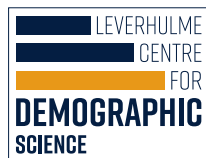


Six Slides for an Interview at Reuben College

7th June, 2024

Charlie Rahal

Associate Professor in Data Science and Informatics



Five Selected Experiences in ML & AI

- Intellectual diversity: ML & AI in (computational) social science!

Five Selected Experiences in ML & AI

● Intellectual diversity: ML & AI in (computational) social science!

Bigger Data Forecasting:
A computational combinatorial approach

Charles Rahal	Marco Barassi
Department of Sociology	Department of Economics
University of Oxford	University of Birmingham
charles.rahal@sociology.ox.ac.uk	m.r.barassi@bham.ac.uk

This version: **12th June, 2017**
Working paper version

Abstract

Current financial and macro-economic forecasting models which utilise 'Big Data' typically only incorporate a relatively small number of series (N) over time (T). We extend existing methods into a new dynamic forecasting framework which uses 128,712 series from the Federal Reserve Economic Data (FRED) database and evaluate it through a pseudo real-time forecasting application across multiple state-level house price indexes. We estimate approximately 482 million bivariate models drawn from a cross-section of dimension hitherto unseen in the econometric literature, selecting and weighting variables across different horizons, forecast approaches, model set sizes, rolling and recursive windows and combination metrics. Our approach performs extremely favorably against each of the three commonly used benchmarks which we evaluate it against, including a comparable big data macroeconomic (FAVAR) model which is outperformed in approximately 80% of instances. We aim to provide several stylized results to guide big data forecasting in economics and finance, with the algorithm showing a general preference for Bayesian Model Averaging over Smooth Adaptive Information Criteria weightings, iterated multi-step over dynamic multi-step forecast generation and for recursive over rolling estimation windows. There is further strong evidence of forecast error increasing in the volatility of underlying series, and we show that our forecast errors are relatively stable, except for during the recent financial crisis.

Keywords: Big Data, FRED, Forecasting, Macroeconometrics

JEL Classification: C33, C35, R30

For correspondence: Charles Rahal, Department of Sociology, University of Oxford, Manor Road Building, Manor Road, Oxford, OX1 3UQ. The scripts and functions (written in Python 3.x) for this project are available upon request and will be made available on GitHub following publication of the paper, along with a copy of the dataset.

Five Selected Experiences in ML & AI

● Intellectual diversity: ML & AI in (computational) social science!

The InterModel Vigorish (IMV) as a flexible and portable approach for quantifying predictive accuracy with binary outcomes

Benjamin W. Domingue^{a1}, Charles Rahal^{a1}, Jessica Faul^{a2}, Jeremy Freese^{a2}, Clint Kauglia^{a2}, Alexandre Rigou^{a1,a2}, Ben Steinhilber^{a2}, and Ajay Shauhar Tripathi^{a2}

^{a1}Stanford University

^{a2}University of Oxford

^{a3}University of Michigan

^{a4}Institute for Futures Studies, Stockholm

^{a5}Lund University

¹bdomingue@stanford.edu & charles.rahal@sociology.ox.ac.uk

²Alphabetized

October 25, 2023

Acknowledgements

This work was supported by the Jacobs Foundation (B.D.), the Levenshane Centre for Demographic Science (The Leverhulme Trust: C.R.) and Nuffield College (C.R.), and from Haskelbokenstiftelsen forskningsstöd (P.T.-E.H.). The authors would like to acknowledge Dan Baki, Richard Bevan, Kyla Chandler, Derrick Chao, Per Engström, Giuseppe Ieracini, David Redding, Mike Schür, Niklas Törsell, Mark Verhaeghe, Shixiao Wang, and Tobias Woldman for helpful feedback on early drafts of this manuscript, and Taha Yaseen and Victor Malmgren for assistance with the Isotest data. The HRS (Health and Retirement Study) is sponsored by the National Institute on Aging (grant number NIA U01AG000740) and is conducted by the University of Michigan.

Author Contribution Statement

BD conceived the original idea and study design, wrote simulations, conducted empirical analyses, and drafted and revised the manuscript. CR undertook code reviews, empirical analyses, visualizations, and contributed to drafting and revising the manuscript. RK worked on empirical applications, and developed a Shiny application. AST and AR worked on the key motivating features of the IMV, including but not limited to notational aspects and analogies to the Kelly Criterion. JFs, JFs, KK, AR, BS and AST all commented on the manuscript.

Five Selected Experiences in ML & AI

● Intellectual diversity: ML & AI in (computational) social science!

1 Introducing **RobustIPy**: An efficient multiversal library with
2 model selection, averaging, resampling, and out-of-sample
3 analysis.

4 Daniel Valdenegro Barral^{1,2}, Charles Rahal^{1,2,3}, and Jiantai Yan^{1,2,4}

5 ¹ Leverhulme Centre for Demographic Science, University of Oxford

6 ²ESRC Centre for Care, University of Oxford

7 ³Suffield College, University of Oxford

8 ⁴Wolson College, University of Oxford

9 June 6, 2024

10 Abstract

11 The specification curve literature brings integrity into the scientific record, but
12 further work remains to be done to adequately integrate researcher-based degrees of
13 freedom. Here we present RobustIPy, an efficient, stable and highly reproducible
14 tool with a user-friendly interface which allows researchers to conduct advanced
15 multiversal analysis. It introduces several new features: it computes confidence
16 intervals using bootstrap π -sampling; it undertakes combinatorics across possible
17 dependent variables; it undertakes model selection and averaging; and it conducts
18 Joint-Inference tests. It also conducts out-of-sample analysis and compares pre-
19 dictive models of $\hat{y}$ against two baselines (null and fully specified models). We
20 demonstrate the utility of this library through applications to three unpublished
21 examples in the academic literature.

22 **Keywords:** *Model Uncertainty, Statistical Software, Open Science*

● Time series boosting.

● InterModel Vigorish (x4).

● RobustIPy.

For correspondence: Daniel Valdenegro, Leverhulme Centre for Demographic Science and Centre for
Care, Department of Sociology, University of Oxford, OX1 1JD, United Kingdom. Tel: 01863 260370.
Email: daniel.valdenegro@demography.ox.ac.uk. The extensive code library which accompanies this
work can be found at <https://www.github.com/dvalden/robustipy>. The authors are grateful to
commentors on earlier versions of the work from participants at an LCCS Internal Meeting, the Centre for
Care, and participants at the ICSSS conference. The authors are grateful for funding from the ESRC (as
ESRC Large Centre grant), and the Leverhulme Trust, the Leverhulme Centre for Demographic Science
(LCCS), and Suffield College.

Five Selected Experiences in ML & AI

● Intellectual diversity: ML & AI in (computational) social science!

Offshoring Emissions through Used Vehicle Exports

Saul Justin Newman^{1, 2}, Kayla Schulte³, Micol Matilde Morellini^{1, 4, 5},
Charles Rahal^{1, 4}, and Douglas R. Leasure^{1, 4}

¹ Leverhulme Centre for Demographic Science, University of Oxford

² University College, University of Oxford

³ Imperial College London

⁴ Nuffield College, University of Oxford

⁵ Department of Sociology, University of Oxford

February 22, 2024

Please cite this article as:

Newman, S.J., Schulte, K., Morellini, M.M., Rahal, C., and Leasure, D., 'Offshoring emissions through used vehicle exports'. *Nat. Clim. Chang.* (2024).
<https://doi.org/10.1038/s41558-024-01943-1>

Abstract

Policies to reduce transport emissions often overlook the international flow of used vehicles. We quantify the rate at which used vehicles generated CO₂ and pollution for all used vehicles exported from Great Britain – a globally leading used-vehicle exporter – across 2009–2021. Destined for low-middle income countries, exported vehicles fail roadworthiness standards and, even under extremely optimistic ‘functioning-as-new’ assumptions, generate at least 13–53% more emissions than scrapped or on-road vehicles.

Keywords: *Big Data, Machine Learning, Climate Change*

Correspondence: All correspondence and requests for material should be directed to Dr Saul Newman, the corresponding author, at saul.newman@demography.ox.ac.uk. **Acknowledgements:** We would like to thank Professor Sarah Sharpley for her support as chief scientific officer at the Department for Transport, climate change engineer Dr Baber Jibir for his wonderful feedback, Mr Grant Threder for his excellent help and advice, and the Driver and Vehicle Standards Agency and the Department for Transport for their substantial help and support, without which this project would not be possible. **Project Ethics:** This project was approved by the University of Oxford's Departmental Research Ethics Committee (Sociology) under ethics approval SOCJ220127A2J246. A compliance assessment was undertaken by Oxford Population Health. **Replication Materials:** Replication models and code are freely available under a creative commons CC-BY-NC 4.0 license from a *figshare repository* or on request from the corresponding author. **Funding:** The authors are grateful for support from the Leverhulme Trust (Grant RC-2018-092) for the Leverhulme Centre for Demographic Science, and Nuffield College.

- Time series boosting.
- InterModel Vigorish (x4).
- RobustiPy.
- Imputing vehicle exports.

Five Selected Experiences in ML & AI

● Intellectual diversity: ML & AI in (computational) social science!

1 Large Language Models Predict Cognition and Education Close
2 to or Better than Genomics or Expert Assessment

3 Tobias Wolfman¹, Felix C. Tropf^{2,3}, and Charles Rahal⁴

4 ¹Faculty of Sociology, Bielefeld University, Germany

5 ²Centres for Longitudinal Studies, Social Research Institute, University College London, UK

6 ³Department of Sociology, Purdue University, USA

7 ⁴Leverhulme Centre for Demographic Science and Nuffield College, University of Oxford, UK

8 June 4, 2024

9 Abstract

10 Previous research using standard social survey data has emphasized a relative
11 lack of power when predicting educational and psychological outcomes. Leveraging
12 a unique longitudinal dataset, we explore predictive ceilings for educational attainment,
13 cognitive abilities, and non-cognitive traits. Integrating various measures
14 of computational linguistics and large language model-based challenges within a
15 SuperLearner framework trained on short aspirational essays written at age 11, we
16 accurately predict cognition and non-cognitive traits at the same and later age to a
17 similar degree as teacher assessments, and substantially better than genomic data.
18 The same is true for predicting final educational attainment. Combining test, grades
19 and teacher assessments into an ensemble model, we can nearly perfectly predict
20 cognitive ability (70%, close to test-retest reliability) and explain 38% of individual
21 differences in attainment. These findings show that recent advances in large
22 language modeling and machine learning equip behavioural scientists with tools for
23 accurate prediction of psycho-social features.

24 **Keywords:** *Psychological Methods, Longitudinal Surveys, Deep Learning*

Data Availability: All non-genetic data associated with the project can be downloaded from the website of the UK Data Service, segmented by sets of datasets, a joint one for theory-led and additional ones for each subsequent wave. Essays are accessible separately. Access to genetic data requires a separate application to the data access committee of the center for longitudinal studies (CLS). **Code Availability:** All statistical software used for generating the results in this paper are carefully listed in a private GitHub repository. This manuscript has been presented at the Social Science Genetics Network Conference 2022 in Bologna and at the International Society for Intelligence Research Conference 2022 in Berkeley. A Preprint is available at <https://arxiv.org/abs/2405.04877>. **Acknowledgements:** F.C.T. was supported by the Laboratory of Excellence in Economics and Decision Sciences (LabEx 1016) funded under the French National Research Agency (ANR), Investissements d'Avenir (Grant ANR-11-LABX-0047) and the USR project FRODOX (EPYC200001). C.R. is grateful for support from the Leverhulme Trust (Grant HC-2018-002) for the Leverhulme Centre for Demographic Science, and Nuffield College. The authors declare no competing interests. **Correspondence:** Correspondence concerning this article should be addressed to Tobias Wolfman, Faculty of Sociology, Bielefeld University, Universitätsstraße 25, 33615 Bielefeld, Germany. Email: tobias.wolfman@uni-bielefeld.de

● Time series boosting.

● InterModel Vigorish (x4).

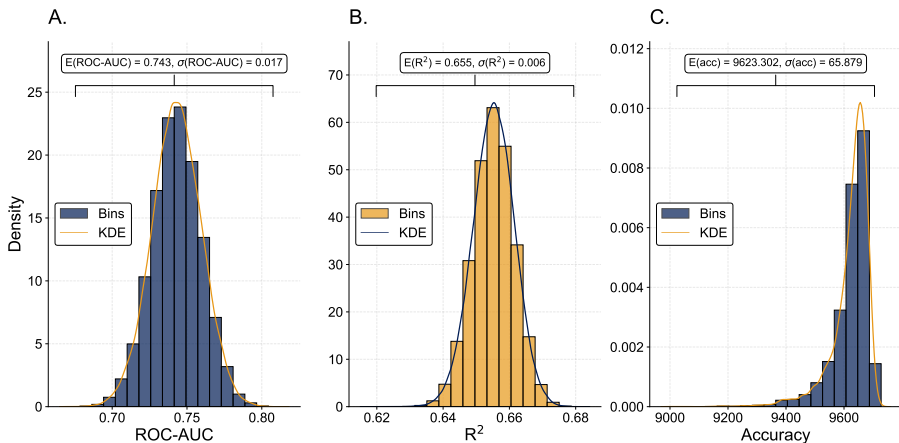
● RobustiPy.

● Imputing vehicle exports.

● LLMs for life course prediction.

● Latter informs a large ERC Grant Application (inc. Ethics/Values).

One Specific Example



‘On the Responsible use of PRGNs in Computational Science’

With Jiani Yan and Mark Verhagen

Student Wellbeing and Development

1. Hold regular open office in college for student support.
 - Experience of supervising smooth and less smooth DPhils!
 - An essential part of this is welfare provision to students.

Student Wellbeing and Development

1. Hold regular open office in college for student support.
 - Experience of supervising smooth and less smooth DPhils!
 - An essential part of this is welfare provision to students.
2. Monthly 'Data Science Drop-Ins' in a welcoming environment.
 - Students submit or bring questions, ML & AI Fellows discuss.
 - I would also offer short programming courses as necessary.

Student Wellbeing and Development

1. Hold regular open office in college for student support.
 - Experience of supervising smooth and less smooth DPhils!
 - An essential part of this is welfare provision to students.
2. Monthly 'Data Science Drop-Ins' in a welcoming environment.
 - Students submit or bring questions, ML & AI Fellows discuss.
 - I would also offer short programming courses as necessary.
3. Foster student community development.
 - Create a space where students share and collaborate.
 - This is what separates Oxford, and is done differentially.

Three Types of Events Related to 'ML and AI'

1. Hackathons are a *great* way to learn, develop, and contribute.
 - They also foster community development.



Three Types of Events Related to ‘ML and AI’

1. Hackathons are a *great* way to learn, develop, and contribute.
 - They also foster community development.



2. (Co-)Convene a seminar series in ‘ML & AI for Society’.
 - Strong integration of ‘Ethics and Values’ stream at Reuben.

Three Types of Events Related to ‘ML and AI’

1. Hackathons are a *great* way to learn, develop, and contribute.
 - They also foster community development.



2. (Co-)Convene a seminar series in ‘ML & AI for Society’.
 - Strong integration of ‘Ethics and Values’ stream at Reuben.
3. Plan to integrate RROx events into and around Reuben.
 - ‘Open Science’, ‘Reproducibility’ and ‘Responsible Research’.

Academic Service at the Heart of Everything

1. Successful recruitment: I *enjoy* admissions related tasks!
 - Experience of admissions/recruitment across multiple divisions.
 - Experience of both student and fellowship recruitment.

Academic Service at the Heart of Everything

1. Successful recruitment: I *enjoy* admissions related tasks!
 - Experience of admissions/recruitment across multiple divisions.
 - Experience of both student and fellowship recruitment.
2. Successful governance: surely key to Reuben's success.
 - Huge range of administrative experience across university.
 - This is unfortunately something which Nuffield didn't afford.

Academic Service at the Heart of Everything

1. Successful recruitment: I *enjoy* admissions related tasks!
 - Experience of admissions/recruitment across multiple divisions.
 - Experience of both student and fellowship recruitment.
2. Successful governance: surely key to Reuben's success.
 - Huge range of administrative experience across university.
 - This is unfortunately something which Nuffield didn't afford.
3. Successful financial management and fundraising:
 - Another area which I am looking for more experience in.

Academic Service at the Heart of Everything

1. Successful recruitment: I *enjoy* admissions related tasks!
 - Experience of admissions/recruitment across multiple divisions.
 - Experience of both student and fellowship recruitment.
2. Successful governance: surely key to Reuben's success.
 - Huge range of administrative experience across university.
 - This is unfortunately something which Nuffield didn't afford.
3. Successful financial management and fundraising:
 - Another area which I am looking for more experience in.
4. Strategic decision making: a chance to sail a new course.
 - Some experience of this at the DSU; excited to contribute more.

Full Involvement in College Life

1. Collaboration: Research projects with students and Fellows.
 - In particular, this would bridge ML & AI and Ethics & Values.

Full Involvement in College Life

1. Collaboration: Research projects with students and Fellows.
 - In particular, this would bridge ML & AI and Ethics & Values.
2. Participation: Regular social/cultural life (dinners, seminars).
 - Dining With Dinsosaurs: 'How predictable is your life?'

Full Involvement in College Life

1. Collaboration: Research projects with students and Fellows.
 - In particular, this would bridge ML & AI and Ethics & Values.
2. Participation: Regular social/cultural life (dinners, seminars).
 - Dining With Dinsosaurs: 'How predictable is your life?'
3. Participation/administrative help with 'extraordinarily' events.
 - Reuben Insights Festival, the Gaudy, Reuben-Linacre ball.

Full Involvement in College Life

1. Collaboration: Research projects with students and Fellows.
 - In particular, this would bridge ML & AI and Ethics & Values.
2. Participation: Regular social/cultural life (dinners, seminars).
 - Dining With Dinsosaurs: 'How predictable is your life?'
3. Participation/administrative help with 'extraordinarily' events.
 - Reuben Insights Festival, the Gaudy, Reuben-Linacre ball.
4. Engagement: Represent college in public forums/outreach.
 - Formally, at administrative events!
 - Informally, at social and sporting events (e.g. Oxford VIII's)!

Thank you for your time!