Measuring and reporting uncertainty of AI and machine learning tools in official statistics

Violeta Calian, Anton Örn Karlsson

Statistics Iceland

Content

- Goal
- Solution
- Results
- Case 1: forecasting with Bayesian model/learning
- Case 2: ML classifiers
- Conclusions

Main goal

To improve the quality and reliability of the official statistics publications

while detecting, controlling and describing the limitations of production processes based on ML/AI algorithms

- Evaluating the *performance* of these methods
- Reporting the *uncertainty & confidence*, biases, failures
- Providing *interpretability* of results
- Preserving transparency (open-source code, when/if possible open-data)

Solution

- *Standard* mathematical statistics methodology,

  addapted and applied to advanced tools/methods/algorithms i.e.:

- Follow standard *steps*:
  - explore data
  - fit/train (model/algorithm)
  - evaluate and optimize (*parameters* <- performance metrics) and/or

    calibrate ($\mathbf{P}\left(\hat{Y} = Y \middle| \hat{P} = p\right) = p$) according to *goals*
  - quantify and report the ***uncertainty*** (due to data variability, model complexity/fit, distributional differences between train/test data measurement, data-model uncertainty interaction), **biases and failures**
  - describe/interpret the results in simpler terms (surrogate models, feature importance, conditional posterior distributions checks)

Analogy

- *(Bayesian/)* ***deep learning*** *(and even LLMs: tranformer-architecture with encoder and/or decoder blocks)*

→ *measures of uncertainty*

&

- *point.est. of mathematical statistics* → *conf.int. / cred.int.*

where uncertainty measures are built by using:

exact or approximative methods

Useful contributions

- **Analytic tools** to understand/evaluate deep learning models:

[*Deep Neural Networks as Gaussian Processes*, Lee, J. etal,
https://arxiv.org/abs/1711.00165]

- Large-scale evaluation of multiple **uncertainty estimation techniques** applied to various LLMs and tasks

[*Look Before You Leap: An Exploratory Study of Uncertainty Measurement for Large Language Models*, Huang, Y. etal,
https://arxiv.org/abs/2307.10236]

E.g.: Question → LLM →

- M1: **single** inference (max/avg prob. or max/avg entropy) → Answer A1
- M2: **Stochastic** inference/sample based/**model** variation (e.g. dropout, deep ensembles) (VR, VR0) → Answers A2.1, A2.2, ...
- M3: **Stochastic**/**Data** Perturbation (maxDiff VR, MaxDiff VR0) → Answers A3.1, A3.2, ...

- Using **Bayesian methods** to:
  - mitigate risks arising from overly confident yet incorrect predictions made by LLMs
  - provide uncertainties over predictions, which can enrich decision-making
  - enable the use of domain-knowledge priors

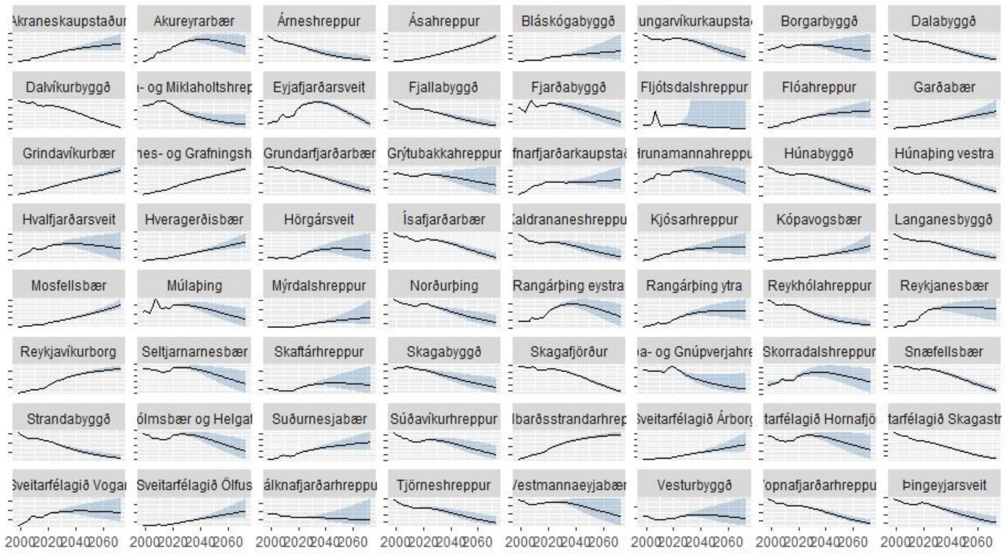
[*Position: Bayesian Deep Learning is Needed in the Age of Large-Scale AI*, Papamarkou, Th. Etal,
https://arxiv.org/abs/2402.00809]

Case 1 -

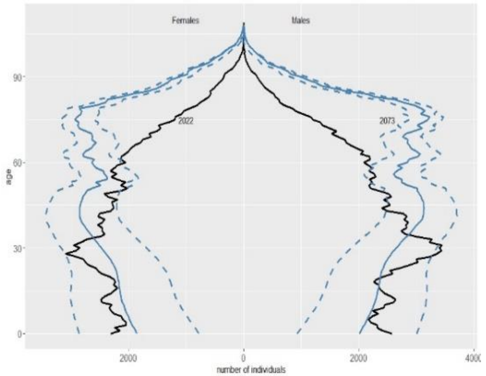
Forecasting (population) with:

Gaussian Process priors as components of

Bayesian hierarchical generalised additive models



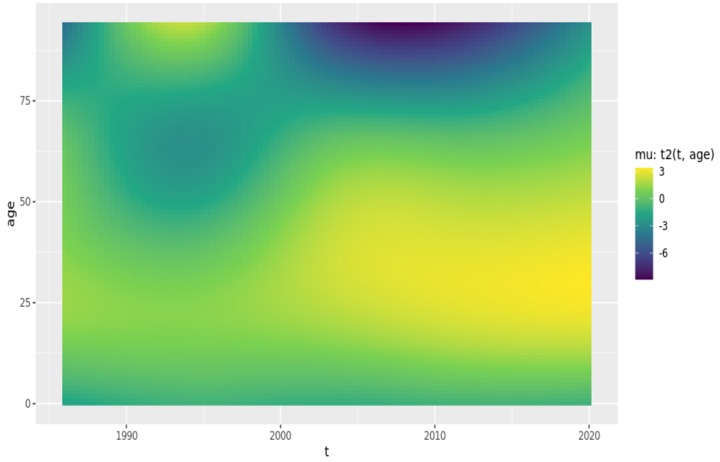
Experimental statistics: Population by municipality 1998-2073



Calian, V., *Methodology of population projections based on hierarchical Bayesian models*, WP-2023, https://hagstofas3bucket.hagstofa.is/hagstofan/media/public/2023/79a217c5-f567-4ddb-bed7-45329a32d531.pdf

and https://github.com/violetacln/SIPP

Case 1 (model and tools)



$$response \sim \boldsymbol{P}(f(..(f(\eta)))$$

$$\eta = \beta \boldsymbol{x} + \gamma \boldsymbol{z} + \sum s_k(a_k, by = c) + \ \dots$$

| | |

| | smooth dependencies ∋ *GP-priors*

| group level attributes and effects

population level attributes and effects

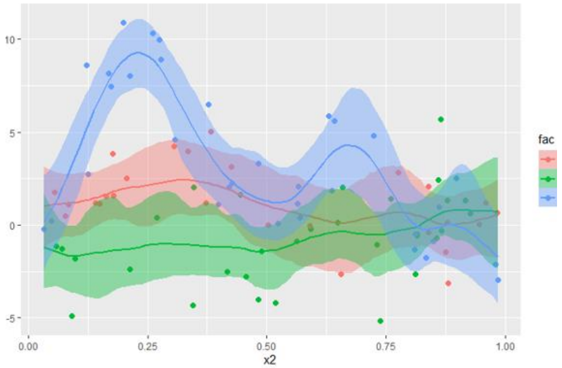
Or **nonlinear**: $\eta = f(x, z, a|c)$ or **unknown**!

R*-tools** (brms*** **and** ***mgcv*** **packages)**:

may include: *me*(…), *mi*(…), mi & me,

mo(…), cs(…), autocor(…)

posterior distribution checks, LOO-validation, …




```
# fit separate gaussian processes for different levels of 'fac'
fit4 <- brm(y ~ gp(x2, by = fac), dat2, chains = 2)
summary(fit4)
plot(conditional_effects(fit4), points = TRUE)
```

https://paul-buerkner.github.io/brms/reference/gp.html

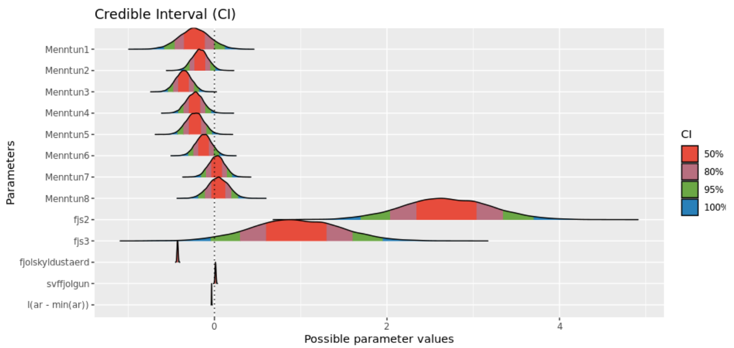
Case 1 (Illustration)

a. *Credible intervals (attributes of fertility rates)*

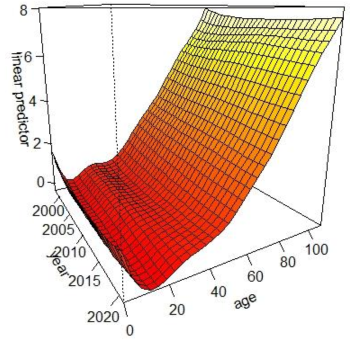
b. *Mortality (log-rates) surface*

c. *Migration uncertainty*

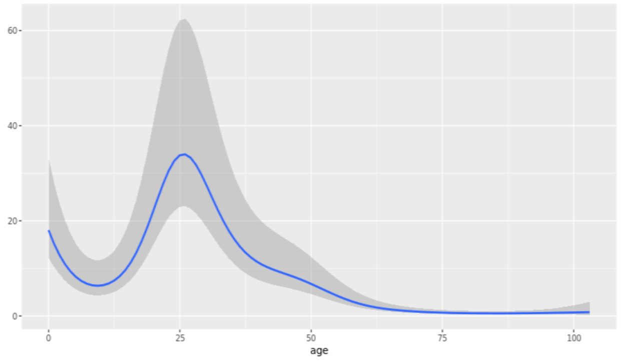
d. Posterior checks



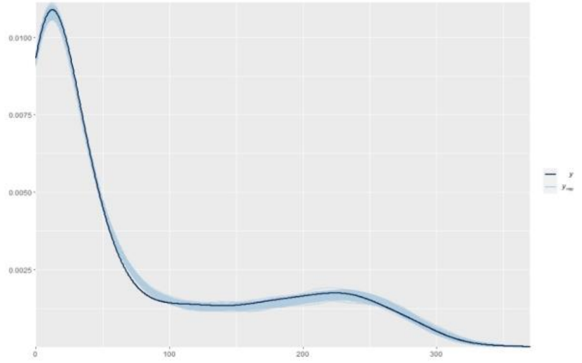
a.



b.



c.



d.

Case 2:
ML classifier for Census and survey optimisation

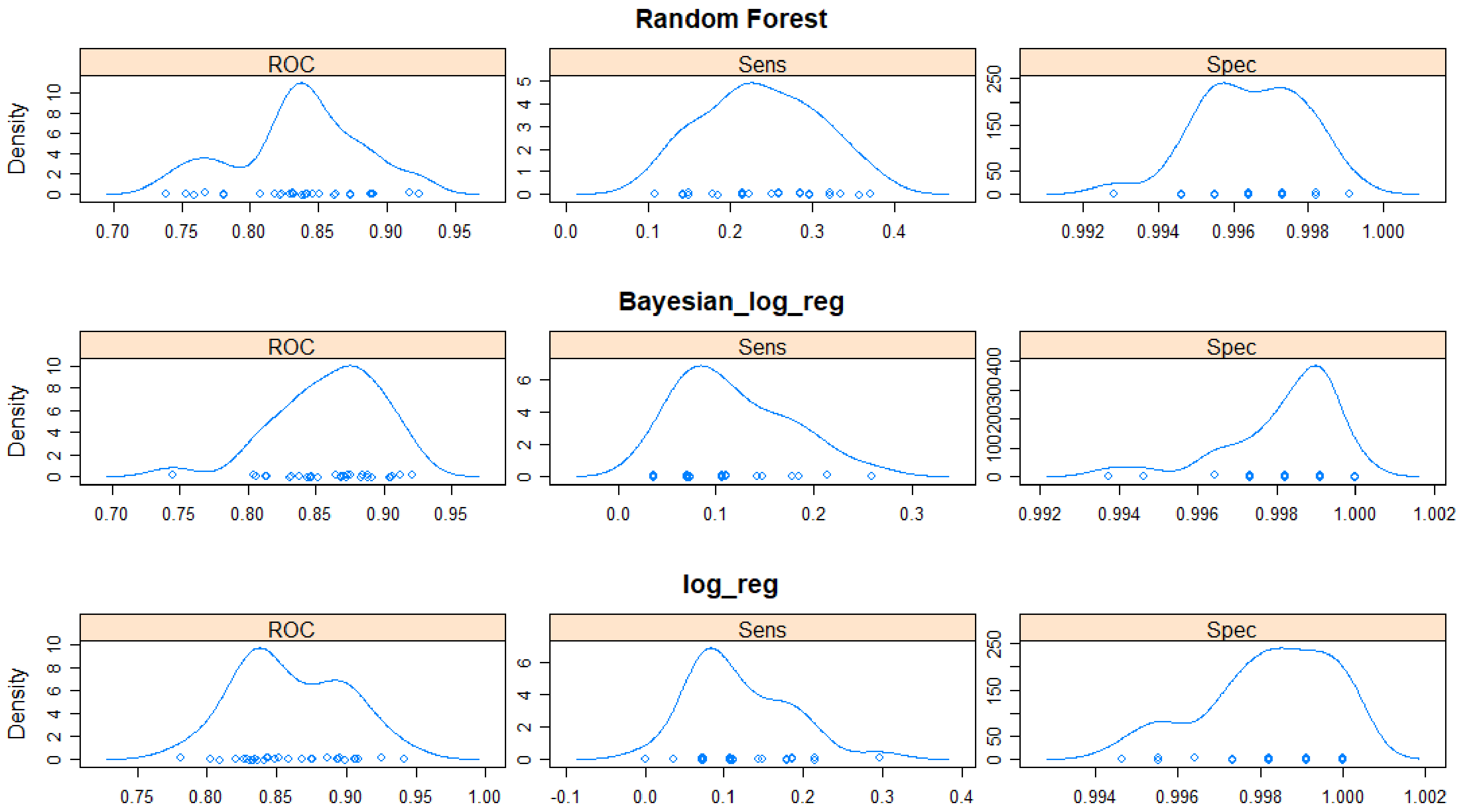
Completed:

- EDA, train/test/cross-validate, optimise/calibrate
- performance evaluation (multiple metrics)
- reporting uncertainty (of results and of performance metrics)
- interpretability tools

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

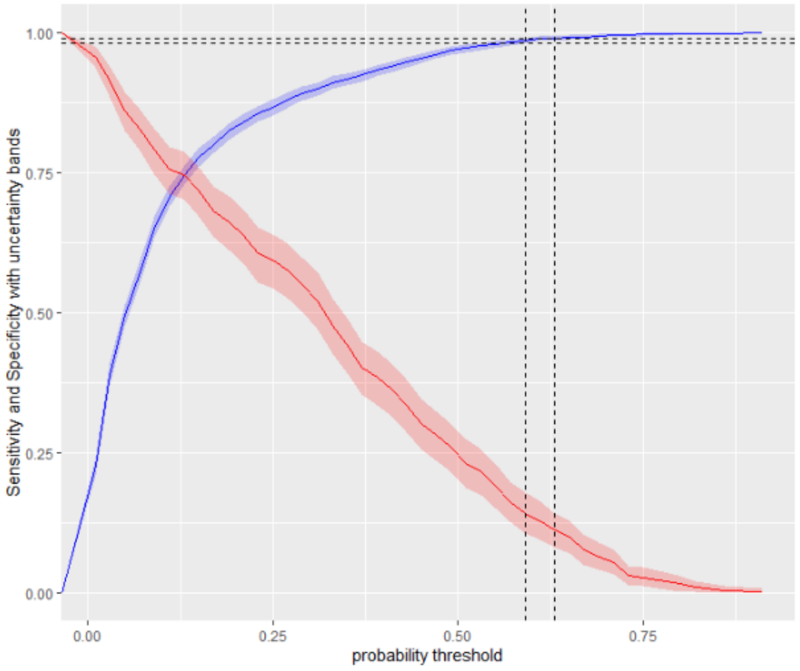
and https://github.com/violetacln/SLOPA

Case 2 (distributions of performance metrics)

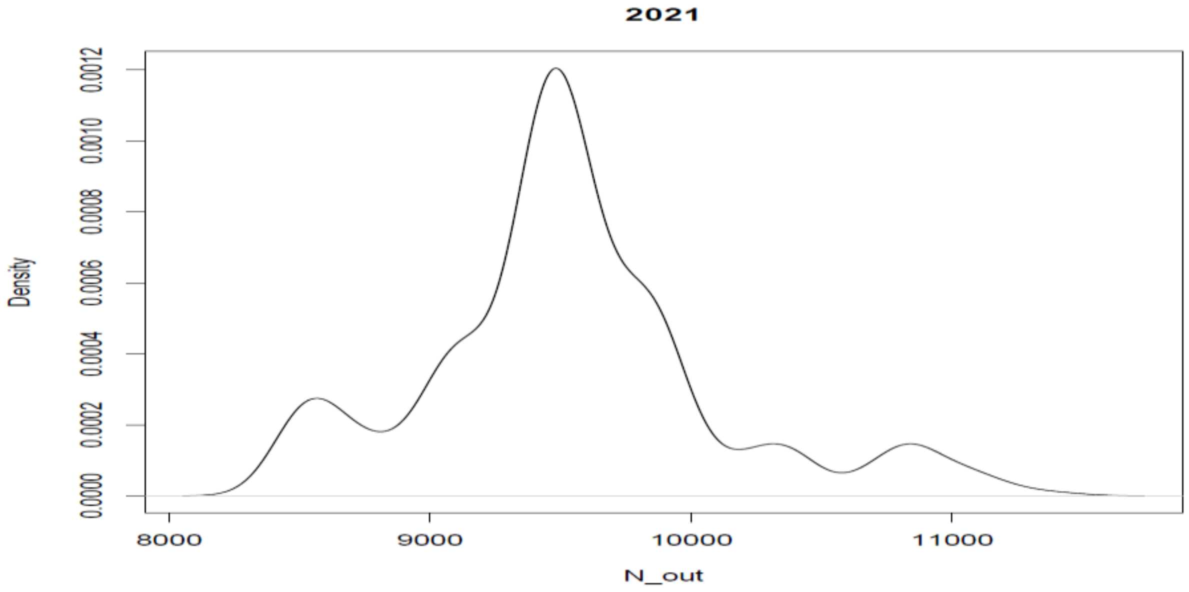


Case 2 *(uncertainty and variability)*

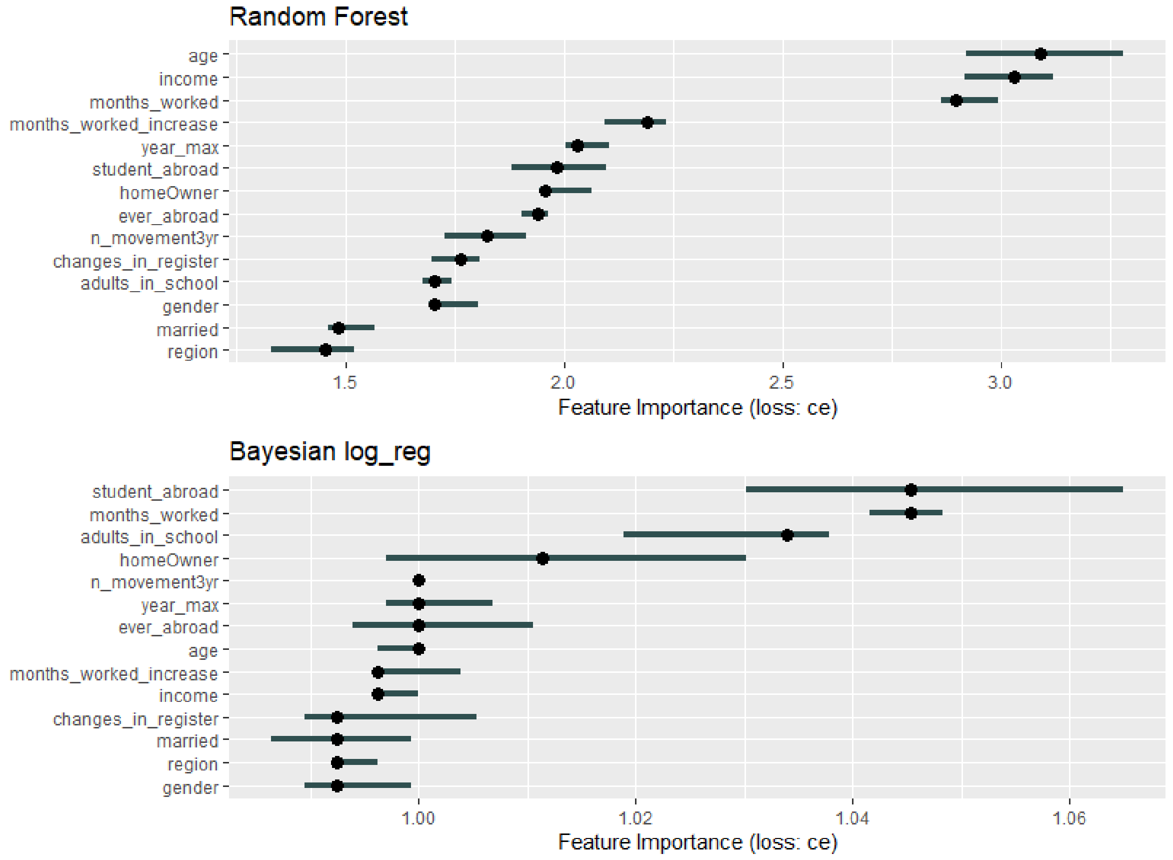
Optimum & uncertainty



Effects of training data variability



Case 2 *(interpretability example: feature importance with uncertainty measures in relation to error increase)*



Conclusions

Both types of statistical products based on new data science technologies (Bayesian and ML/deep learning)

and used for forecasting or classification purposes respectively

can be treated according to robust and transparent methods for **measuring, controlling and reporting uncertainty**.

The only limitations:

from insufficient computational resources, input data or incomplete domain/interpretation knowledge.

Thank you!

Violeta.Calian@hagstofa.is

https://github.com/violetacln