Traditional Statistics and Machine Learning in Social Network Analysis: A Comparative Reanalysis of Social Network Data on Energy Transition Decisions

Mart Verhoog
Marketing and Communication Department
IU International University of Applied Sciences
Cologne, Germany
e-mail: mart.verhoog@iu.org

Abstract - **The goal of this idea contribution is to provide a systematic head-to-head comparison of regression-based inference and Machine Learning (ML) prediction in applied Social Network Analysis (SNA) for energy transition research, addressing a gap that has not yet been explored. The problem is relevant because methodological choices affect how actor influence and decision-making are interpreted in networked household energy-efficient refurbishments. While regression models offer explanatory clarity, ML models often deliver higher predictive accuracy; yet their joint evaluation in this domain remains missing. This study proposes a structured pipeline combining regression baselines with ML models, such as Random Forests, Support Vector Machines (SVMs), and Gradient Boosting. Model performance will be evaluated using R² and the Receiver Operating Characteristic – Area Under the Curve (ROC-AUC), while interpretability will be assessed through SHapley Additive exPlanations (SHAP) values. The expected outcome is a sharper understanding of trade-offs and complementarities between inference and prediction in energy transition networks, informing methodological integration in computational social science.**

Keywords - Social Network Analysis; Machine Learning; Traditional Statistics; Comparative Methods; Interpretability and Prediction

I. Introduction

Statistical reasoning has long underpinned empirical social science. Yet today, the term “artificial intelligence” is often used loosely, conflating traditional methods with data-driven techniques simply because they run in complex software environments. This paper critically engages with that trend by comparing traditional statistics and ML approaches within a shared empirical context - Social Network Analysis (SNA) of decision-making in energy-efficient home refurbishments.

While statistical modeling prioritizes explanatory clarity and hypothesis testing, ML focuses on pattern detection and predictive power. Recent studies (e.g., Hossain [1], Sakib et al. [2]) have shown how ML methods like LASSO, Random Forests, and SVMs improve predictive accuracy, especially in complex or high-dimensional settings. However, to our knowledge, no prior work has conducted a head-to-head comparison of regression-based inference and ML-based prediction in applied SNA of energy transition networks. This study seeks to fill that gap by re-analyzing a large, previously published dataset using both methodological paradigms.

The remainder of this idea contribution first introduces the methodological paradigms (Section II), then outlines the dataset and empirical frame (Section III), presents the research question (Section III), and concludes with expected contributions and next steps (Section IV).

II. Traditional Statistics and Machine Learning: Definitions and Tensions

A. Traditional Statistical Inference

Methods like linear and logistic regression rely on assumptions, such as linearity, homoscedasticity, and independence. These methods enable interpretability and quantification of uncertainty—key strengths in hypothesis-driven research.

B. Machine Learning and Predictive Modeling

ML models like Random Forests and SVMs are assumption-light and often outperform traditional models in predictive contexts. Though often less interpretable, new tools (e.g., SHAP values) are improving transparency.

C. Convergence and Complementarity

Techniques like regularized regression (LASSO, Ridge) and decision trees bridge the gap between interpretability and flexibility. These “hybrid” models illustrate growing convergence.

D. Prior Comparative Research

Studies in psychology, epidemiology, and sociology (e.g., Jang & Lee [3]; Di Franco & Santurro [4]) show ML often provides superior predictive power while traditional models offer theoretical alignment. Comparable SNA studies remain scarce - underscoring this study’s relevance.

III. Social Network Analysis: The Empirical Frame

A. Dataset and Context

The analysis draws on a dataset from Verhoog (2017) covering approximately 700 cases of household decisions regarding energy-efficient refurbishment. The data capture not only household characteristics but also the role of professional stakeholders, such as building merchants, engineers, energy consultants, and financial institutions. Alongside these actor variables, the dataset includes the technical and efficiency status quo of the dwelling and the homeowners' attitudes toward refurbishment.

For instance, a household's decision to invest in energy-efficient refurbishment may depend not only on the building's technical condition and the homeowner's attitudes, but also on the involvement of professionals—whether consultants, engineers, merchants, or financial institutions—within the decision network. This combination of contextual, attitudinal, and network-related information provides a suitable empirical frame for comparing traditional statistical inference and Machine Learning (ML) prediction in applied Social Network Analysis (SNA)

B. *Feature Engineering*

Multiple SNA metrics (degree, betweenness, closeness centrality; network size; density; interaction intensity) will serve as predictors. These are complemented by contextual variables and preprocessed via normalization and encoding for compatibility with both methods.

C. *Modeling Approach*

The original analysis employed regression and factor methods. The reanalysis will apply ML models, such as Random Forests, Support Vector Machines (SVMs), and Gradient Boosting, with optional extensions to Neural Networks. The comparative pipeline will include preprocessing of SNA metrics (normalization and encoding), model training with cross-validation, hyperparameter tuning via grid search, and evaluation on a hold-out test set. Model performance will be evaluated using R^2 and the Receiver Operating Characteristic – Area Under the Curve (ROC-AUC), while interpretability will be assessed through SHapley Additive exPlanations (SHAP) values. Results from these ML models will be benchmarked directly against baseline regression models to enable a head-to-head comparison of inference and prediction.

D. *Research Question*

This study asks to what extent can Machine Learning uncover structural patterns in energy transition networks that traditional statistics may overlook, and under what conditions do their results converge or diverge?

This overarching question is addressed through three dimensions:

1. Network conditions – when do ML models achieve higher predictive accuracy than regression models?
2. Feature relevance – which network features emerge as most influential in ML compared to regression?
3. Interpretability – can tools, such as SHAP reconcile predictive accuracy with explanatory clarity in applied SNA for energy transition research?

IV. Expected Contribution and Next Steps

This idea contribution aligns with SIMUL 2025 themes by providing one of the first head-to-head comparisons of regression-based inference and ML-based prediction in applied SNA for energy transition research. In line with the research question, the study will clarify (i) under which network conditions ML offers superior predictive accuracy, (ii) how feature importance differs between ML and regression, and (iii) whether interpretability tools can bridge predictive and explanatory approaches. While the analysis centers on one empirical dataset, the approach is adaptable to other energy transition networks, which may enhance its broader relevance. These insights will highlight methodological synergies and trade-offs, advocating a hybrid perspective that combines theory-driven and data-driven approaches in studying decision-making in energy transition networks.

Immediate next steps are the preparation of the dataset and the construction of the modeling pipeline (Python: scikit-learn, NetworkX, SHAP). This will involve feature engineering of network metrics, cross-validation and hyperparameter tuning for multiple ML models, and systematic benchmarking against regression baselines to generate comparative performance and interpretability results. The findings will inform a full paper for the computational social science and organizational modeling community.

References

[1] A. Hossain, "Utilizing machine learning and causal graph approaches to address confounding factors in health science research: a scoping review", *F1000Research*, vol. 14, art. no. 129, 2025, doi: 10.12688/f1000research.159632.1.

[2] S. Sakib et al., "Comparative analysis of machine learning algorithms used for translating aptamer-antigen binding kinetic profiles to diagnostic decisions", *ACS Sensors*, vol. 10, no. 2, pp. 907–920, Feb. 2025, doi: 10.1021/acssensors.4c02682.

[3] D. Jang and B. Lee, "When machine learning meets social science: A comparative study of ordinary least square, stochastic gradient descent, and support vector regression for exploring the determinants of behavioral intentions to tuberculosis screening", *Asian Communication Research*, vol. 19, no. 3, pp. 101–118, Dec. 2022, doi: 10.20879/acr.2022.19.3.101.

[4] A. Di Franco and M. Santurro, "From big data to machinel learning: an empirical application to the sociology of health", *Athens Journal of Social Sciences*, vol. 10, no. 1, pp. 33–50, 2023.

[5] M. Verhoog, *Controlling Actors and Decisions in Construction Networks*, Springer Gabler, Wiesbaden, 2017.

ISBN: 978-1-68558-300-2