A Machine Learning Analysis of Social Capital in the United States: Spatial Structure, Temporal Persistence, and Key Predictors (1990–2014)

Wanning Kang Yusha Bai Britney Zhu
Stony Brook University, NY, USA
{wanning.kang, yusha.bai, britney.zhu}@stonybrook.edu

Abstract

The present study examines long-term trends and determinants of social capital across U.S. counties from 1990 to 2014. We utilize publicly available data provided by the Northeast Regional Center for Rural Development to create a harmonized panel dataset and evaluate national, regional, and temporal patterns in community trust and civic engagement. We compare Elastic Net regression models, random forests models, and XGboost models as methods to identify the socioeconomic factors most strongly associated with social capital. High temporal persistence of social capital, marked regional disparities, and a mild national decline over the period of 24 years are revealed. Both random forest and XGBoost capture meaningful nonlinear relationships and yield strong predictive performance, identifying association density, voter turnout, and population size as key contributors. On the contrary, elastic Net-only with extensions that incorporate squared and interaction terms reveals very low predictive accuracy, suggesting that linear models are poorly suited for this domain.

Overall, the results point to the geographically structured and institutionally embedded nature of U.S. social capital and underscore the importance of organizational infrastructure and civic participation in maintaining community trust. These insights inform potential policy interventions aimed at strengthening local institutions, supporting civic engagement, and enhancing social cohesion.

Keywords — social capital, machine learning, temporal persistence, feature importance, spatial analysis

1. Introduction

Social capital, understood as a combination of networks, trust, civic engagement, and social cohesion, has a foundational impact on community well-being, economic development, and political stability. Scholars and policymakers during the last several decades have voiced concerns that social trust has gradually weakened in the United States.Robert Putnam's Bowling Alone [1] famously documented the decline of civic engagement, volunteerism, and institutional participation. Yet, most studies on social capital evaluate it cross-sectionally or at one point in time, leaving many important questions about temporal dynamics and long-run predictors unanswered.

This paper offers a data-driven, comprehensive analysis of social capital across U.S. counties from 1990 to 2014. Using county-level indicators from the Northeast Regional Center for Rural Development, we integrate five benchmark years-1990, 1997, 2005, 2009, and 2014-into a single unified panel dataset. We analyze national trends, regional disparities, temporal persistence, and the socioeconomic determinants of social capital by using a suite of machine learning models that include Elastic Net regression, Random Forest, and XGBoost. [1–3]

Our key research questions are:

1. How has social capital changed over the period 1990–2014?
2. What socioeconomic factors most strongly predict social capital levels?
3. To what extent does social capital exhibit spatial clustering and temporal persistence?
4. Can a nonlinear machine learning model reliably capture the determinants of social capital?

The work aims to deepen the empirical knowledge in the study of social trust, detect its structural determinants, and provide input for public policy towards enhancing social cohesion.

2. Data and Methods

2.1. Dataset

Data were obtained from the Northeast Regional Center for Rural Development (NERCRD)'s county-level measures of social capital.[1] We use five benchmark years: 1990, 1997,

[1] See https://nercrd.psu.edu/data-resources/county-level-measure-of-social-capital/.

2005, 2009, and 2014. The combined panel includes county-level measures of:

- **relig**: density of religious organizations,
- **civic**: density of civic organizations,
- **assn**: association density,
- **pvote**: presidential voter turnout,
- **respn**: census response rate,
- **nccs**: nonprofit organizations,
- **pop**: population size,
- **sk**: Social Capital Index.

All variables were standardized and merged into a longitudinal dataset suitable for machine learning analysis.

3. Framework

We evaluate three predictive models:

3.1. Elastic Net Regression

Elastic Net combines the Lasso (L1) and Ridge (L2), hence it balances feature selection and multicollinearity control. We test:

- base linear features
- augmented features: squared terms, interaction terms

With these modifications, however, Elastic Net still obtains rather poor predictive performance, suggesting that the underlying relationships between socioeconomic variables and social capital are nonlinear, and not well captured by a linear model.

3.2. Random Forest Regression

A Random Forest regressor was used to model the Social Capital Index as a nonlinear function of the above predictors. Random Forest is well-suited for this task due to:

- Its robustness to multicollinearity
- Ability to model nonlinear interactions
- High predictive accuracy
- Built-in measures of feature importance

Model evaluation utilized train-test splitting, reporting R², Mean Squared Error (MSE), and Mean Absolute Error (MAE).

3.3. XGBoost Regression

XGBoost is a gradient-boosted decision tree algorithm optimized for high predictive accuracy and efficient handling of nonlinear relationships. Compared with Random Forest, XGBoost builds trees sequentially and focuses on reducing residual errors, making it well-suited for capturing complex patterns in socioeconomic data. Key advantages include:

- ability to model nonlinearities and interaction effects;
- Built-in regularization to prevent overfitting
- Effective handling of multicollinearity and heterogeneous feature scales
- Strong predictive performance on structured tabular data

The model is trained using pre-2014 observations and evaluated using R², Mean Squared Error (MSE), and Mean Absolute Error (MAE), consistent with the procedures applied to the other models.

4. Results

4.1. Model Performance Comparison

Table 1 summarizes train and test performance for the three models.

Table 1. Model performance comparison.

Model	Train R^2	Test R^2	Notes
Elastic Net (base)	0.0622	0.0525	underfits
Elastic Net (augmented)	0.0796	0.0731	little gain
Random Forest	0.9675	0.8837	strong fit
XGBoost	0.9859	0.8881	best overall

4.1.1. Elastic Net Summary

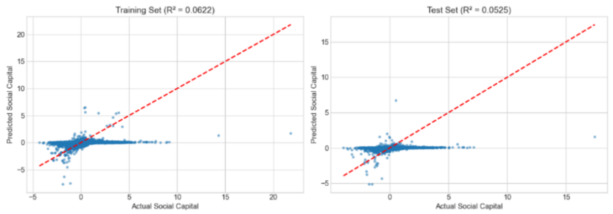

Figure 1.1. Actual vs. Predicted Social Capital (Training and Test Sets) of the Elastic Net (base) Model.

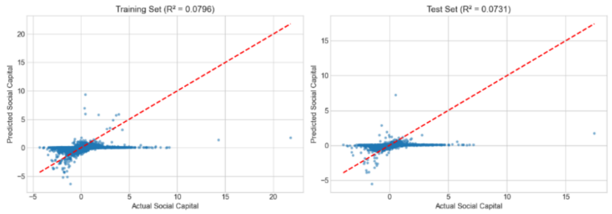

Figure 1.2. Actual vs. Predicted Social Capital (Training and Test Sets) of the Elastic Net (augmented) Model.

The Elastic Net model performed significantly worse than expected. Even after augmenting features:

- squared terms (e.g., $civic^2$, $pvote^2$)
- Interaction terms (e.g., $civic * pvote$),

The test R² remained low. This shows that social capital is not able to be explained by either linear or polynomial effects alone and nonlinear machine-learning models fit better.

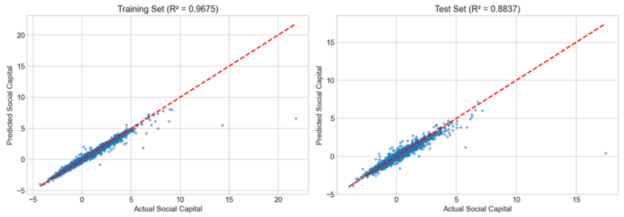

Figure 1.3. Actual vs. Predicted Social Capital (Training and Test Sets) of Random Forest

4.1.2. Random Forest Summary

The Random Forest model demonstrates excellent fit:

- Training R² = 0.9675
- Test R² = 0.8837

The close alignment of predicted and actual values indicates that the model captures the main structural determinants of social capital with high accuracy. This confirms the viability of nonlinear machine learning methods for social-science prediction tasks.

4.1.3. XGBoost Summary

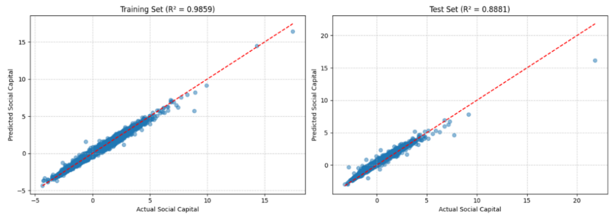

Figure 1.4. Actual vs. Predicted Social Capital (Training and Test Sets) of the XGBoost Model

The XGBoost model delivers the strongest overall performance among the three approaches.

- Training R² = 0.9859
- Test R² = 0.8881

The tight clustering of predicted values around the 45-degree line shows that XGBoost captures complex nonlinear relationships with exceptional accuracy. Its boosted tree structure allows the model to learn subtle interactions that both linear models and single-tree ensembles cannot fully represent. These results highlight XGBoost as a highly effective method for predicting social capital, providing both high predictive precision and strong generalization to out-of-sample years.

4.2. Correlation Structure

Key observations:

- Strong correlations exist among relig, civic, nccs, and pop ($r > 0.85$).
- Correlations between most predictors and sk are modest (-0.13 to $+0.01$)

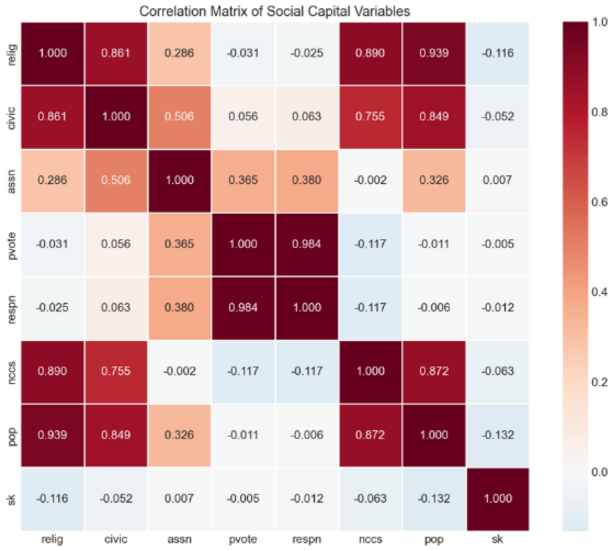

Figure 2. Correlation Matrix of Social Capital Variables

- The weakness of pairwise correlations with sk suggests that social capital arises from multivariable interactions, not single-variable effects—justifying the use of Random Forests.

This explains why Elastic Net fails: **the outcome is not driven by single variables linearly, but by complex multivariate interactions,** which tree-based models capture more effectively.

4.3. Time-Varying Correlations

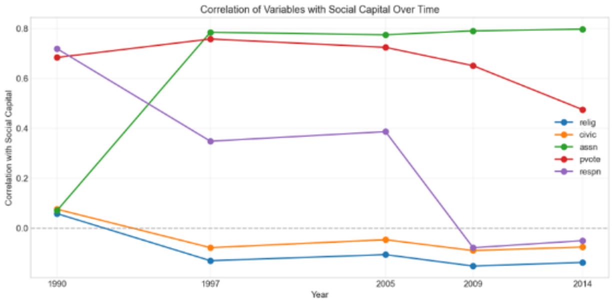

Figure 3. Correlation of Variables with Social Capital Over Time

Findings:

- assn maintains the strongest and most stable association with social capital (0.78–0.81)
- pvote shows a declining relationship with sk over time.
- Variability in relig and civic correlations indicates temporal sensitivity to sociopolitical contexts.

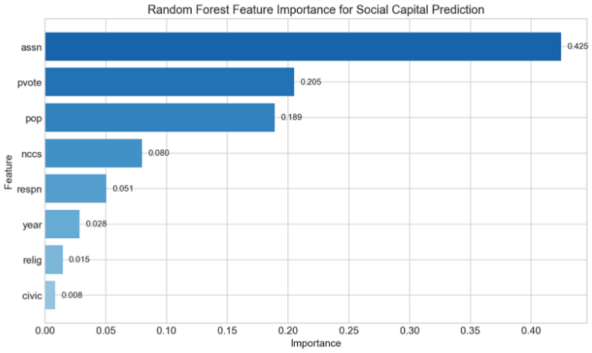

Figure 4.1. Random Forest Feature Importance for Social Capital Prediction

Table 2. Random Forest feature importance.

Feature	Importance
assn	0.425
pvote	0.205
pop	0.189
nccs	0.080
respn	0.051
year	0.028
relig	0.015
civic	0.008

4.4. Feature Importance

4.4.1. Random Forest Feature Importance

Association density accounts for 42.5% of the predictive power—more than double any other predictor—highlighting its central role in community-level social capital formation.

4.4.2. Elastic Net Feature Importance

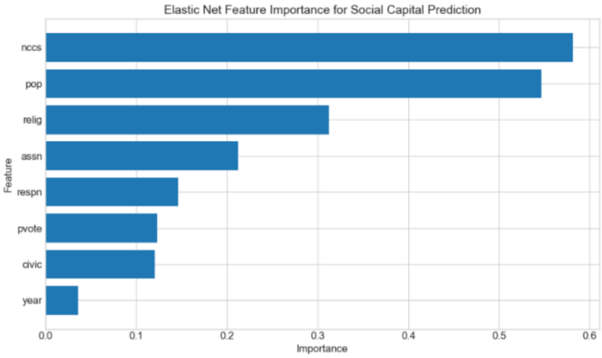

Figure 4.2. Elastic Net (Base) Feature Importance Plot

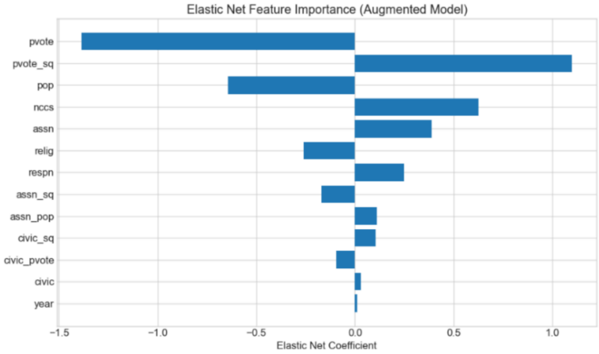

Figure 4.3. Elastic Net (Augmented) Feature Importance Plot

Two visualizations for feature importances were produced to help understand the Elastic Net model. The results of the augmented model (Figure 4.3) showed that it has a substantial quadratic effect since the values of the coefficients of pivot and pvote_sq were extremely large in magnitude but with opposite signs. However, despite adding a quadratic term, Elastic Net, which this model uses, performed poorly in terms of prediction with a test R of about 0.05. The simple model (Figure 4.2) found nonlinear patterns with more stable coefficients, in which nonprofit density (nccs), population (pop), and religious organizations (relig) were most closely related. However, this model also performed poorly, further illustrating that Elastic Net is not a proper modeling technique for this data set over that of either Random Forest or XGBoost.

4.4.3. XGBoost Feature Importance

Because XGBoost builds trees sequentially and optimizes residual reduction at each step, its importance scores reflect variables that consistently improve model fit across boosted trees.

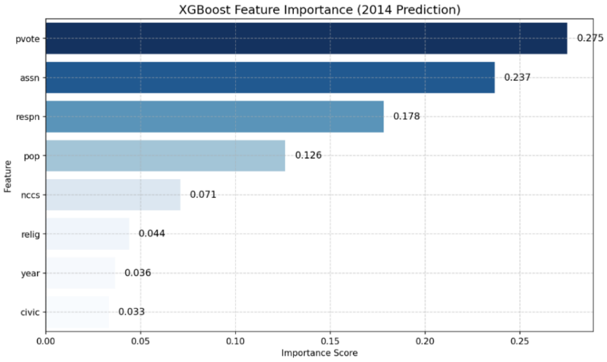

Figure 4.4. XGBoost feature importance scores.

The results show that **voter turnout** is the most influential determinant of social capital in the boosted-tree framework, followed closely by **association density**.

XGBoost also assigns substantial importance to **census response rates** and **population size**, indicating that it captures interaction effects and nonlinear influence patterns not

Table 3. XGBoost feature importance.

Feature	Importance
pvote	0.275
assn	0.237
respn	0.178
pop	0.126
nccs	0.071
relig	0.044
year	0.036
civic	0.033

visible in linear models.

These findings complement the Random Forest results and reinforce the evidence that organizational density and political participation are key structural components underlying community-level social capital.

4.4.4. SHAP-Based Model Interpretability (XGBoost)

To further interpret the nonlinear mechanisms captured by XGBoost, we apply SHAP (SHapley Additive exPlanations) to quantify each predictor's marginal contribution to the Social Capital Index. SHAP provides both global and local interpretability by decomposing individual predictions into additive feature attributions.

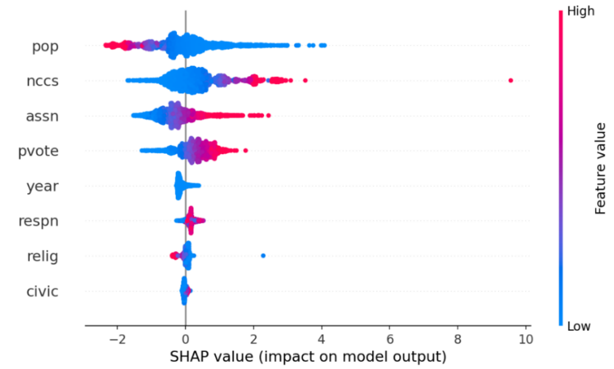

Figure 4.5. SHAP summary (beeswarm) plot for XGBoost model.

The beeswarm plot (Figure 4.5) reveals several consistent patterns. Higher levels of association density (assn), voter turnout (pvote), and nonprofit presence (nccs) systematically increase predicted social capital, confirming their dominant role in community civic infrastructure. In contrast, population size (pop) exhibits mixed effects: while larger counties may generate broader social networks, they may also experience reduced interpersonal cohesion depending on demographic heterogeneity. Variables such as relig and civic show comparatively modest influence, consistent with the model's gain-based importance rankings.

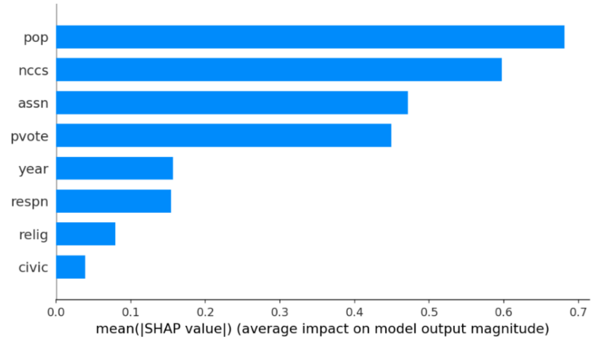

Figure 4.6. SHAP Global Importance (Mean |SHAP| Values)

Figure 4.6. Global SHAP importance (mean absolute SHAP values) for XGBoost.

The global SHAP importance plot (Figure 4.6) aligns closely with the model's internal feature importance, identifying organizational participation and institutional trust variables—particularly assn, pvote, and respn—as the primary determinants of social capital. The agreement between SHAP and model-based importance strengthens confidence that the XGBoost model captures meaningful structural relationships rather than fitting noise. Overall, the SHAP analysis demonstrates that XGBoost not only provides strong predictive performance but also yields interpretable, theory-consistent insights into the drivers of social capital across U.S. counties.

4.5. National Trends

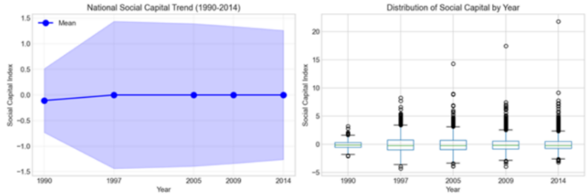

Figure 5. National Social Capital Index trend (1990–2014).

The national Social Capital Index(Figure 5):

- Begins slightly below zero in 1990
- Rises marginally by 1997
- Stabilizes from 2005 onward

Although fluctuations exist, the overall pattern indicates a flattening and mild decline in county-level social capital.

4.6. Regional Differences

State-level analysis (Figure 6) further highlights inequality:

- The Midwest consistently exhibits the highest social capital.
- The South maintains the lowest social capital, with values between -0.5 and -0.8.

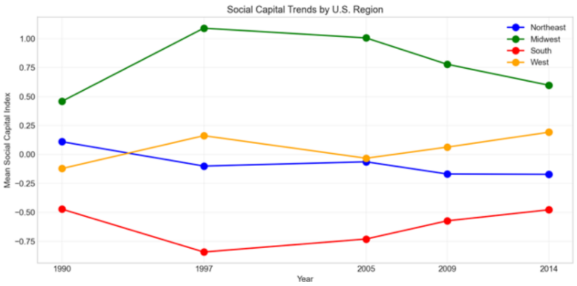

Figure 6. Social capital trends by U.S. region.

- Northeast and West show moderate and stable levels.

These disparities suggest long-standing cultural and institutional differences that shape regional social cohesion.

4.7. State-Level Inequality

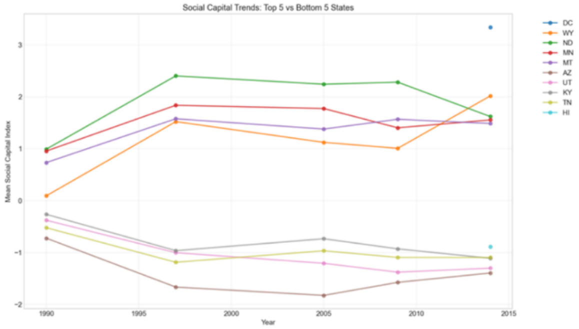

Figure 7. Social Capital Trends: Top 5 vs Bottom 5 States

High-social-capital states (ND, MN, MT) display persistent and elevated levels.Low-social-capital states (KY, TN, AZ) remain consistently below -1.0.

The divergence illustrates deep structural inequality, reinforcing the need for targeted policy intervention.

4.8. Temporal Persistence

4.8.1. Predictive Stability Across Years

To evaluate how well the model generalizes across different historical periods, we perform a panel time-series cross-validation.

For each benchmark year, XGBoost is trained on all previous years and tested on the next available year.

- The very poor fit for 1990 to 1997 ($R = -2.87$) reflects major structural changes in the definition and measurement of association density around 1990.
- After 1997, the predictive performance becomes highly stable, with R² consistently above 0.84, demonstrating that social capital is structurally persistent, and XGBoost can reliably predict future values once the measurement definitions stabilize.

Table 4. Panel CV performance of XGBoost across years.

Train period	Test year	RMSE	R^2
$\leq$ 1990	1997	2.83	-2.87
$\leq$ 1997	2005	0.46	0.89
$\leq$ 2005	2009	0.52	0.85
$\leq$ 2009	2014	0.42	0.89

This panel-CV analysis reinforces the conclusion that social capital evolves gradually, driven by slow-moving institutional and demographic factors, which tree-based models capture effectively.

4.8.2. Temporal Prediction Using Random Forest (2009 to 2014)

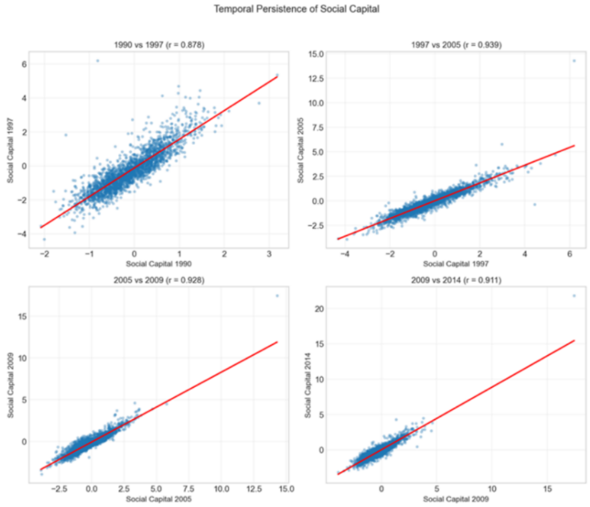

Figure 8.1. Temporal Persistence of Social Capital

Table 5. Inter-year Pearson correlations

Period	**r**
1990 $\rightarrow$ 1997	0.878
1997 $\rightarrow$ 2005	0.939
2005 $\rightarrow$ 2009	0.928
2009 $\rightarrow$ 2014	0.911

These exceptionally high correlations confirm that social capital is highly persistent, suggesting that structural features—economic, cultural, and institutional—dominate over short-term fluctuations.

4.8.3. Temporal Prediction with Elastic Net

To investigate whether or not future social capital can be forecasted from prior socio-economic indicators, we estimated a temporal Elastic Net model using features from 2009 to predict 2014 social capital for counties that appear in all years.

Unlike the cross-sectional Elastic Net model, which showed poor performance due to nonlinear relationships and complex interactions, the accuracy for the temporal Elastic Net model was substantially higher, reflecting the strong persistence of social capital over time.

The model used the following predictors from 2009: relig, civic, assn, pvote, respn, sk (2009), population, and nonprofit variables. Performance Indicators:

- R^2: 0.8832
- RMSE: 0.4634

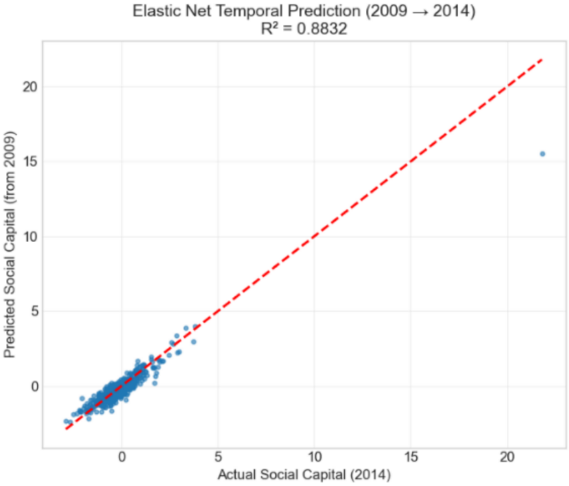

Figure 8.2. Elastic Net temporal prediction of 2014 Social Capital Index from 2009 covariates.

This strong performance evidences the fact that, though being unable to capture complex cross-sectional heterogeneity, Elastic Net is very effective when the prediction focuses on temporal changes due to the stable structural and institutional factors dominating year-to-year variation in social capital. Indeed, a scatterplot comparing actual vs. predicted 2014 values shows a near-linear relationship with little error (Figure 8.2).

4.9. Trends of Component Variables

- Religious organizations increase steadily.
- Civic organizations decline sharply.
- Association density collapsed after 1990 due to definitional or structural changes.
- Voter turnout and census responses also fall dramatically.
- Despite fluctuations in components, the composite Social Capital Index remains relatively stable—indicating compensatory effects among indicators.

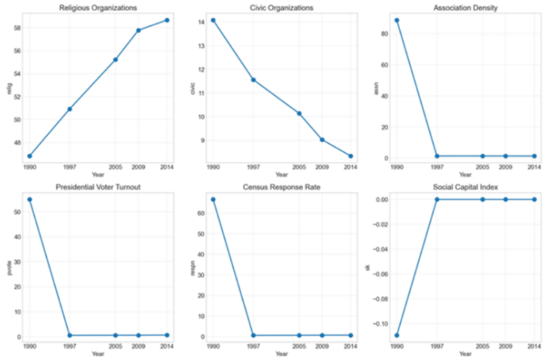

Figure 9. Trends of core variables: relig, civic, assn, pvote, respn, and the Social Capital Index.

5. Discussion

The findings of this study reveal several important insights into the structure and evolution of social capital in the United States:

Structural persistence. Social capital exhibits strong inter-year stability ($r > 0.87$) for all periods, suggesting that deep, long-term institutional and cultural factors dominate over short-term economic or political changes.

Regional and state-level inequality. Persistent differences between regions and states emphasize the role of geography, demographic patterns, and historical trajectories in shaping social trust.

Multidimensional determination. Weak simple correlations but strong Random Forest performance indicate that social capital depends on nonlinear interactions among multiple socioeconomic factors. XGBoost further reinforces this conclusion, as its strong predictive accuracy and SHAP-based interpretability highlight the importance of nonlinear effects and institutional participation variables in shaping social capital.

Decline in civic engagement. Sharp decreases in civic organizations, voter turnout, and census response rates may reflect weakening participatory norms, potentially contributing to the broader decline in trust and cohesion observed in national surveys.

Central Role of Association Density Association density emerges as the most important predictor, highlighting the significance of organizational infrastructure in sustaining social networks and trust.

6. Conclusion

This analysis offers a multi-model investigation of American social capital from 1990 through 2014. On a positive note, it appears that American social capital has seen relatively modest declines over the long term, offering a platform that suggests that deep structural roots, as opposed to recent events, play a key role in determining American social cohesion.

On all the models, it has been shown that:

- Even with squared and interaction terms, Elastic Net's performance remains poor, which suggests that a linear relationship cannot be a proper explanation for social capital.
- Random Forest and XGBoost provide significantly high predictive accuracy, verifying that nonlinear interaction and complex feature correlation are important in modeling social capital.
- In addition, XGBoost provides clear interpretability through SHAP values, revealing theoretically consistent mechanisms underlying social capital formation.
- Association density, voter turnout, and population size are found to be key determining factors.
- Results from temporal modeling reveal strong persistence of social capital, where past values of social capital strongly predict future values.
- There is evidence showing clear regional and state-level disparities in U.S. counties.
- There is a mild but notable national decline.
- Nonlinear models such as Random Forest and XGBoost show robust predictive performance.

These results taken together imply that a long-term, structural approach to policy that seeks to enhance civic institutions, community groups, and gross-roots networks that foster and sustain trust and participation would be most effective. Because social capital develops slowly and unevenly, place-based policies rather than national policies with broad-based impact are most likely to be effective. The results also underscore the structural and institutional nature of social capital and offer guidance for policymakers seeking to strengthen community cohesion through investments in civic infrastructure, local associations, and participatory institutions.

References

[1] Carl Boggs. Social capital and political fantasy: Robert putnam's "bowling alone". https://www.jstor.org/stable/657878, 2001. Accessed on JSTOR. 1

[2] Anil Rupasingha and Stephan J. Goetz. Social and political forces as determinants of poverty: A spatial analysis. *The Journal of Socio-Economics*, 36(4):650–671, 2007.

[3] Anil Rupasingha, Stephan J. Goetz, and David Freshwater. The production of social capital in us counties. *The Journal of Socio-Economics*, 35(1):83–101, 2006. 1