AI, as it naturally outputs predictive distributions rather than mere point estimates. Consequently, any prediction generated by a Bayesian neural network inherently contains information about the model's confidence, perfectly aligning with the strict quality frameworks required for official statistics.

3. Concrete Applications at Statistics Iceland

We illustrate these principles with several concrete applications currently in continuous evolution at Statistics Iceland. All projects are implemented using the open-source R statistical computing environment [5] and are publicly available to encourage transparency and collaboration.

3.1. Synthetic Datasets for Confidentiality Protection

Statistical Disclosure Control (SDC) is a legal and ethical requirement for NSIs. To protect confidentiality, we evaluated the use of deep neural networks—specifically autoencoders and Generative Adversarial Networks (GANs)—to produce synthetic datasets [5].

The synthetic data generators we employed are implemented by the R-packages *NPBayesImputeCat*[1], *synthpop*[2], *synMicrodata*[3], *synthesizer*[4], RGAN[5]. The DP tools we tested are available through open-source packages (*diffpriv*[6], *DPpack*[7]) implemented the Laplace, analytic/- Gaussian, exponential mechanisms. Organising the results is still work in progress, but we share the code for generating all synthetic sets and DP outcomes on our dedicated repository [6].

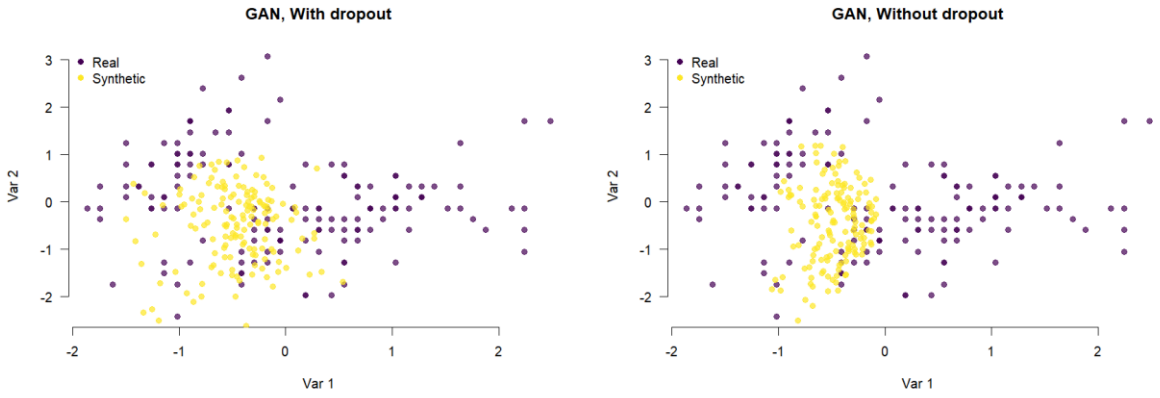

Figure 1. Synthetic data example with/without dropout regularization of GAN

In Figure 1, we see two synthetic data sets (yellow) produced by a RGAN from a unique, original data input (purple), in two cases: with (lefthand side) and without

[1] Hu, J. et al. (2026). NPBayesImputeCat: Non-Parametric Bayesian Multiple Imputation for Categorical Data. CRAN: Package NPBayesImputeCat

[2] Nowok, B., Raab, G. M., & Dibben, C. (2016). synthpop: Generating Synthetic Versions of Sensitive Microdata for Statistical Disclosure Control. *Journal of Statistical Software*, 74(11). CRAN: Package synthpop or the synthpop Homepage.

[3] Lee, J. (2026). synMicrodata: Synthetic Microdata Generator. CRAN: Package synMicrodata

[4] van der Loo, M. (2025). synthesizer: Fast, Robust, and High-Quality Synthetic Data Generation with a Tuneable Privacy-Utility Trade-Off. CRAN: Package synthesizer

[5] Neunhoeffer, M. (2022). RGAN: Generative Adversarial Nets (GAN) in R, on CRAN or GitHub.

[6] Rubinstein, B. I. P., & Aldà, F. (2017). diffpriv: Easy Differential Privacy. *Journal of Machine Learning Research*, 18. CRAN: Package diffpriv

[7] Giddens, S., & Liu, F. (2024). DPpack: Differentially Private Statistical Analysis and Machine Learning. CRAN: package DPpack

(righthand side) dropout, i.e. the regularization mechanism employed by the GAN algorithm to prevent overfitting and to increase the diversity of the output. The regularized version of the synthetic data has higher variability and utility.

The challenge in synthetic data generation lies in balancing the disclosure *risk* against the dataset's analytical *utility*. By applying uncertainty quantification frameworks to these generative models, we can rigorously tune and evaluate this risk/utility trade-off. Measuring the variance in generated samples ensures that the synthetic data accurately mimics the original distribution without confidently reproducing identifiable outlier profiles.

Applying a Bayesian formula for the disclosure risk one may obtain *closed forms* of parameters for specific risk profiles as well as *numerical solutions* for more general ones (see [4]). The argument uses the relative risk $r_i(p_i, q_i, t^*)$, (where p_i – the prior probability that the adversary identifies the individual *i* and q_i - the prior probability that the adversary makes a disclosure about the individual *i* while t^* – released results). It is defined as the ratio between the posterior probability (given the adversary model and the released information) and the prior probability of disclosure. The relative risk should be bound by the official statistics agency-allowed values ($r^*(p_i, q_i)$) and thus $\epsilon = min_{\{p_i, q_i \in (0,1)\}} \epsilon_i(p_i, q_i)$. We tested this hypothesis on example cases and may confirm the advantage of analytic/theoretical choices for the critical values of the epsilon (/and delta) differential privacy parameter(s).

3.2. Automatic Validation and Anomaly Detection

Data validation is traditionally one of the most resource-intensive phases of statistical production. To automate this, we implement deep isolation forests and autoencoders to detect anomalies within incoming data streams [7]. In an NSI context, blindly trusting a black-box anomaly detector can lead to high false-positive rates, overwhelming human reviewers.

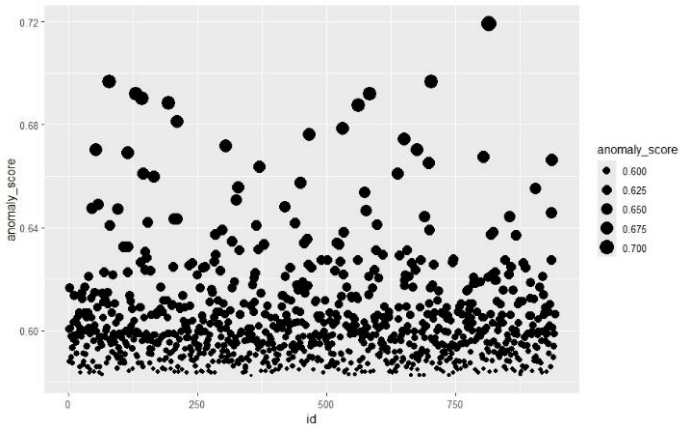

Figure 2. Anomaly scores on a microdata sample

By measuring the uncertainty of the anomaly scores (shown for an arbitrary dataset in Figure 2), the system can provide confidence bounds (see the posterior probabilities in Figure 3) alongside its flags.

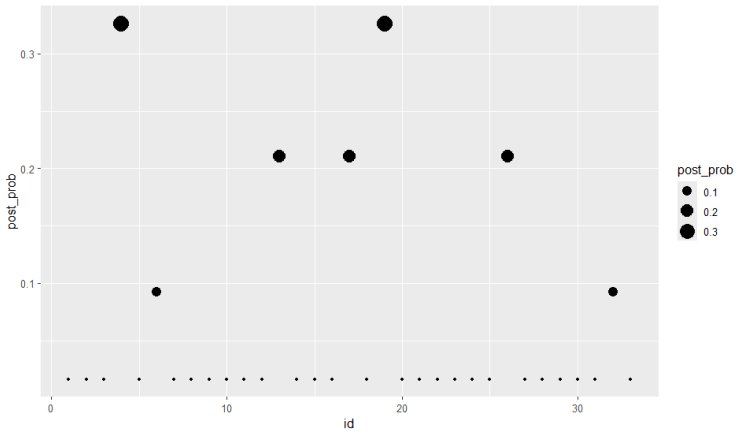

Figure 3. Posterior probability of anomaly scores illustration

This ensures that the validation process remains transparent and that human intervention is directed only toward highly uncertain or highly anomalous records.

3.3. Time Series Forecasting

Forecasting economic and social indicators requires strict adherence to reliability standards. We have performed time series forecasting utilizing Bayesian hierarchical models with stochastic (Gaussian-) process priors [8] for demographic predictions.

Interestingly, these priors are mathematically equivalent to certain deep and/or wide neural networks, and this is exactly the basis of our most recent methodology, under evaluation at this stage. An example of deep neural network model fitting and medium horizon predictions with confidence bands (reflecting both data and model variability) is shown in Figure 4, for migration numbers of incoming and outgoing individuals, several age groups and two genders.

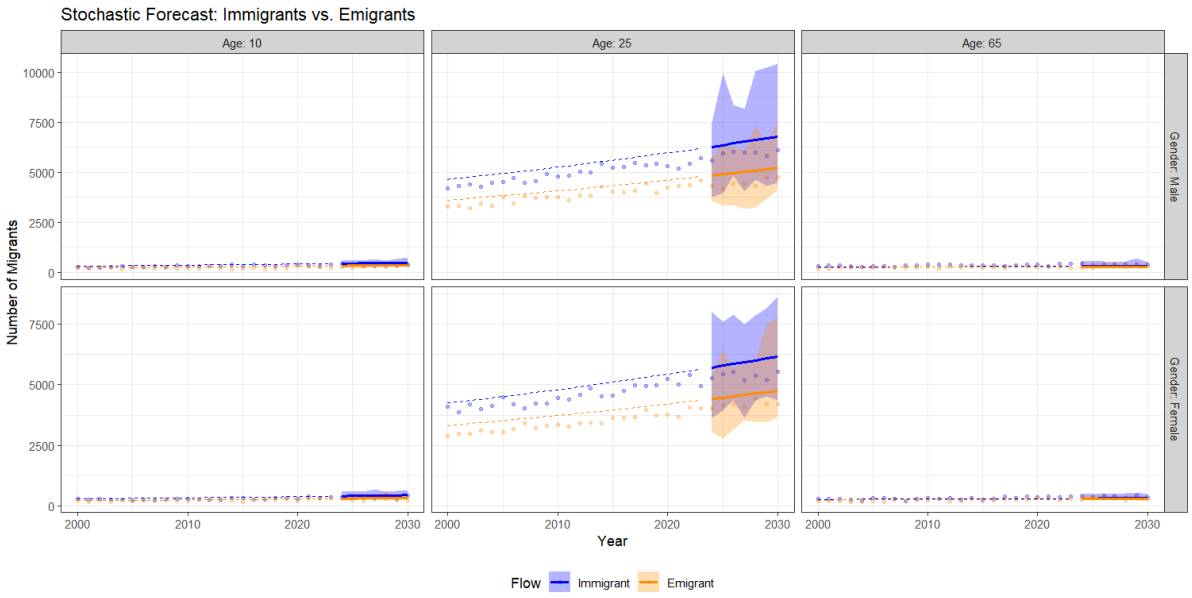

Figure 4. Prediction intervals of forecast im-/e- migration numbers depending on age and gender

We share the R-code which generates and tests stochastic population projections based on Bayesian regularized neural networks (using R-package *brnn*[8]) at repository [8], where our previous Hierarchical Bayesian models are linked as well.

By testing these Bayesian Neural Network forecasters against traditional methods, we exploit the non-linear pattern recognition capabilities of deep learning while retaining

[8] Perez Rodriguez, P., & Gianola, D. (2025). brnn: Bayesian Regularization for Feed-Forward Neural Networks. R package version 0.9.4.

the rigorous, built-in uncertainty bounds of Bayesian statistics. This prevents again the dissemination of overly confident, misleading point forecasts.

4. Conclusion

The modernization of official statistics through machine learning is an ongoing necessity, but it cannot come at the expense of transparency and reliability. By embracing analytical, data-driven, and Bayesian methods to quantify uncertainty, statistical agencies can safely deploy advanced algorithms like GANs, autoencoders, deep isolation forests and Bayesian neural networks.

The progress of these projects at Statistics Iceland, alongside all related materials and open code, is actively shared on GitHub repositories [6, 7, 8]. We encourage the official statistics community to utilize, critique, and contribute to this approach, ensuring that the future of statistical production is not only highly automated but also fundamentally accountable.

References

[1] Lee, J. et al. (2017). *Deep Neural Networks as Gaussian Processes*. arXiv preprint. URL: https://arxiv.org/abs/1711.00165

[2] Huang, Y. et al. (2023). *Look Before You Leap: An Exploratory Study of Uncertainty Measurement for Large Language Models*. arXiv preprint. URL: https://arxiv.org/abs/2307.10236

[3] Papamarkou, T. et al. (2024). *Bayesian Deep Learning is Needed in the Age of Large-Scale AI*. arXiv preprint. URL: https://arxiv.org/abs/2402.00809

[4] Kazan, Zeki, and Jerome Reiter. "Prior-itizing privacy: A Bayesian approach to setting the privacy budget in differential privacy." *Advances in Neural Information Processing Systems* 37 (2024): 90384-90430.

[5] R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

[6] Statistics Iceland Open Code: *Testing SDC Tools*. GitHub Repository. URL: https://github.com/violetacln/testingSDCtools

[7] Statistics Iceland Open Code: *SDMX_ML_Validation*. GitHub Repository. URL: https://github.com/violetacln/sdmx_ML_validation

[8] Statistics Iceland Open Code: *SIPP (Time Series/Forecasting for population projections)*. GitHub Repository. URL: https://github.com/violetacln/SIPP