How to deal with machine learning bias in economic history

Torben Johansen (University of Southern Denmark)
Julius Koschnick (University of Southern Denmark)
Christian Vedel (University of Southern Denmark)[1]

Draft: March 8, 2026 • 51st Annual Meeting of the Social Science History Association

Keywords: Machine learning, Large language models, bias correction, economic history
JEL codes: N01, C55, C80, C81, C82, C45

Machine learning (ML) and large language models (LLMs) have rapidly expanded the empirical toolkit available to economic historians. These tools allow researchers to process large historical corpora, extract structured information from unstructured sources, and construct novel quantitative measures from text. Tasks that previously required extensive manual coding, such as record linkage, classification of occupations, sentiment analysis of political texts, or topic assignment in publication datasets, can now be executed at scale with automated models. As a result, ML tools have significantly expanded the feasible size of datasets and opened new research questions that were previously inaccessible due to data constraints.

Despite these opportunities, the use of ML tools raises the concern that ML predictions may exhibit mild to severe statistical bias that can amplify in downstream analyses. Prediction errors from ML models are rarely random. Instead, they are likely to be systematically related to linguistic features of the input, as well as limitations of the model or training data. Even highly accurate models can therefore generate biased estimates when their predictions are used as inputs in downstream tasks such as regression analyses. Indeed, prediction errors can substantially distort downstream estimates, even when models achieve accuracy rates above 90 percent (Egami, Hinck, Stewart, & Wei, 2024).

This issue is particularly acute in economic history. Historical texts often differ substantially from the modern corpora used to train contemporary language models. They contain archaic language, references to historical institutions or debates unfamiliar to modern models, as well as stylistic conventions that differ markedly from modern style. These differences increase the risk of temporal contamination and anachronisms in ML predictions. Combined with the selective survival of historical texts and the overrepresentation of Western sources in digital archives, these factors create substantial potential for systematic prediction bias when ML tools are applied to historical data.

We argue that the discipline needs to respond to this challenge of bias from ML predictions by estimating the structure of the bias and correcting for it in downstream tasks. Crucially, simple validation is not sufficient to account for structural errors. Concretely, we recommend the adoption of debiasing frameworks (Angelopoulos, Bates, Fannjiang, Jordan, & Zrnic, 2023; Carlson & Dell,

[1]University of Southern Denmark, Economics Department, Campusvej 55, Odense, Denmark

2025; Egami, Hinck, Stewart, & Wei, 2023; Egami et al., 2024; Ludwig, Mullainathan, & Rambachan, 2025). In this approach, a randomly sampled subset of observations is annotated by historically skilled expert coders to construct a gold-standard reference. These annotations are then used to estimate the bias structure of ML predictions and to correct downstream estimates accordingly. The resulting estimators remain consistent while preserving the efficiency gains and statistical power afforded by large-scale ML predictions.

But these tools also increases the need for the traditional careful historian's skillset. To meaningfully apply the debiasing framework within economic history, we emphasize the role of the *careful historian* who is responsible for overseeing the team generating gold standard labels. We argue that correctly labeling historical data involves the skills and toolset of the skilled historian (see also Campbell, 2025). Before generating labels, documents should be subjected to source criticism where documents are located within the historical, political, and social context. Moreover, the *careful historian* should ensure that expert labelers have a strong domain knowledge in the historically relevant period to avoid anachronisms. All in all, we argue that the debiasing framework can constitute a productive interface between quantitative and careful qualitative work in economic history.

To facilitate the adoption of the debiasing framework in economic history, the paper provides a short overview of the literature applying ML tools in economic history and derives recommendations of how to deal with bias from ML predictions. For this, we develop a taxonomy of ML uses cases in economic history where we distinguish between three types of ML applications:

The first category is *ML to replace human annotators*. In these settings, ML models perform tasks that could in principle be completed by expert coders, such as classifying occupations, linking historical records or labeling political sentiment (see e.g. Arora & Dell, 2024; Dahl, Johansen, & Vedel, 2026; Giommoni, Loumeau, & Tabellini, 2026; Liu, Papanikolaou, Schmidt, & Seegmiller, 2025; Silcock, Arora, D'Amico-Wong, & Dell, 2024). The key advantage is scalability: ML allows researchers to annotate millions of observations at low marginal cost. However, because these predictions are imperfect, their direct use in empirical analyses may introduce systematic bias. For this class of applications, we argue that debiasing methods provide a general solution to bias from ML predictions. Researchers can construct a randomly sampled gold-standard dataset with expert annotations, use this dataset to estimate prediction bias, and correct downstream estimates accordingly. Importantly, this approach integrates ML tools with traditional historical methodology: The process of creating gold-standard labels requires careful source criticism and domain expertise, placing the work of the historian at the center of the empirical pipeline.

The second category involves *ML to account for missing data*. In many historical settings, researchers combine information from multiple sources that differ in coverage or completeness. ML models can be used to impute missing variables or to harmonize information across datasets (see e.g. Clark, Cummins, & Curtis, 2024; Paker, Stephenson, & Wallis, 2025). These applications can often be framed as a "source–target" problem, where one dataset contains complete information while another dataset contains missing values for variables of interest. How can researchers deal with ML bias in these cases? We argue that in a relevant set of cases, these problems can be reformulated as type 1 problems which makes it possible to apply a debiasing framework. In other cases, researchers might want to use existing data as a gold standard for missing data. Since this data is not missing at random, researchers need to either present convincing argument that the missing data is as-good-as-random or explicitly model their selection process (Carlson & Dell, 2025).

The third category consists of *ML for new measures*. Recent research increasingly uses embedding-based representations of text to derive measures of innovation, sentiment, political ideology, or knowledge spillovers (see e.g. Ash, Chen, & Ornaghi, 2024; Gorin, Heblich, & Zylberberg, 2025; Koschnick, 2025). These approaches treat the embedding space of language models as a geometric representation of meaning, allowing researchers to quantify semantic relationships between texts. Unlike annotation tasks, these measures often cannot be validated against a human-coded ground truth, because the underlying concept—such as semantic similarity or textual novelty—cannot be directly observed. For these applications, debiasing with gold-standard labels is typically infeasible simply because we lack a ground truth. Instead, we recommend extensive validation exercises that compare new ML-based indicators to established proxies and test the robustness of empirical results to alternative modeling choices.

Across all three application types, the paper emphasizes that ML tools should complement rather than replace the core methodological principles of historical research. Careful source criticism, transparent documentation of data construction, and explicit discussion of assumptions remain essential. In fact, the integration of ML within the debiasing framework may strengthen these principles by requiring researchers to explicitly document how gold-standard labels are constructed and how historical context informs the annotation process.

To conclude, it seems that ML has the power to transform the discipline of economic history. ML tools allow researchers to process very large datasets at low cost. Moreover, they also make it possible to create fundamentally new measures of economic activity. However, we argue that in order to make this research credible, the discipline needs to actively account for bias from ML predictions. In this paper, we argue that the debiasing framework offers a promising route forward. We further emphasize that the debiasing framework can combine the power of large scale quantitative analysis with the role of the careful qualitative historian. To facilitate this, we discuss different types of ML applications in economic history and suggest different strategies to actively account for bias in each of these cases.

References

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., & Zrnic, T. (2023). Prediction-powered inference. *Science*, *382*(6671), 669–674.

Arora, A., & Dell, M. (2024). *Linktransformer: A unified package for record linkage with transformer language models.* (arXiv:2309.00789)

Ash, E., Chen, D. L., & Ornaghi, A. (2024). Gender attitudes in the judiciary: Evidence from us circuit courts. *American Economic Journal: Applied Economics*, *16*(1), 314–350.

Campbell, C. (2025). The historian in the age of ai. *Transactions of the Royal Historical Society*, 1–11. doi: 10.1017/S0080440125100509

Carlson, J., & Dell, M. (2025). A unifying framework for robust and efficient inference with unstructured data. Retrieved from https://arxiv.org/abs/2505.00282

Clark, G., Cummins, N., & Curtis, M. (2024). Three new occupational status indices for england and wales, 1800–1939. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *57*(1), 41–66. Retrieved from https://doi.org/10.1080/01615440.2024.2368458 doi: 10.1080/01615440.2024.2368458

Dahl, C. M., Johansen, T., & Vedel, C. (2026). *Breaking the hisco barrier: Automatic occupational standardization with occcanine.* Retrieved from https://arxiv.org/abs/2402.13604

Egami, N., Hinck, M., Stewart, B., & Wei, H. (2023). Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. *Advances in Neural Information Processing Systems*, *36*, 68589–68601.

Egami, N., Hinck, M., Stewart, B. M., & Wei, H. (2024). Using large language model annotations for the social sciences: A general framework of using predicted variables in downstream analyses. *Preprint from November*, *17*, 2024.

Giommoni, T., Loumeau, G., & Tabellini, M. (2026, February). *Extractive taxation and the french revolution* (Working Paper No. 34816). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w34816 doi: 10.3386/w34816

Gorin, C., Heblich, S., & Zylberberg, Y. (2025, June). *State of the art: Economic development through the lens of paintings* (Working Paper No. 33976). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w33976 doi: 10.3386/w33976

Koschnick, J. (2025). *Did a feedback mechanism between propositional and prescriptive knowledge create modern growth?* (arXiv:2512.16587)

Liu, H., Papanikolaou, D., Schmidt, L. D., & Seegmiller, B. (2025). *Technology and labor markets: Past, present, and future; evidence from two centuries of innovation* (Tech. Rep.). National Bureau of Economic Research.

Ludwig, J., Mullainathan, S., & Rambachan, A. (2025). *Large language models: An applied econometric framework* (Working Paper No. 33344). National Bureau of Economic Research.

Paker, M. M., Stephenson, J. Z., & Wallis, P. (2025, June). *Predictive modeling the past* (Economic History Working Papers No. 379). London, UK: London School of Economics and Political Science, Economic History Department. Retrieved from https://researchonline.lse.ac.uk/id/eprint/128852/1/WP-379.pdf

Silcock, E., Arora, A., D'Amico-Wong, L., & Dell, M. (2024). Newswire: A large-scale structured database of a century of historical news. In A. Globerson et al. (Eds.), *Advances in neural information processing systems* (Vol. 37, pp. 49768–49779). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2024/file/58f52a01a516609336da78ff17e6f81f-Paper-Datasets_and_Benchmarks_Track.pdf doi: 10.52202/079017-1575