EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is partly
financed by the European Union

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

EUROPEAN CONFERENCE ON
QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Privacy protection, data validation and secure machine learning

with new statistical methods while preserving transparency

Violeta Calian

Statistics Iceland

violeta.calian@hagstofa.is

Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is partly
financed by the European Union

Content

Formulation of the problem and proposed solution

Results!

Examples:

- Data validation
- SDC
- ML classifier (for Census and surveys)

Formulation of the problem

Using advanced, up-to-date statistical *methods* to:

- Validate input data
- Produce high quality statistics/analysis
- Ensure statistical disclosure control (SDC)

While:

- Evaluating the *performance* of these methods
- Reporting the *uncertainty*, biases, failures
- Providing *interpretability* of results

And preserving transparency (open-source code)

Solution

Standard mathematical statistics methodology,
applied & addapted to advanced tools/methods/algorithms!

Celebrated examples:

- machine learning
- deep learning
- Bayesian modelling

Solution continued

Standard steps:

- explore
- train
- evaluate and optimise 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)
- describe/interpret the results in simpler terms (surrogate models, feature importance, conditional posterior distributions checks)

Results

Illustration of solution

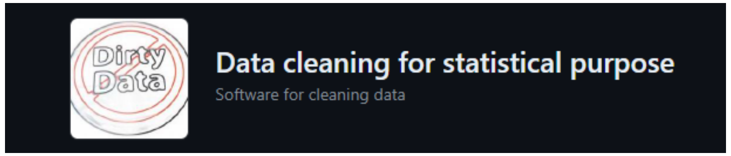
1. Data validation
2. SDC
3. ML classifier

with Bayesian modelling, deep learning and ML & *uncertainty*+performance reporting

Data validation

Classical approach – advantages and implementation (multi-step)

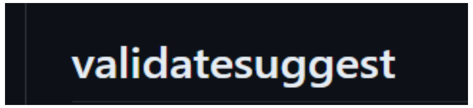
R-

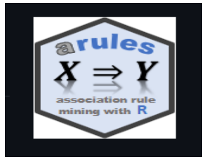

Data cleaning for statistical purpose
Software for cleaning data

https://github.com/orgs/data-cleaning/repositories

New methods – motivation and implementation

- rule discovery: **R-**


validatesuggest



- simultaneous, Bayesian, edit and imputation for continuous and categorical microdata

Data validation continued

Classical methods

Main steps

- Data and rules confrontation
- Error location
- Imputations

Reference

Statistics Netherlands: theory and R-packages (validate, errorlocate, simputation, validatesuggest)

New methods *& uncertainty*

ML (e.g. apriori, eclat algorithms) for rule discovery for confrontation step, plus error location and imputations

Bayesian hierarchical models:

(i) a Dirichlet process mixture of multinomial distributions (if categorical) or flexible joint probability (if continuous) as the model for the underlying true values of the data, with support restricted to the set of theoretically possible combinations,

(ii) a model for latent indicators of the values that are in error, and

(iii) a model for the reported responses for values in error.

https://www.tandfonline.com/doi/abs/10.1080/01621459.2015.1040881

https://dmanriqu.pages.iu.edu/preprints/LCM_Zeros_EdImp.pdf

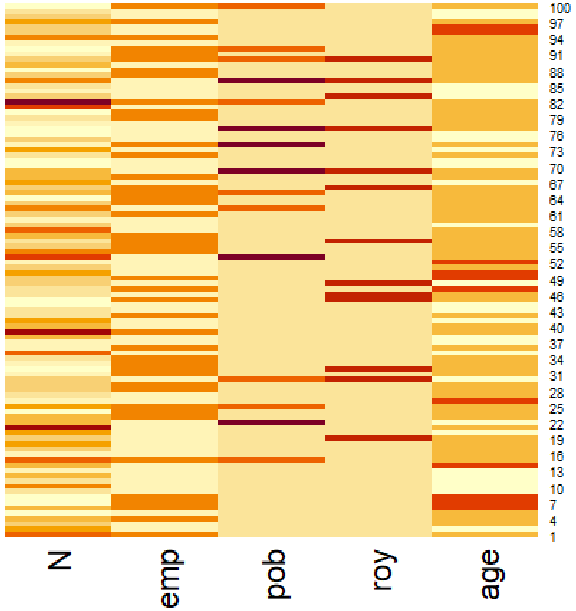
SDC

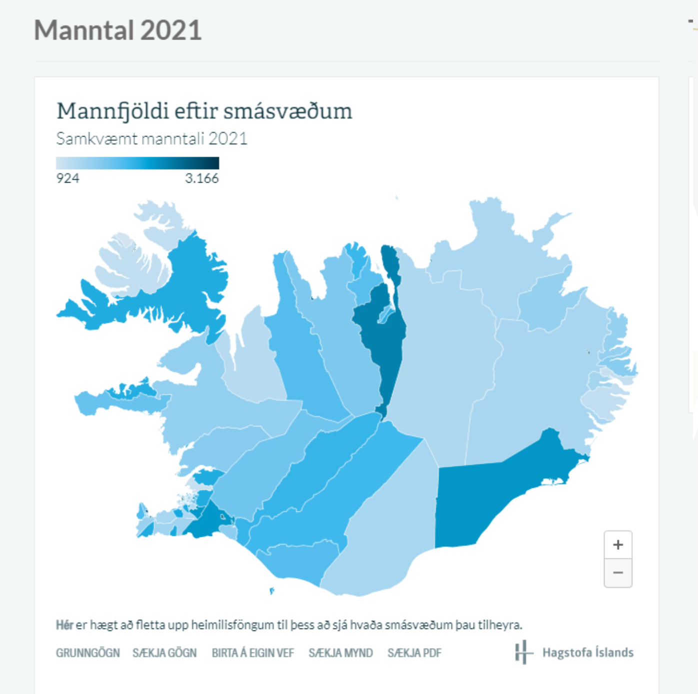
Evaluation of *classical* method

- Risk:
  identification
  attribute disclosure
  differencing
- Methods
  (non-/perturbative, variants, critical parameters)
- Residual risk & Information loss

Additional problems/*issues* –
examples (grid cell swapping)

- https://github.com/sdcTools




Manntal 2021

Mannfjöldi eftir smásvæðum

Samkvæmt manntali 2021

Hér er hægt að fletta upp heimilisföngum til þess að sjá hvaða smásvæðum þau tilheyra.

SDC continued

New methods/ideas - under evaluation:

- using Bayesian modelling for generating *synthetic* data
- Bayesian framework - most suitable reasoning:

  calculate predictive probabilities and disclosure *risk* (of original, protected, synthetic data)

  under model uncertainty (with e.g. model averaging) while using joint data distributions
- using *deep-learning* and/or cryptography inspired methods such as adversarial neural networks
- using *differential privacy* and its Bayesian variant which can guard against difficult scenarios built on deep learning

  “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”

ML classifier – multiple algorithms

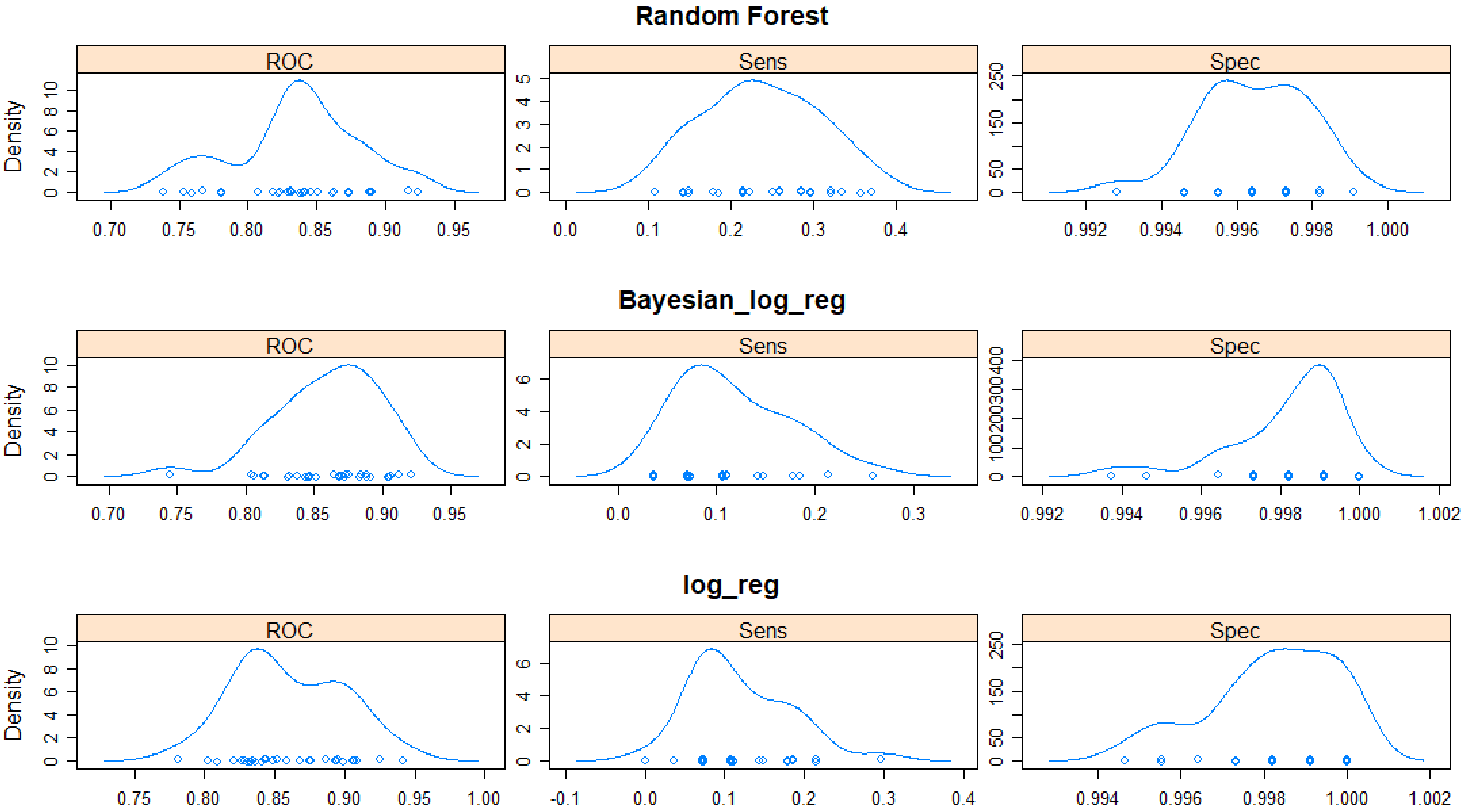
Completed:

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

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

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

ML classifier – example, performance metrics distributions



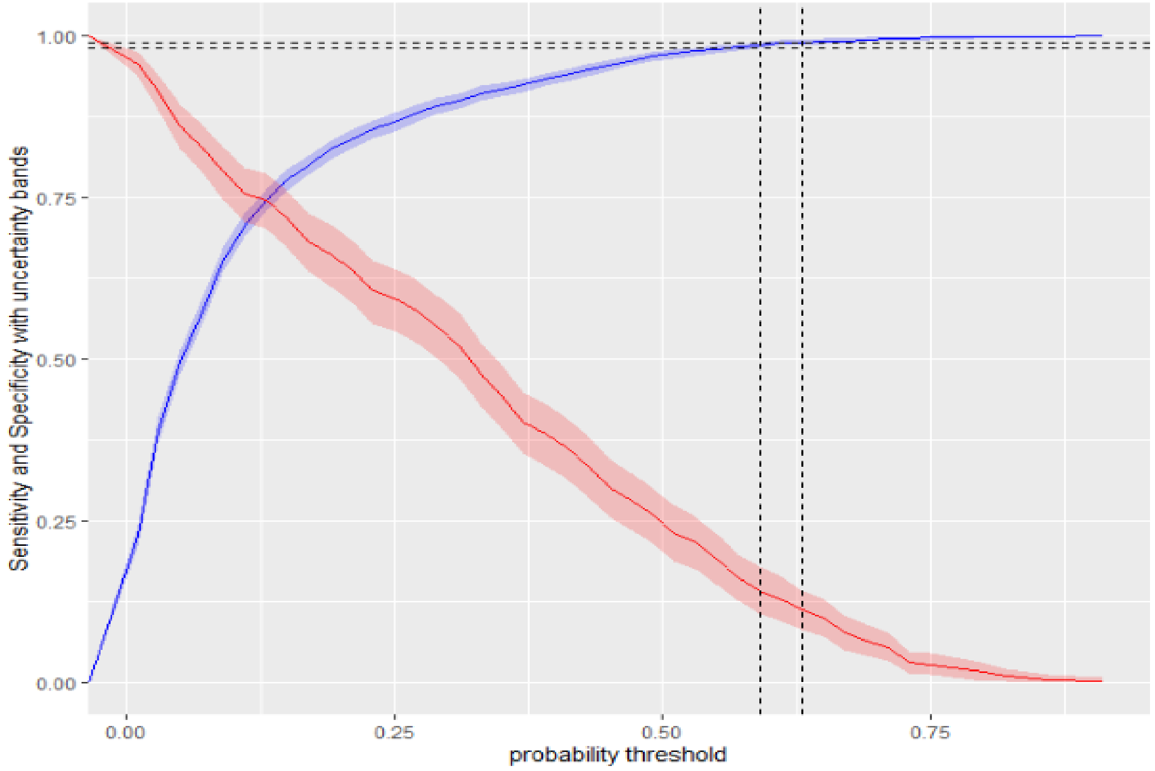
EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Instituto Nacional de Estatística
Statistics Portugal

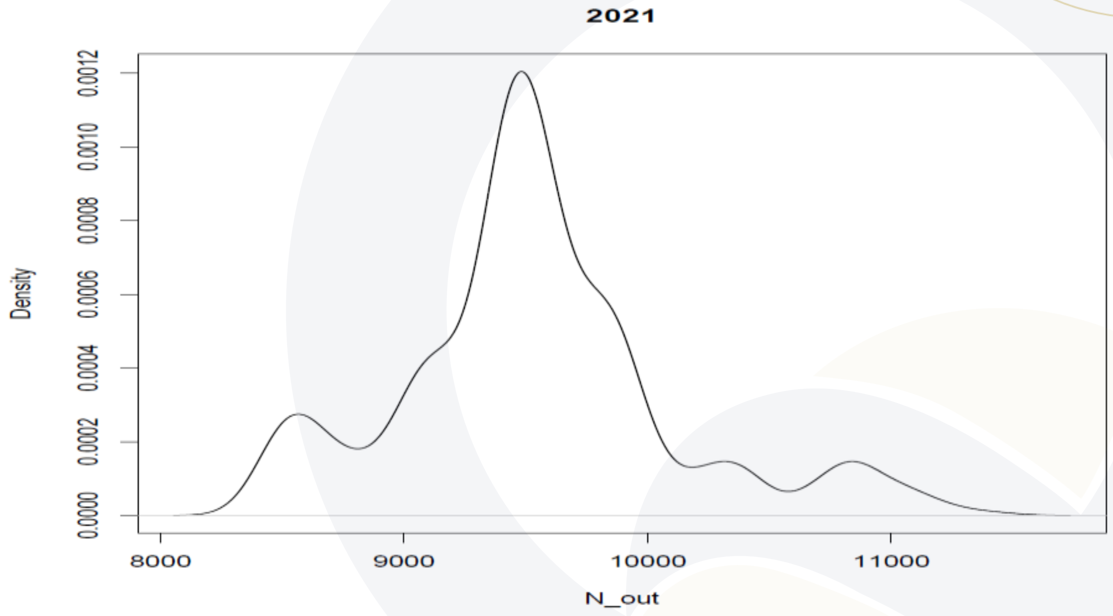
eurostat

The conference is partly financed by the European Union

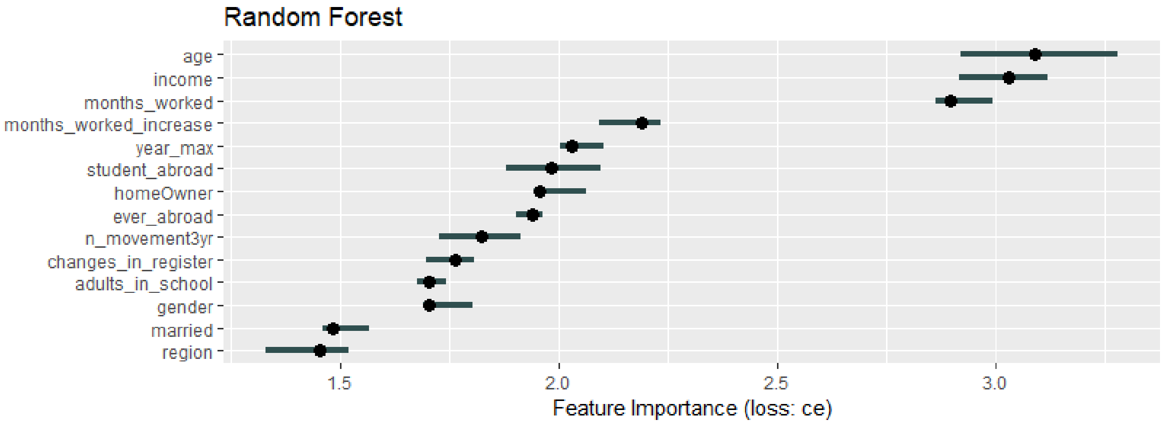
Confidence bands of RF performance metrics

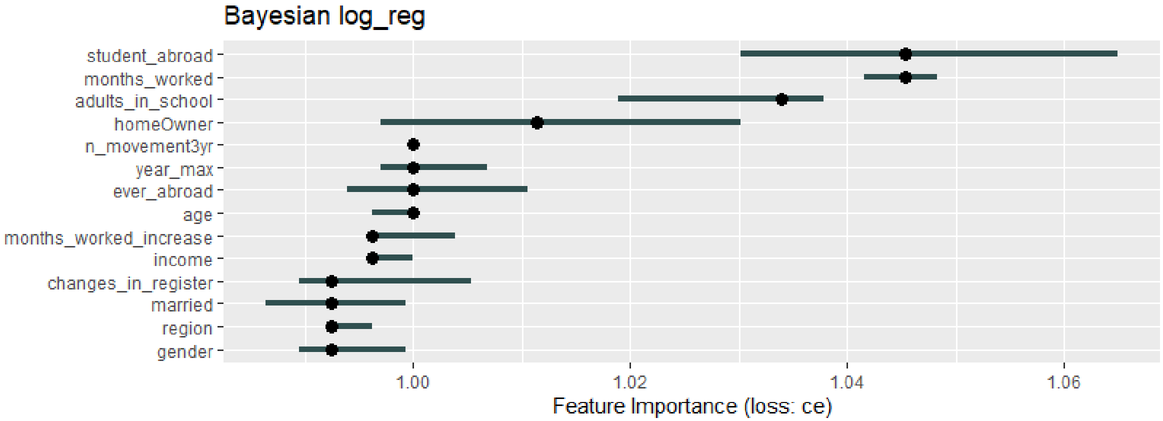


Effect of data variability on predicted outcome for a decision tree



ML classifier – example, feature importance





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

References

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

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. (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

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

Thank you!

EUROPEAN CONFERENCE ON
QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is partly
financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS 2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union