https://doi.org/10.26219/heal.aueb.9465

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

ΜΕΤΑΠΤΥΧΙΑΚΟ ΣΤΑΤΙΣΤΙΚΗ
MSc IN STATISTICS

Modeling and forecasting construction material time series using Dynamic Panel data models

By
Lazaros Nikolaos

A THESIS
Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
May 2025

https://doi.org/10.26219/heal.aueb.9465

https://doi.org/10.26219/heal.aueb.9465

Μοντελοποίηση και πρόβλεψη χρονοσειρών δομικών υλικών με τη χρήση μοντέλων δυναμικών πάνελ

Νικόλαος Ιωάννης Λάζαρος

ΔΙΑΤΡΙΒΗ
Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Διπλώματος Μεταπτυχιακών Σπουδών στη Στατιστική

Αθήνα
Μάϊος 2025

https://doi.org/10.26219/heal.aueb.9465

https://doi.org/10.26219/heal.aueb.9465

DEDICATION

To my parents and my brother,
Your support, love, and encouragement have been my greatest strength. Thank you.

https://doi.org/10.26219/heal.aueb.9465

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Professor I.Vrontos for his guidance, constant support and readiness to assist throughout every stage of this thesis. His expertise have been instrumental in the completion of this work.

https://doi.org/10.26219/heal.aueb.9465

https://doi.org/10.26219/heal.aueb.9465

ABSTRACT

This thesis investigates the modeling and forecasting of U.S. cement import using dynamic econometric methods, focusing on short- and long-term trends over four central states: California, Florida, North Carolina, and Texas. Our interest stems from the fact that construction materials have always played a critical role in infrastructure planning, while their prices have been characterized by considerable volatility. Our approach is based on the construction of a panel dataset covering the period 2008–2022 and the use of time series and panel data techniques. The use of import volume instead of prices is chosen in this case due to address data availability issues. Besides, imports dynamics reflect the fluctuations of construction activity and, therefore, those of domestic demand and prices of materials.

After transforming the data to ensure stationarity, Lasso regression is used for the selection of state-specific predictors. These variables are incorporated into ARIMAX models to improve forecasting performance. We evaluate each model's predictive ability using metrics such as RMSE, MAE, and Mean Error, with ARIMAX generally producing strong results, particularly in California and Texas.

In addition, a number of panel modeling approaches, including Pooled OLS, Fixed Effects, and Seemingly Unrelated Regressions (SUR) are explored. While these models offer interpretive insights, their predictive performance varies across states, with the SUR model examined primarily as an exploratory framework for capturing potential cross-sectional dependencies.

The results highlight the effectiveness of combining variable selection and ARIMA modeling for state-level forecasting of construction imports. They also emphasize the importance of adapting model structures to regional characteristics and data behavior, offering guidance for both practitioners and policymakers in the construction and economic planning sectors.

https://doi.org/10.26219/heal.aueb.9465

https://doi.org/10.26219/heal.aueb.9465

ΠΕΡΙΛΗΨΗ

Η παρούσα διπλωματική εργασία διερευνά τη μοντελοποίηση και πρόβλεψη των τιμών εισαγωγών τσιμέντου στις Ηνωμένες Πολιτείες, με χρήση δυναμικών οικονομετρικών μεθόδων και έμφαση στις βραχυπρόθεσμες και μακροπρόθεσμες τάσεις σε τέσσερις βασικές πολιτείες: Καλιφόρνια, Φλόριντα, Βόρεια Καρολίνα και Τέξας. Το ενδιαφέρον εστιάζεται στη σημασία των δομικών υλικών για τον σχεδιασμό υποδομών και στη μεταβλητότητα των τιμών τους. Δημιουργείται ένα panel dataset για την περίοδο 2008–2022 και εφαρμόζονται τεχνικές χρονοσειρών και δεδομένων panel. Η χρήση του όγκου των εισαγωγών στη θέση των τιμών έχει επιλεγεί σε αυτή την περίπτωση για την αντιμετώπιση θεμάτων διαθεσιμότητας στοιχείων. Εξάλλου, η δυναμική των εισαγωγών αντανακλά τις διακυμάνσεις της κατασκευαστικής δραστηριότητας και συνεπώς, αυτές της εσωτερικής ζήτησης και των τιμών των υλικών.

Αρχικά εφαρμόζονται μετασχηματισμοί για τη διασφάλιση στασιμότητας των χρονοσειρών. Ακολουθεί η χρήση της μεθόδου Lasso για την επιλογή προβλεπτικών μεταβλητών ανά πολιτεία, οι οποίες ενσωματώνονται σε μοντέλα ARIMAX για τη βελτίωση της προγνωστικής ικανότητας. Η απόδοση αξιολογείται με δείκτες όπως RMSE, MAE και Μέσο Σφάλμα, με τα μοντέλα ARIMAX να παρουσιάζουν ιδιαίτερα ικανοποιητικά αποτελέσματα, ειδικά στην Καλιφόρνια και το Τέξας.

Παράλληλα εξετάζονται μοντέλα panel όπως τα Pooled OLS, Fixed Effects και Seemingly Unrelated Regressions (SUR). Αν και τα panel μοντέλα προσφέρουν χρήσιμα ερμηνευτικά στοιχεία, η προγνωστική τους απόδοση διαφοροποιείται μεταξύ των πολιτειών, με το SUR να αξιολογείται κυρίως ως διερευνητικό πλαίσιο.

Τα αποτελέσματα αναδεικνύουν την αποτελεσματικότητα του συνδυασμού επιλογής μεταβλητών με ARIMA για πρόβλεψη σε επίπεδο πολιτείας, καθώς και τη σημασία της προσαρμογής των μοντέλων στα χαρακτηριστικά της εκάστοτε περιοχής. Η μελέτη προσφέρει χρήσιμα συμπεράσματα για επαγγελματίες και φορείς χάραξης πολιτικής στον τομέα των κατασκευών και της οικονομικής πρόβλεψης.

https://doi.org/10.26219/heal.aueb.9465

https://doi.org/10.26219/heal.aueb.9465

Table of Contents

https://doi.org/10.26219/heal.aueb.9465

https://doi.org/10.26219/heal.aueb.9465

LIST OF TABLES

LIST OF FIGURES

https://doi.org/10.26219/heal.aueb.9465

https://doi.org/10.26219/heal.aueb.9465

1 Chapter 1: Introduction

1.1 Overview of the U.S. Construction Industry

The construction industry in the United States represents a cornerstone of its economy, contributing approximately 4% to the national GDP (VAPGDPC 2024), with an annual expenditure exceeding 1.6$ trillion as of recent years and, according to the latest data, reaching 2.1$ trillion for the year 2024 (TTLCONS 2024). Construction encompasses a wide spectrum of activities, from residential and commercial construction to infrastructure and industrial projects. As a labor-intensive sector with significant capital and material investments, construction heavily influences related industries, including manufacturing, transportation and energy. Such diversity reflects not only the industry's resilience but also its susceptibility to economic, geopolitical, and environmental factors. For instance, during economic downturns (COVID-19), reduced construction spending often mirrors broader economic contractions (Vrontos 2023).

1.2 Economic Volatility and the Price of Construction Materials

Construction material prices are among the most volatile cost elements within the industry. Materials like cement, steel, asphalt, and lumber are indispensable, often accounting for up to 50% of a project's total cost (Afolabi and Abimbola 2022). This volatility stems from a complex interplay of global supply chains, energy prices, geopolitical tensions, and environmental regulations (Miller 2023). Recent events such as the COVID-19 pandemic and the Russia-Ukraine conflict have heightened these fluctuations, leading to unprecedented spikes in material prices. For example, Ilbeigi (2017) states that the price of asphalt cement in the state of Georgia showed extreme short-term variability, demonstrating both seasonal and cyclical trends. Understanding these dynamics is crucial for stakeholders, from project managers and contractors to policymakers.

1.3 The Need for Accurate Forecasting

Forecasting construction material prices is pivotal for several reasons:

1. Cost Estimation Accuracy: Precise cost predictions enable stakeholders to budget effectively, minimizing financial risks associated with underestimation or overestimation (Kaur 2020)
2. Project Scheduling: The price of materials can influence project timelines. For instance, unexpected cost increases might lead to delays or scope reductions.
3. Investment Decision-Making: Developers and financiers rely on accurate forecasts to assess the viability of construction projects (Vrontos 2023).
4. Policy Formulation: Governments use price forecasts to allocate funding for public infrastructure projects and to implement policies addressing material shortages or surpluses (Afolabi and Abimbola 2022).

1.4 Methodologies in Forecasting Construction Material Prices

Modern advancements in forecasting methodologies have revolutionized the process of making material price predictions. Econometric models, machine learning algorithms, and time-series analyses have emerged as key tools. Techniques like Autoregressive Integrated Moving Average (ARIMA), regression analysis, and ensemble machine learning methods have demonstrated varying degrees of success in addressing the industry's unique challenges. Studies have shown that hybrid approaches, combining artificial intelligence with traditional statistical models, offer significant improvements in accuracy. Many of these methods will be presented in Chapter 2 of the Literature Review.

 https://doi.org/10.26219/heal.aueb.9465

1.5 Cement: A Case Study

For the sake of this study the focus will be on Cement as the primary construction material. Cement serves as a bellwether for material price trends due to its ubiquity in construction. The U.S., despite being a major producer globally, also imports significant quantities to meet its substatial domestic demand, making cement prices in the domestic market particularly sensitive to global market conditions.. As a result, the U.S. has attained the position of being the largest cement importer in the world ($2.28B as of 2022 (OEC 2022)). Given that cement consumption is directly linked to the course of the construction industry, the additional imported quantities reveal the ups and downs of the U.S. construction industry. Increasing imports reflect rising domestic demand and prices. The analysis aims to effectively model and predict U.S. cement imported quantity across time and identify important predictor variables that affect their productivity and market price.

1.6 Structure of the Thesis

This thesis is organized into five chapters. In Chapter 1 the research topic is introduced, outlining the importance of modelling and forecasting construction material prices and imports, cement in particular, and identifying the research objectives and scope. Chapter 2 will provide a comprehensive review of the relevant literature explored, starting with an overview of panel data and dynamic panel data models, followed by a discussion of existing forecasting methods and the research gap this study seeks to address. Chapter 3 will describe the dataset that will be used for this analysis, detailing its characteristics, sources and some preprocessing steps. Chapter 4 will include the outline of the methodological framework, including dynamic panel data models and some alternative techniques that will be applied to address the challenges posed by the specific structure of our data. Finally, Chapter 5 will conclude the study, summarizing the key insights, implications that this study faced and proposals for future research.

https://doi.org/10.26219/heal.aueb.9465

2 Chapter 2: Literature Review

2.1 Introduction

In this chapter of the thesis the focus is on the review of past research on modeling and forecasting the cost and imports of construction materials as well as the importance of dynamic panel data models in capturing key features of our data.

2.2 Review of Panel Data

Panel data also referred to as longitudinal data contain observations of different cross sections across time. Some examples that could make up panel data series include countries, firms, individuals, or demographic groups. Panel data, like Time series data contains observations collected at a regular frequency chronologically. Also like Cross-Sectional data, Panel data contains observations across a collection of individuals. This dual structure makes panel data a powerful tool for understanding dynamic relationships and heterogeneity across entities while leveraging the temporal dimension for richer analysis (Hsiao 2014). There are numerous variants of the panel data models that will be presented.

2.3 Structure of Panel Data

Panel data can be categorized into two types based on their structure (Wooldridge 2012): First, the Balanced Panels, where all cross-sectional units are observed across the same time periods, which simplifies the estimation methods considerably as no observations are missing. Secondly, the Unbalanced panels, where there are missing observations of different units for specific time periods. These are more common in real-world data and can complicate estimation methods.

2.4 Advantages of Panel Data

Panel data have significant advantages over conventional cross-sectional or time-series datasets. As highlighted by Hsiao (2014), these include several advantages:

- To begin with, panel data provide a large number of observations, thereby enhancing the degrees of freedom and reducing collinearity among explanatory variables, which makes them efficient for econometric parameter estimation.

- Also their ability to combine interindividual differences with intraindividual dynamics allows for the construction of more realistic behavioral hypotheses, enabling researchers to capture complex relationships and changes over time.

- In addition, panel data facilitate the uncovering of dynamic relationships, such as lagged effects and behavioral inertia[1], which are critical for modeling dynamic economic processes.

- They also control for unobserved heterogeneity through fixed-effects or random-effects models, thus reducing omitted variable bias and yielding more reliable estimates.

[1] The tendency of economic variables to exhibit resistance to sudden changes.

- Furthermore, by pooling observations across entities and time, panel data improve the accuracy of predictions for individual outcomes, particularly when behaviors are conditionally similar.

- Finally, panel data simplify statistical inference, as the central limit theorem applied across cross-sectional units ensures asymptotic normality even for nonstationary data, making hypothesis testing more robust and intuitive.

2.5 Disadvantages of Panel Data

Above, the numerous advantages that panel data offer were discussed, without considering the challenges they present:

- One notable issue is the risk of heterogeneity bias, which arises when differences across individuals or entities are not adequately accounted for in the model.

- Additionally, while fixed-effects models are helpful in controlling for unobserved heterogeneity, they remove time-invariant variables from the analysis, making it impossible to estimate their impact.

- In the context of dynamic panel data, which include lagged dependent variables, models often suffer from Nickell bias, particularly in datasets with a short time dimension.

- Furthermore, cross-sectional dependence, where units in the dataset influence one another, can lead to misleading conclusions if ignored.

- Measurement errors and missing data can further complicate the analysis, while nonstationarity in time-series components can result in spurious relationships if not properly addressed.

Despite these challenges, with rigorous testing and careful modeling, panel data remain a powerful tool for understanding dynamic economic processes.

2.6 General Panel Data Models

Panel data models analyze the relationship between a dependent variable and one or more expanatory variables while accounting for unobserved heterogeneity. The general structure of a panel data regression model is the following:

$$y_{it} = \alpha_i + \beta_{1i}\, x_{1,it} + \beta_{2i}\, x_{2,it} + \cdots + \beta_{k\iota}\, x_{k,it} + \epsilon_{it}$$

Where:

- y_{it}: Dependent variable for entity i at time t,
- βi: k x 1 vector of slope coefficient
- x_{it}: k x 1 vector of explanatory variables for entity i at time t
- α_i: Entity-specific effect capturing unobserver time-invariant heterogeneity,
- ϵ_{it}: Error term.

The entity-specific effect (α_i) is crucial, as it allows the model to account for individual differences across entities, which would otherwise result in omitted variable bias.

https://doi.org/10.26219/heal.aueb.9465

2.7 Types of Panel Data Models

2.7.1 Pooled Ordinary Least Squares (Pooled OLS):

- Assumes homogeneity across entities ($\alpha_i = \alpha \; for \; all \; i$) and pools all observations together, ignoring individual-specific effects.
- Formula:

$$y_{it} = \alpha + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \cdots + \beta_k x_{k,it} + \epsilon_{it}$$

- y_{it}: Dependent variable for entity i at time t,
- β: k x 1 vector of slope coefficient (common among all entities-states)
- x_{it}: k x 1 vector of explanatory variables for entity i at time t
- α : Common intercept (same for all entities)
- ϵ_{it}: Error term.
- Limitations: Ignores heterogeneity across entities anf fails to account for individual-specific effects, often resulting in biased estimates.

2.7.2 Fixed Effects Model (FEM):

- Accounts for time-invariant unobserverved heterogeneity by allowing each entity to have its own intercept (α_i).
- Formula:

$$y_{it} = \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \cdots + \beta_k x_{k,it} + \epsilon_{it}$$

- y_{it}: Dependent variable for entity i at time t,
- β: k x 1 vector of slope coefficient (common among all entities-states)
- x_{it}: k x 1 vector of explanatory variables for entity i at time t
- α_i : Entity specific intercept
- ϵ_{it}: Error term.
- Estimation: Typically uses within-transformation to remove α_i:

$$y_{it} - \overline{y_i} = \beta_1(x_{1,it} - \bar{x}_{1,i}) + \beta_2(x_{2,it} - \bar{x}_{2,i}) + \cdots + \beta_k(x_{k,it} - \bar{x}_{k,i}) + (\epsilon_{it} - \overline{\epsilon_i})$$

  where $\overline{y_i}$ and $\overline{x_i}$ are entity-specific means.
- Stenghts: Controls for unobserved heterogeneity correlated with regressors.
- Limitations: Cannot estimate the effects of time-invariant variables.

2.7.3 Random Effects Model (REM):

- Assumes α_i is a random variable drawn from a distribution and uncorrelated with the regressors.
- Formula:

$$y_{it} = \alpha + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \cdots + \beta_k x_{k,it} + u_i + \epsilon_{it}$$

  Where $u_i \sim N(0, \sigma_u^2)$ represents the random effect which is constant over time.
- y_{it}: Dependent variable for entity i at time t,
- β: k x 1 vector of slope coefficient
- x_{it}: k x 1 vector of explanatory variables for entity i at time t
- α : Common intercept
- ϵ_{it}: Error term.
- Estimation: Uses Generalized Least Squares (GLS) for efficiency.
- Strengths: Allows for the inclusion of time-invariant variables.
- Limitations: Assumes no correlation between $\alpha_i$and regressors, which may not hold in practice.

https://doi.org/10.26219/heal.aueb.9465

2.7.4 Dynamic Panel Data Models

Dynamic panel data models extend static models by including one or more lagged dependent variables as explanatory variables. This allows the model to capture persistence (inertia), where past outcomes influence current behavior. The general form of a dynamic panel data model is:

$$y_{it} = \alpha_i + \rho_1 y_{i,t-1} + \rho_2 y_{i,t-2} + \rho_3 y_{i,t-3} + \rho_4 y_{i,t-4} + \beta_1 x_{1,it} + \beta_2 x_{2,it} \cdots + \beta_k x_{k,it} + \epsilon_{it}$$

Where:

- $y_{i,t-1}$: Lagged dependent variable.
- ρ: Coefficient representing the effect of past outcomes on current outcomes.

Dynamic models address important phenomena such as delayed responses in economic systems. However, the unclusion of $y_{i,t-1}$ introduces endogeneity, as $y_{i,t-1}$ is likely correlated with ϵ_{it} leading to biased estimates with OLS or standard fixed-effects estimator. This problem is known as the Nickell Bias. The magnitude of the bias diminishes as T increases, making in less problematic for panels with large time periods. However, for small-T panels the bias is substantial and cannot be ignored. Nickel emphasized the need for alternative stimation methods to address this issue which mitigated the bias by using lagged variables as instruments (Nickell 1981). His work provided the foundation for undestanding the limitation of traditional fixed-effects estimators in dynamic models and laid the groundwork for robust alternatives such as the Difference GMM by Arellano and Bond (Arellano and Bond 1991) and System GMM by Blundell and Bond (Blundell and Bond 1998).

2.8 Forecasting and Modeling Construction Material Prices

The accurate forecasting and modeling of construction material prices are critical for project planning and cost control. A variety of approaches appear in the literature, ranging from traditional econometric methods to machine learning models. This section will review some of the key studies that have addressed this issue, focusing on the techniques employed and their suitability for capturing both short-term and long-term price dynamics.

Elmousalami(2019) uses several quantitative models and techniques for construction project cost estimation such as Fuzzy Logic (FL) where he utilizes imprecise information to generate decision-making models similar to human reasoning, Artificial Neural Networks (ANN) that simulate the way the human brain analyzes and processes information are best for modeling complex patterns and relationships that may be difficult to explicity program. Regression models are also used to estimate the relationships among variables and are used as a baseline in comparison to more complex AI models. Hashemi (2020) also uses machine learning techniques for cost estimation and prediction in construction projects. On the quantitative approaches he uses statistical methods such as Regression analysis, analogous models where he uses similar construction projects to estimate costs for new projects and analytic models where he analyzes the cost of each project component. Jiang et al examines the construction cost by using time series data to predict the Construction Cost Index (CCI) and its influencing factors. The models being utilized are the Simple Moving Average (SMA), Exponential Smoothing (ES), and Autoregressive Integrated Moving Average (ARIMA) the latter being highlighted as the most effective model due to its high model-fit correlation for CCI prediction (Jiang, Awaitey and Xie 2022). A worthy mention is the methodology presented by Ng, Cheung, Lam and Skitmore (2000) utilizing Multivariate Discriminant Analysis (MDA) to forecast direction changes in the Tender Price Index (TPI). The report focuses on MDA efficacy in volatile economic conditions, by accurately classifying TPI movement into upward, constant, or downward. Using eight leading economic indicators they show that MDA can surpass traditional regression models. About short term fluctuation Ilbeigi (2017) utilized various advanced time-series forecasting models to address the challenge of predicting asphalt cement price fluctuations. Instead of the common practises for cost estimation of construction

https://doi.org/10.26219/heal.aueb.9465

projects, like adding a fixed percentage of the estimated total cost of cement, inflating the estimated cost of cement or conducting Monte Carlo simulation to analyze the uncertainty of the upcoming cement prices he works with the autocorrelation that exists in the previous cement prices. He implements Holt Exponential Smooting (ES), Holt-Winters (ES), Autoregressive Integrated Moving Average (ARIMA) and seasonal ARIMA. Afolabi and Abimbola (2022) created a web-based system that integrates machinne learning to predict future cement prices. Using historical data they trained a model using linear regression that demonstrates 80% accuracy on predicting the future cement prices.

2.9 Forecasting and Modeling Using Panel Data

There are various methodologies currently used in forecasting and modeling construction material prices, such as time-series analysis, regression models, and machine learning approaches. In this part of the literature review we are going to shift our focus on the dynamic panel data models and their application in economic forecasting. As presented in the paper by Kholodilin (2008) which focuses on the regional GDP growth are across the 16 German states. In this paper Kholodilin employs dynamic panel data models to forecast annual growth rates of GDP for each German state in a way to account for both spatial dependence and individual variations across regions. Each region has difference economic performance that is influenced by its geographical neighbors due to factors like trade flows and labor mobility. In the study Autoregressive Models (AR) are estimated for each region to account for past values influence on current GDP growth rates. The AR models are also used as benchmark models. Next Pooled Panel Model (POLS) is applied which imposes homogeneity across all regions, assuming a common intercept and slope for all regions, which simplifies the models but restricts regional specificity. To allow for individual intercepts for each region, accommodating inherent differences between regions while maintaining common slope coefficient the author uses Fixed Effects Model (FE). To account for the potential endogeneity of the lagged dependent variable used, Generalized Method of Moments (GMM) is used, which can provide more robust (unbiased) estimates than the Ordinary Least Squares (OLS) estimator.
Next the study by Matei (2024) explores the complex dynamics between government pandemic responses and economic growth across 105 countries from 2017 to 2021. So, we have both a lot of entities and time periods. Utilizing a non-linear panel analysis, specifically Panel Smooth Transition Regression (PSTR), Matei identifies how the effects of these measures vary significantly depending on a country's development level and institutional quality. The PSTR model captures the non-linear impact of pandemic policies, revealing that countries with higher institutional quality experience less negative impacts from stringent measures. Additionally, the Generalized Method of Moments (GMM) is employed to address endogeneity concerns, ensuring robustness in the findings.
Ben-Ahmed(2022) employs dynamic panel data models to analyze the effect of COVID-19 on the stock returns of digital companies. This study specifically explores how the pandemic influenced market behaviors and investment strategies, using monthly growth data of total confirmed cases and total death cases of COVID-19 to assess their impact on the financial performance of digital firms. To achieve that, the researchers use Dynamic panel regression to incorporate past stock returns and market-to-book ratios as explanatory variables, along with COVID-19 infection rates and mortality data. This allows them capture the persistence and temporal dynamic of stock performance in response to rapidly changing pandemic conditions. Also they use the Two-Step Generalized Method of Moments (GMM) to account for potential endogeneity problems, where explanatory variables are correlated with the error term. This method is effective in panels where individual effects and autocorrelation might bias the estimates as discussed in Kholodilin’s (2008) study. The methodologies used in this paper are advanced econometric techniques that can be applied to a range of economic and financial

https://doi.org/10.26219/heal.aueb.9465

research topics, including the forecasting of construction material prices in response to external economic shocks.

As presented on the next chapter the data will be in the format of a long panel. There will be several states with a large number of time periods, so the dataset is composed of long time periods (T) and few units (N) which are common in macroeconomic studies. The study by Labra and Torrecillas (2018) presents an in-depth examination of the methodological challenges and solutions in using dynamic panel data models for long time periods with a small number of observations. They particularly propose System GMM because it is advantageous in settings with long panels as it leverages additional moment conditions that can provide more efficient estimators, when the error terms are not strictly exogenous. There is also a discussion about strategies to combat the commom problem of too many instruments in dynamic panel models which can lead to overfitting and weak identification. By limiting the number of instruments or using only those that are most relevant (parsimonious instrumenting), researchers can improve the precision and validity of their models. In this study they also propose considering Seemingly Unrelated Regression (SUR) when dealing with multiple equations that could be estimated more efficiently if the cross-equation correlations are account for, which is typical in system with interrelated variables across different sectors or regions. Next up, the paper by Taeil (2023) delves into the analysis of how the COVID-19 pandemic has affected international trade dynamics among key global trading players using the Seemingly Unrelated Regression (SUR). SUR as discussed on the previous paper is more efficient than individual regressions due to its ability to account for the error correlation across different models. It is particularly apt as it allows for a nuanced analysis of interconnected trade activities.

Another study by Vrontos (2024) using a mix of linear and quantile predictive regression models methodically assesses how various macroeconomic and pandemic-specific indicator influence GDP trajectories across multiple countries. A Markov chain Monte Carlo (MCMC) stochastic search algorithm is utilized for model selection in order to explore model space and identify relevant predictors for the model. The use of Bayesian methods for model selection in the context of dynamic panel data provides a robust framework for dealing with model uncertainty. On the second part of this study Dynamic panel models are used to incorporate time-series data across multiple entities to account for both past values and individual entity characteristics that influence outcomes. There are also MCMC methods for Bayesian inference to further strengthen the analysis (Vrontos, Galakis, et al. 2024). SUR is also employed in another paper by Vrontos (2023) in which he explores sophisticated econometric and machine learning techniques to model and predict U.S. cement importer quantities over time. With the use of Predictive models like multiple regression-type models and machine learning techniques such as penalized likelihood regression models (Ridge, Lasso, Elastic Net), regression trees, and Random Forest as well as Binary Logit Models like Logit regression-type models and penalized likelihood binary Logit regression models the study focuses on medium-term forecast horizon and evaluation of the predictive ability of the modeling approaches using statistical metrics such as Mean Square Forecast Error (MSFE) and Mean Absolut Forecast Error (MAFE).

A notable Addition to the SUR methodology is presented in the paper by Ogaki (2003) which extends the traditional dynamic panel data model by incorporating Seemingly Unrelated Regressions (SUR) to account for cross-sectional correlation in the error structure. Unlike standard dynamic panel data approaches that assume independence across panel units, this method jointly estimates a system of dynamic equations one for each cross-sectional unit while allowing for contemporaneous correlation among their disturbances. This is particularly advantageous in settings where unobserved shocks are likely to affect multiple units simultaneously such as regional markets for construction materials. This method appears well

 https://doi.org/10.26219/heal.aueb.9465

suited to capture both the dynamic temporal structure and potential cross-state interdependencies in cement import movements.

3 Chapter 3: Data

3.1 Intro

In this chapter, the dataset utilized for analyzing the dynamics of construction material markets, specifically cement, within the United States from 2008 to 2022 is presented. The data source is a reputable industry database which monitors import and export quantities pertinent to construction materials etc. This source was chosen for its comprehensive coverage and reliability in tracking construction material import quantities over extended periods, ensuring the robustness of the analysis. The dataset comprises of quarterly observations of Portland Cement, Aluminous Cement, Slag Cement import quantities in kilograms (Kg), Articles of Cement Concrete or Artificial Stone in Kg, Articles of Plaster or Items Based on Plaster, Marble, Granite, Ceramic and many other material import quantities that can act as our dependent variables, alongside various economic indicators and variables that were selected as independent variables like the New Privately Owned Housing Units Authorized, Construction Employees in Thousands, Homeownership Rates by State and many more. All the available variables can be found in the appendix. Each of these variables are available by State. We are going to work with a long panel from 2008 to 2022 and the chosen 4 states, California, Florida, North Carolina, Texas. The data is structured as a balanced panel data, enabling the exploitation of both cross-sectional and time-series variabilities to understand complex interactions and temporal dependencies inherent in the import quantities of construction materials.

3.2 Data Cleaning

To beging with the data is checked for missing values and removed if any are found. The dataset is from a trustworthy and organized source so its complete and not much is needed in terms of cleaning. The data are grouped data based on the State they belong to. We have CAL for data from California, FL for Florida, NC for North Carolina and TX for Texas. The variable we are mainly interested is the 2523 Portland Cement Aluminous Cement Slag Cement Etc (Kg). This variable will be our dependent variable Y.

3.3 Exploratory Data Analysis

In Table 1 there are summary statistics for the dependent variable for each state which provides key summary statistics for cement imports across four major key states. These statistics include the mean, median, standard deviation (SD) and the range of imports, reflecting the central tendencies and variability in the data across different market conditions. During the period of our data Texas appears to have the highest imports of cement while North Carolina data show a significantly lower usage. California shows a high average import value of approximately 140 million but a much lower median of 76 million, suggesting a skewed distribution with some very high values pulling the mean upwards. This could indicate occasional large-scale imports due to significant infrastructure projects or seasonal construction spikes. Florida has a higher mean import value at approximately 257 million and a median closer to this mean at 175 million, suggesting a somewhat more consistent import pattern than California but still with substantial variability. North Carolina shows an unusually low mean and median compared to the other states, with mean imports at about 2.3 million and a median of just 500. This dramatically low median suggests that there are many months with minimal to no cement

https://doi.org/10.26219/heal.aueb.9465

imports. Texas shows the highest mean import value at approximately 664 million, with a median also substantial at 531 million, suggesting Texas as a major hub for cement imports likely due to its large construction market and extensive commercial activities. The variability in cement imports across these states can be attributed to a variety of economic, logistical, and regulatory factors. States with higher median and mean values, such as Texas and Florida, likely have robust construction sectors with consistent demand. In contrast, the low values in North Carolina suggest a smaller-scale or more cyclical construction sector. The data also suggest potential areas for deeper investigation, such as understanding the economic drivers behind the high variability in California and the minimal imports in some periods in North Carolina. These insights could be vital for stakeholders in the construction industry, policymakers, and economic planners to understand regional market dynamics and make informed decisions.

Table 1. Summary statistics for Cement imports by State (kg in millions)

State	California	Florida	North Carolina	Texas
Mean	139	257	2	664
Median	76	175	0,0005	531
SD	158	232	5	479
Min	9	2	0	174

Next is the analysis of the cement import time series across our four states. On the plot bellow we observe that each state shows distinct patterns in term of level, trend and volatility. Firstly in California the plot shows significant volatility with sharp peaks and troughs. There appears to be a cyclic pattern, suggesting periodic fluctuations. High volatility is evident with sudden spikes in cement imports. In Florida there is a steadily increasing trend in cement imports over time. Relative to California there is less volatility and more consistent growth. North Carolina displays a sudden spike after a period of very low activity in cement imports. This pattern might suggest an infrequent but significant demand for cement possibly due to a major project or a catastrophy that drove the demand high. Extremely high volatility is noted with almost no activity followed by a sharp spike. Lastly in Texas there is a moderately incresasing trend with some fluctuations. The overall upward trend suggests growing demand possibly linked to economic expansion and robust construction activity. There is moderate volatility with a notable drop and recovery.

https://doi.org/10.26219/heal.aueb.9465

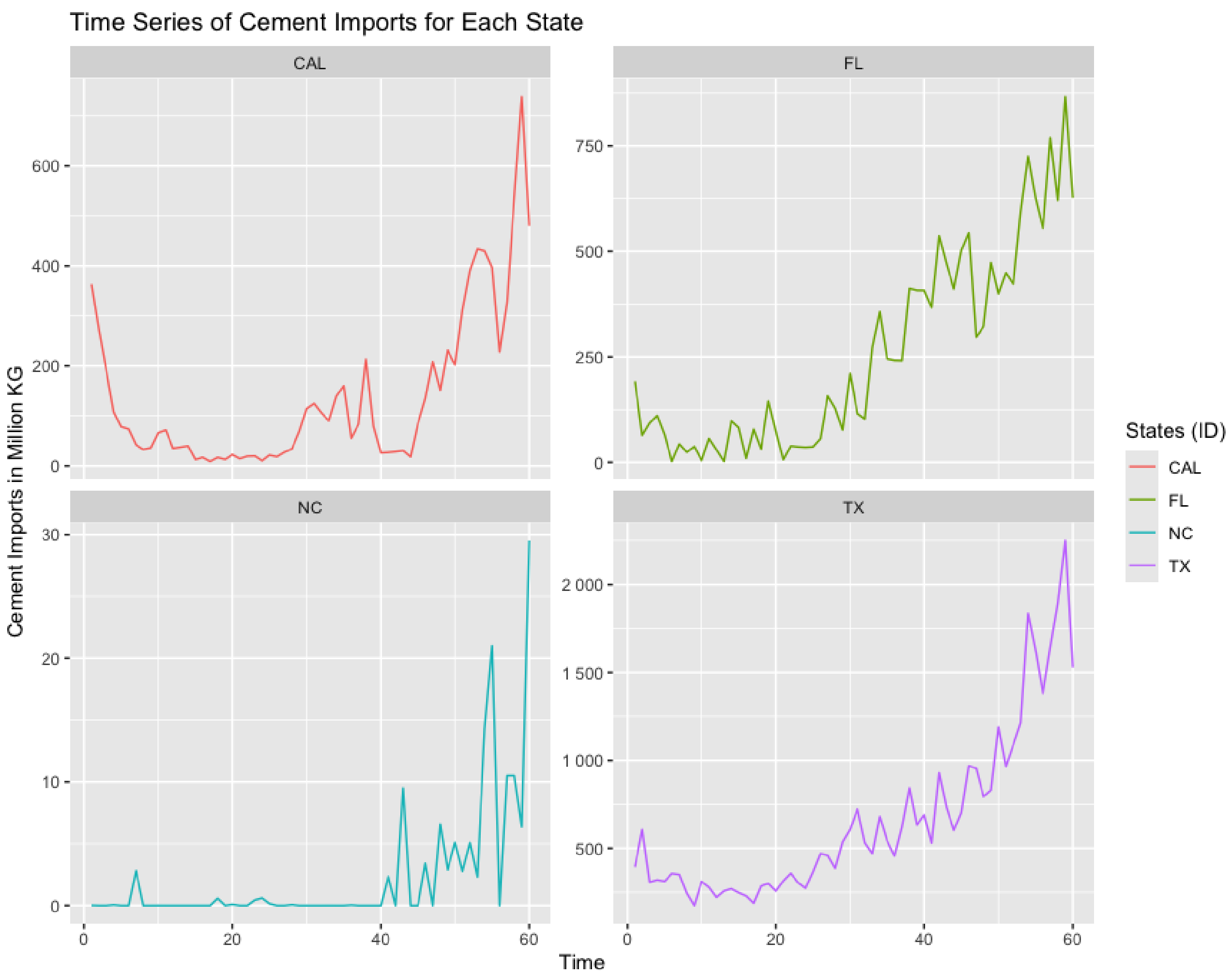

Figure 1 Time Series of Cement Imports by State

After visually inspecting the time series data comes the verification of the stationarity of each series. Stationarity testing is essential to ensure that the assumptions required for the subsequent time series modeling are met. The tests indicated that most of the independent variables were non-stationary.
To address this issue and stabilize the variance within our data, it is necessary to opt for a transformation known as Log-Differencing. This method involves taking the logarithm of the series to reduce variance amplification across higher values and then differencing the logarithmically transformed data to remove trends and achieve stationarity. This transformation is particularly effective in dealing with non-stationary data that exhibits exponential growth patterns or proportional variability. Bellow there are the dependent variables after the log transformation for each state.

https://doi.org/10.26219/heal.aueb.9465

- California:

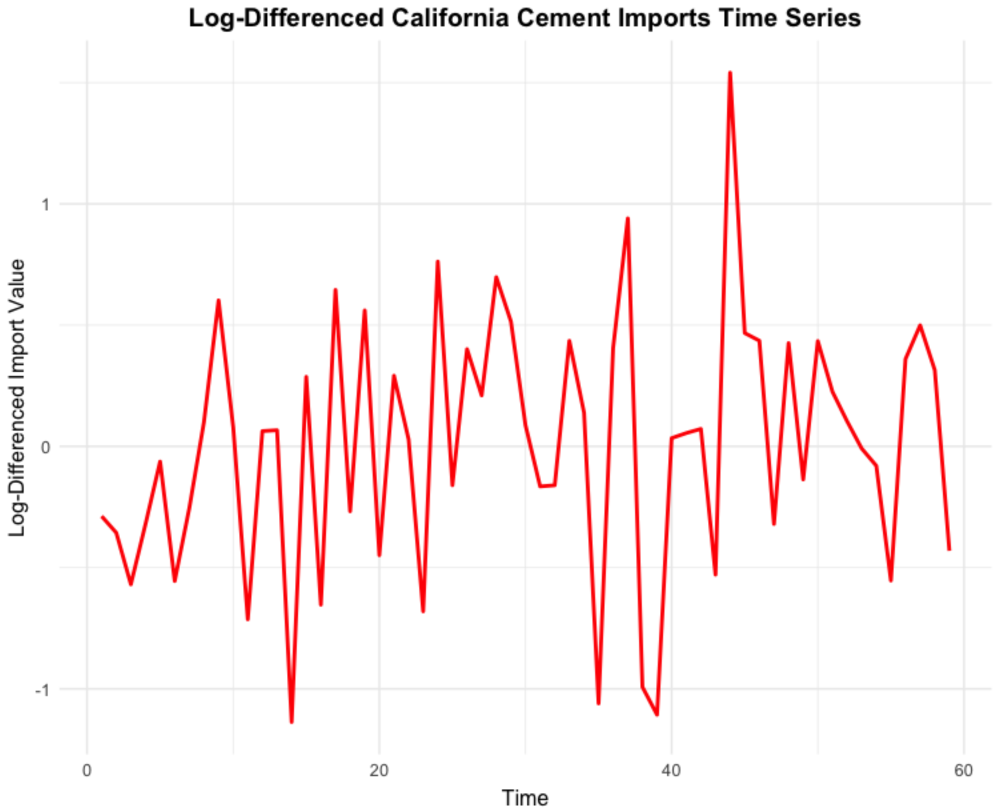

Figure 2 Log_differenced California Cement Imports Time Series

The figure above is the result of our log-differenced time series of cement imports into California. The transformation applied was intended to stabilize variance and achieve a stationary series suitable for analytical modeling. The series exhibits considerable volatility with significant peaks and troughs throughout the observed periods. This variability could be influenced by external economic factors or market demand fluctuations. The transformation appears successful in removing any long-term trends for the original data as there is no persistent upward or downward movement in the series. The data characteristics, particularly its volatility and the apparent lack of trend point towards models that can accommodate sudden changes and are robust to volatility such as ARIMA or GARCH models.

https://doi.org/10.26219/heal.aueb.9465

- Florida:

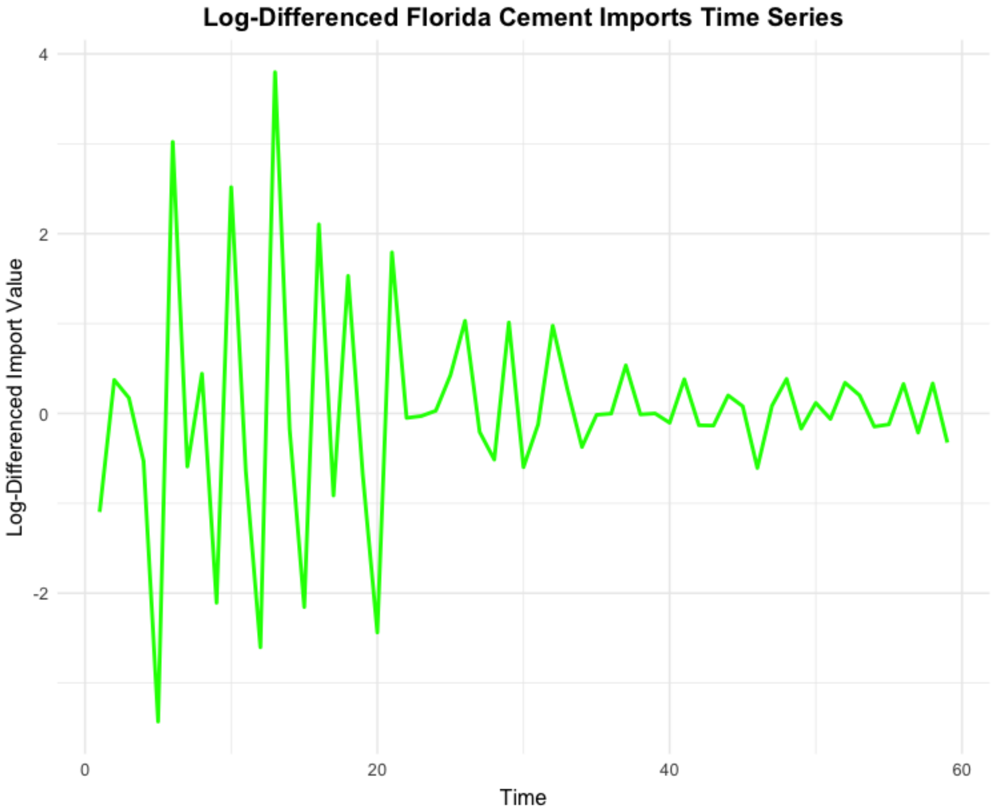

Figure 3 Log-Differenced Florida Cement Imports Time Series

This plot now displays significant volatility within the log-differenced series, particularly in the first half of the series. The magnitude of fluctuations suggests substantial variability in cement volumes. After the initial volatile period, however, the series exhibits more subdued fluctuations, suggesting a possible stabilization in the factors affecting cement imports.

https://doi.org/10.26219/heal.aueb.9465

- North Carolina:

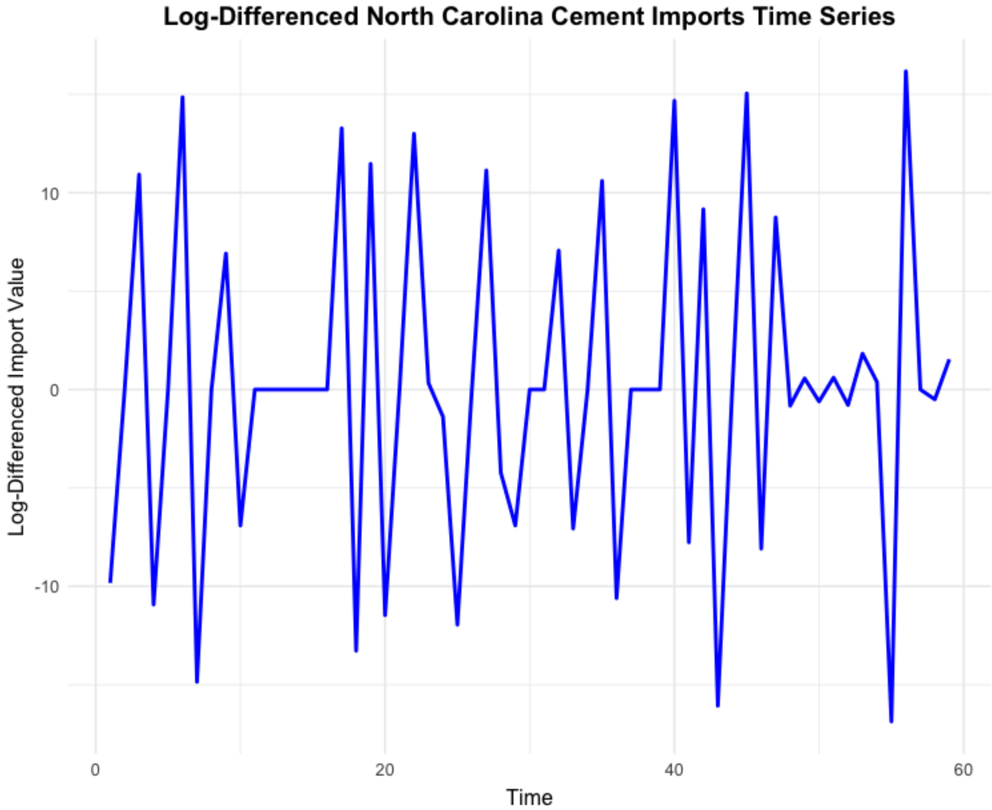

Figure 4 Log-Differenced North Carolina Cement Imports Time Series

The series for North Carolina displays a pattern of sharp peaks and troughs with a wide range of values, suggesting intense fluctuations in cement import volumes. Unlike Florida, this series maintains a high level of variability throughout, without a clear sign of stabilization. The extreme values, especially the negative dips, could indicate periods of significantly reduced import activity. There is no evident reduction in the variability throughout the series, which could reflect ongoing market instability or recurrent impoactful events such as catastrophes. The presence of extreme negative and positive values might reflect underlying cyclic economic actiivities.

 https://doi.org/10.26219/heal.aueb.9465

- Texas:

Last but not least Texas. Texas data show a relatively more contained fluctuation pattern, with less extreme variability compared to Florida and North Carolina. The fluctuations are somewhat regular, with a lower amplitude, suggesting the the import activities are influenced by less volatile market conditions or that the transformation effectively normalized more of the varince.

Each states transformed time series reveals different characteristics of the cement import market. Florida and North Carolina and California exhibit higher volatility, which could be indicative of more reactive or sensitive market conditions to external shocks. Texas displays a more consistent import pattern suggesting a more stable market.

None of the series display clear or persistent trend over time, which is a good indicator of stationarity. All series fluctuate around a constant mean and after running statistical tests such as the Augmented Dickey-Fuller(ADF) test and KPSS test we can formally assess the stationarity of the above series.

https://doi.org/10.26219/heal.aueb.9465

https://doi.org/10.26219/heal.aueb.9465

4 Chapter 4: Models

4.1 Importance of Modeling

In the real of economic and financial analysis, modeling time series and dynamic panel data serves as a cornestone for understanding complex temporal and cross-sectional dynamic. The ability to accurately model these dynamics is crucial for predicting future trends, making informed decisions, and testing theoretical assertins against empirical data. This chapter will build on the preliminary analysis and transformation of the data from chapter 3. The primary aim of this section is to deelop robust models that can efficiently capture the underlying patterns and relationships within the data. Given the unique characteristics of the dataset–a long panel with extensive temporal data across four states-this chapter focuses on deploying sophisticated econometric models tailored to unravel the dynamic and complex interactions within the data. Specifically, the objectives will be:

- **Identify Suitable Models**: Based on the characteristics of the stationary time series data and the structure of the panel data, identify and justify the most suitable econometric models.
- **Estimate Model Assumptions**: Implement these models to estimate parameters that best describe the relationships and dynamics present in the data.
- **Validate Model Assumptions**: Through diagnostic checks, ensure that the models meet the necessary statistical assumptions, providing reliable and valid results.
- **Forecast and Simulate:** Utilize the models to make forecasts and simulate potential future scenarios, which are invaluable for strategic planning and policy formulation.

4.2 Models to be Explored

The chapter will explore the following models, each chosen for its relevance to the specific challenges and characteristics of our panel data:

- Autoregressive Integrated Moving Average models (ARIMA)
- Pooled OLS Models
- Panel Autoregressive Distributed Lag (ARDL) Models: To capture both short-term and long-term relationships within the data.
- Seemingly Unrelated Regressions (SUR): To efficiently estimate models with potentially correlated error terms across different states.
- Fixed Effects and Random Effects Models: To account for unobserved heterogeneity across states that could influence import volumes.
- Dynamic Panel Data Models:

4.3 ARIMA

To begin with,simple models will be used using only the imports for each state separately. The methodology will be the Autoregressive Moving Average (ARMA). This modeling technique will enable to quantitatively predict future trends in cement imports across the four chosen states. The parameters will be chosen based on the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) analyses. The models will not only serve as a predictive tool but will also enrich the understanding of the dynamic relationships within the data, potentially revealing unique insights into economic activities and market behaviors specific to each state. Given the complexity of the dataset and the potential influence of multiple exogenous factors, the necessity of efficiently selecting the independent variables is recognized. This will be accomplished by employing two primary methods: correlation analysis and Lasso regression. First we will apply correlation analysis for a glimpse of the relation between the dependent and independent variables and after the application of Least

https://doi.org/10.26219/heal.aueb.9465

Absolute Shrinkage and Selection Operator(LASSO) regression will choose the most important variables for the analysis.

4.3.1 Correlation analysis

The correlation plots reinforce the notion that different states exhibit distinct relationships between explanatory variables and cement imports, underscoring the importance of region-specific modeling. In California, we observe that variable (24), Weather and Climate Billion-Dollar Disasters, has the most pronounced negative correlation with cement imports, confirming that natural disasters substantially disrupt construction activity in the state. In Florida, while disasters have a modest effect, a more intriguing pattern emerges with variable (9), the Rental Vacancy Rate**.** Its correlation with cement imports is positive at short lags but turns negative at longer lags. This suggests that rising vacancies may initially coincide with increased cement demand, potentially due to short-term responses like speculative building or stimulus-supported activity, yet prolonged high vacancy rates eventually dampen construction, as oversupply becomes persistent. Variable (10), Homeowner Vacancy Rate, exhibits the strongest positive correlation with cement imports among all variables in North Carolina. This implies that higher homeowner vacancy rates are closely associated with increased cement imports in the state. Most notably**,** Texas exhibits the richest correlation structure of all four states. It shows **a** large number of variables with statistically meaningful correlations, both positive and negative, across multiple lags. This implies that cement imports in Texas are influenced by a broader set of economic, demographic, and housing-related factors**,** reinforcing the need for more flexible, state-specific dynamic modeling**.**

.

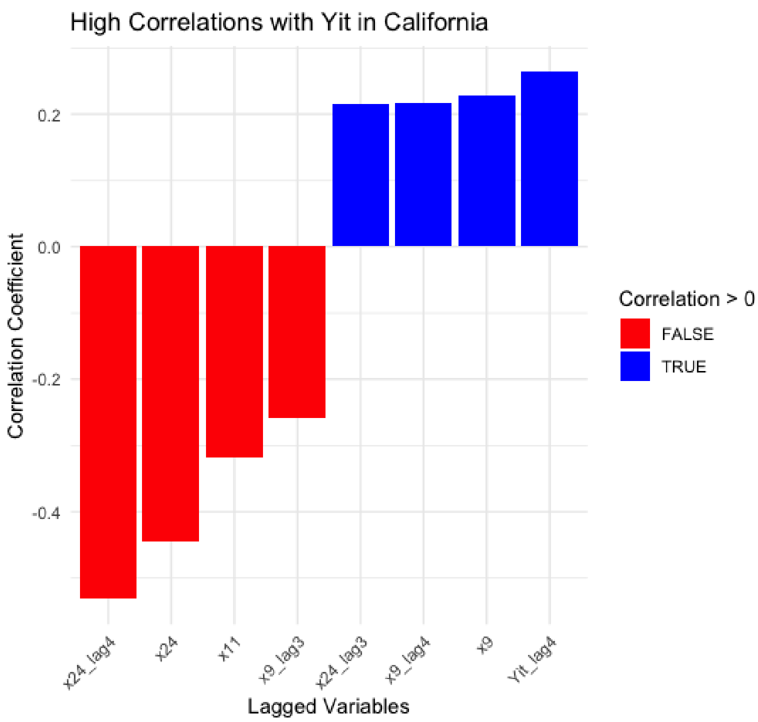

Figure 5 Correlation Plot of California

https://doi.org/10.26219/heal.aueb.9465

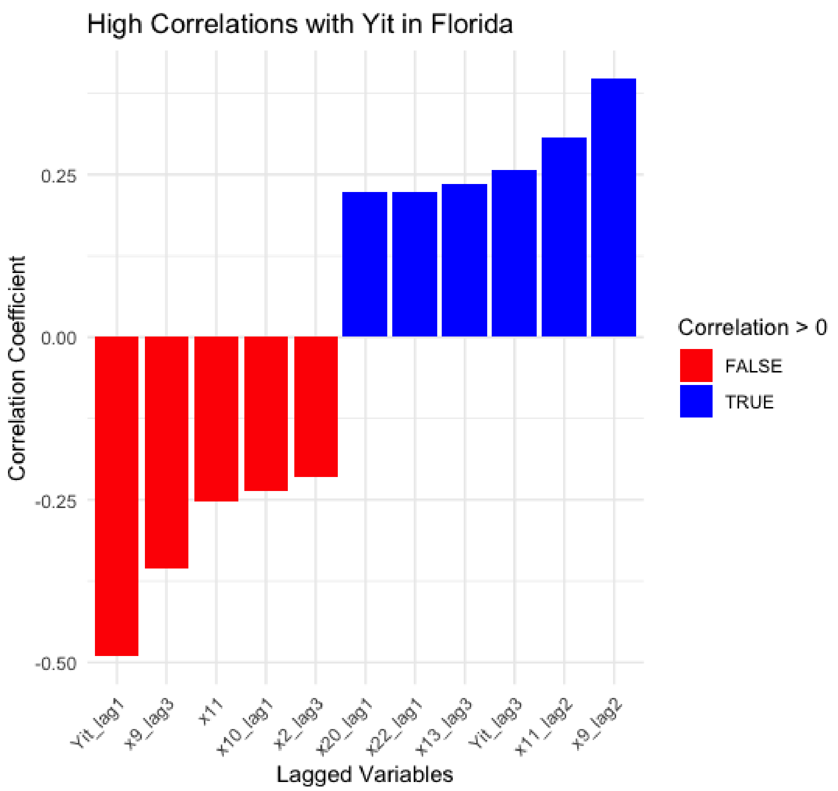

Figure 6 Correlaton Plot of Florida

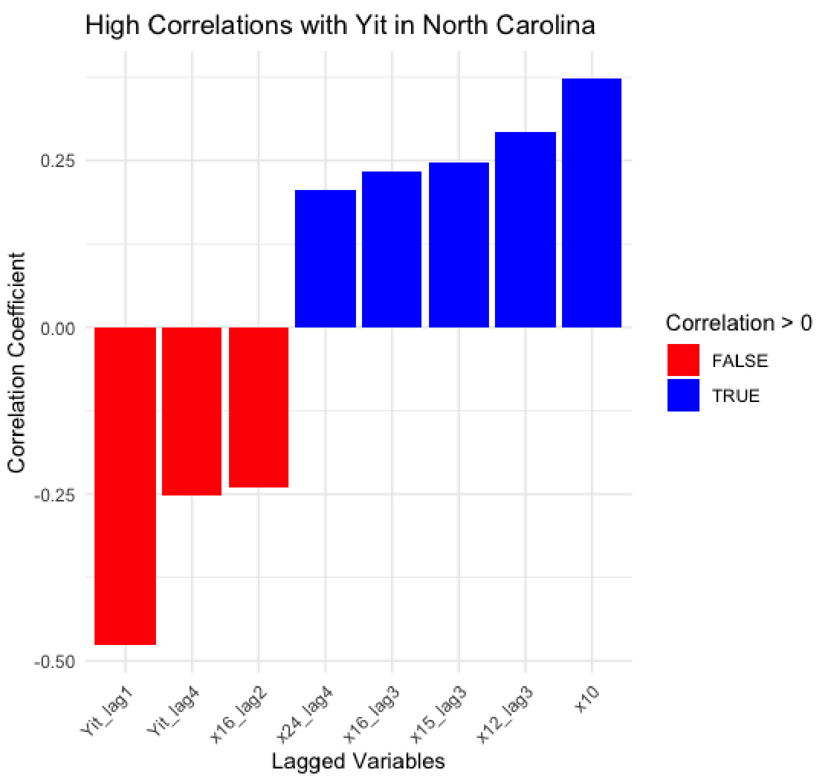

Figure 7 Correlation plot of North Carolina

https://doi.org/10.26219/heal.aueb.9465

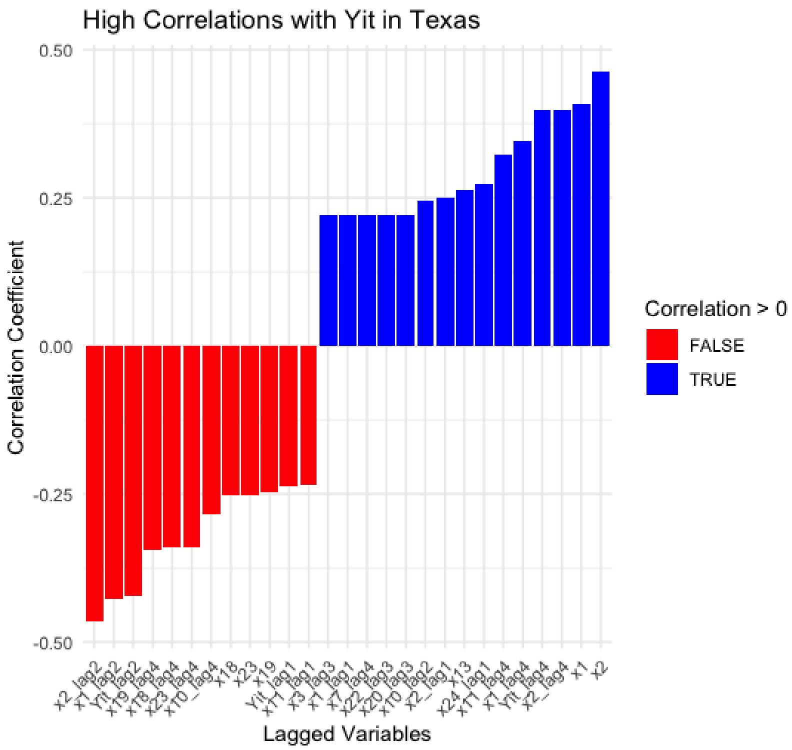

Figure 8 Correlation plot of Texas

https://doi.org/10.26219/heal.aueb.9465

4.3.2 Lasso Regression Analysis

Moving into the Lasso regression analysis, the goal is to perform variable selection tailored to each state's ARMA model. Given the distinct regional characteristics and the varying influence of predictors across states (as shown in the correlation plots), it is reasonable to expect that different variables will be relevant for different regions. Lasso regression is used here not as a predictive model itself, but as a variable selection tool. By applying an ℓ1 penalty, Lasso shrinks the coefficients of less influential variables toward zero, effectively filtering out noise and highlighting the most important predictors. This helps reduce overfitting and improves the forecasting performance of the subsequent ARMA models by ensuring that each state's model includes only its most relevant predictors.

Lasso Regression Formula:

$$Minimize\{\frac{1}{2T}\sum_{t=1}^{T}\left(y_t-\sum_{j=1}^{p}x_{t,j}\beta_j\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\}$$

Where:

- t is the number of time periods
- y_t is the dependent variable at time t
- $x_{t,j}$ is the value of the j-th variable at time t
- β_j is the coefficient for the predictor variable x_j
- λ is the regularization parameter that controls the strength of the penalty
- p is the number of the predictor variables

Lasso Results:

Table 2 Lasso Regression Variables

California	Florida	North Carolina	Texas
Weather and Climate Billion-Dollar Disasters to affect (lag 4)	**Rental Vacancy Rates by State (lag 2)**	**Wages and salaries by place of work (Thousands of dollars) (lag 2)**	**New Private Housing Units Authorized by Building Permits, Units, Seasonally Adjusted**
Homeownership Rates by State	**Rental Vacancy Rates by State (lag 3)**	**Homeowner Vacancy Rates by State**	**New Private Housing Units Authorized by Building Permits, Units, Seasonally Adjusted (lag 2)**
	Homeownership Rates by State		

Lasso regression was employed to select the most impactful predictors for cement imports in our chosen States. The regression analysis revealed a distinct set of predictors for each state, reflecting the unique economic and logistical factors influencing cement import dynamic regionally.

https://doi.org/10.26219/heal.aueb.9465

4.3.3 Models and Evaluation

After using auto.arima with xreg input (ARIMAX) for the exogenous variables, the following models for each state are formed:

- California: ARMA (2,1) with random residuals (Ljung-Box test P-value =0.27)
- Florida: ARMA(2,1) with random residuals (Ljung-Box test P-value = 0.47)
- North Carolina: ARMA(3,2) with random residuals (Ljung-Box test P-value = 0.15)
- Texas: ARMA(2,2) with random residuals (Ljung-Box test P-value = 0.05)

The residuals were checked visually and resemble white noise (randomness)
Also checking for Normality(shapiro.test) and Homoscedasticity(bptest) of residuals and both assumptions hold for all the States models residuals.

Table 3 ARMA Models Metrics

Metric	California	Florida	North Carolina	Texas
Mean Error (ME)	0.14	-0.14	0.56	0.04
Root Mean Squared Error (RMSE)	0.33	0.64	8.06	0.15
Mean Absolute Error (MAE)	0.27	0.53	2.74	0.11

This section evaluates the predictive performance of the ARIMA models which have been tailored with variables selected via Lasso regression for each states cement import data. Above is the table with the metrics and the forecast plots for each State.
The forecasting performance of the ARIMA models, enhanced with Lasso-based variable selection, was evaluated across the four selected U.S. states using standard accuracy metrics: Mean Error (ME), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). The model for California demonstrated strong performance with low bias (ME = 0.14) and relatively low error values (RMSE = 0.33, MAE = 0.27), indicating accurate and consistent forecasts. Texas yielded the best overall results, exhibiting minimal bias (ME = 0.04) and the lowest error metrics (RMSE = 0.15, MAE = 0.11), reflecting excellent predictive accuracy. Florida showed slightly worse performance with a modest underprediction bias (ME = -0.14) and moderate error levels (RMSE = 0.64, MAE = 0.53). In contrast, North Carolina experienced substantial difficulties, with high bias (ME = 0.56) and very large errors (RMSE = 8.06, MAE = 2.74), suggesting a poor model fit. This could be attributed to unaccounted structural breaks, volatility, or suboptimal variable selection. Overall, while the models perform well for most states, further refinement—particularly for North Carolina—is recommended to improve robustness and forecasting reliability.Overall, the models perform adequately in stable environments but struggle with anomalies or extreme changes. This necessitates further exploration into model robustness, possibly through the integration of additional predictive indicators or the adjustment of model complexities to enhance their adaptive capabilities for each state’s specific characteristics. Such improvements could potentially lead to more reliable forecasts, particularly in states where the current models underperform.

https://doi.org/10.26219/heal.aueb.9465

Forecast Plots:
Each plot contrasts the actual data points with the forecasted values over the test period, highlighted by a confidence interval tha indicates the range of expected forecasts. For California and Florida the forecasts follow the actual trends showing some effectiveness in these states. In contrast, the plot for North Carolina exhibits a significant divergence during a peak, suggesting the model's limitations in handling sudden or extreme changes. Texas displays moderate predictive accuracy.

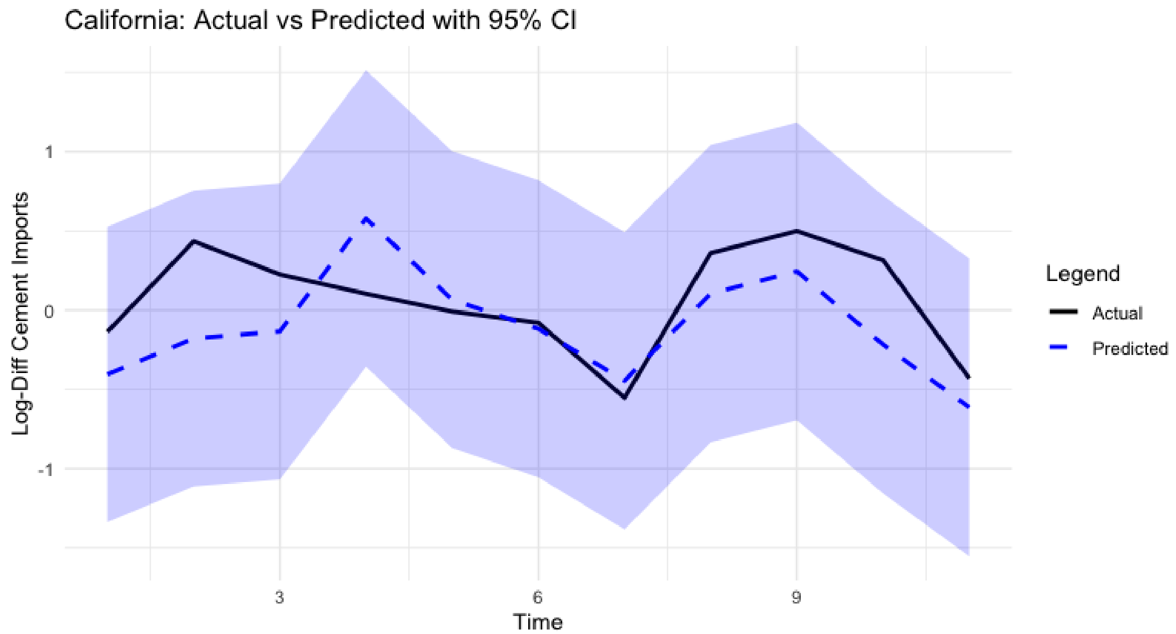

Figure 9 California ARMA Forecast

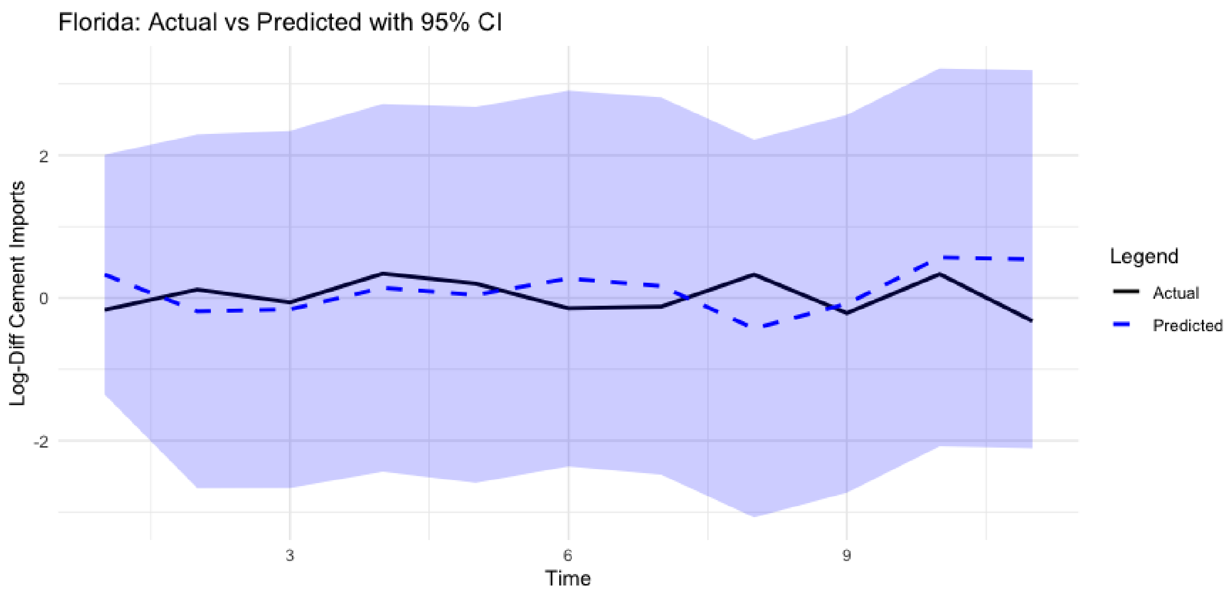

Figure 10 Florida ARMA Forecast

https://doi.org/10.26219/heal.aueb.9465

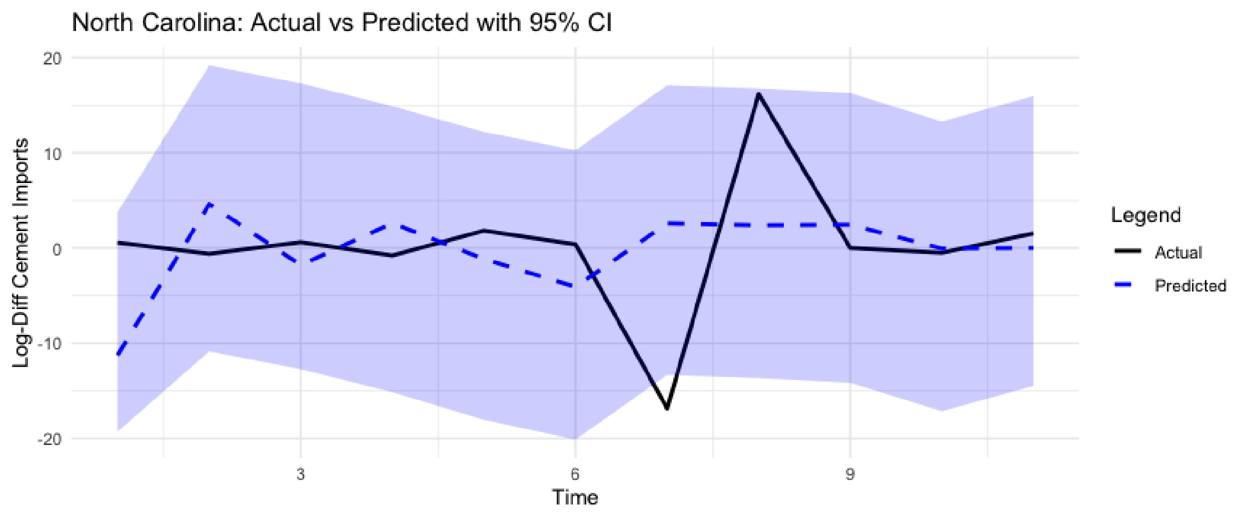

Figure 11 North Carolina ARMA Forecast

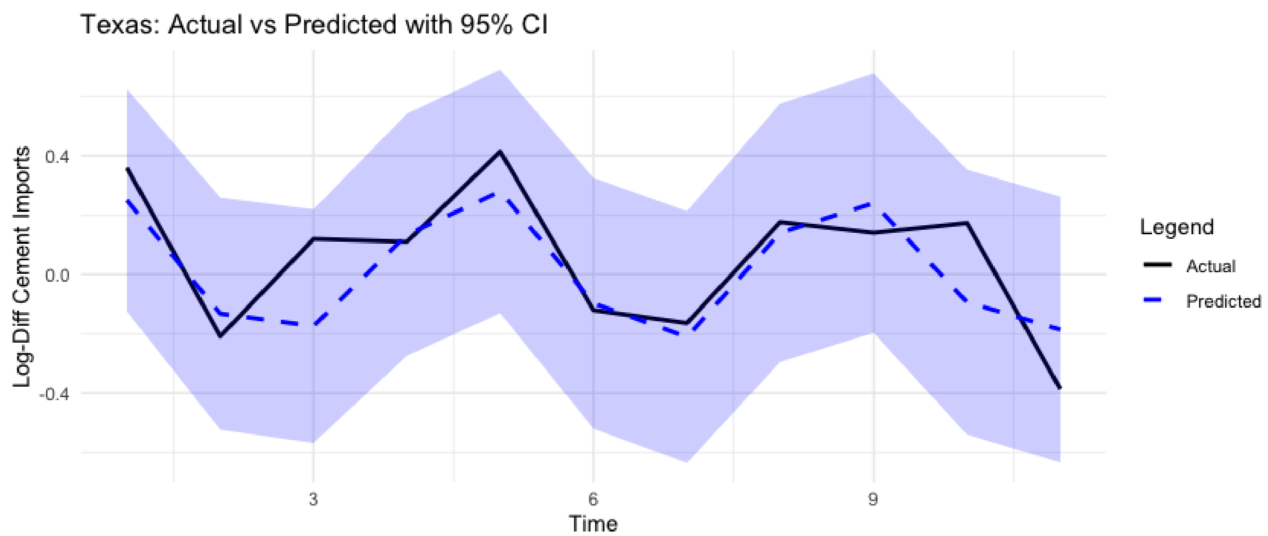

Figure 12 Texas ARMA Forecast

4.3.4 Conclusion

In the exploration of ARMA modeling for cement imports across difference states the application of variable selection through LASSO regression ensured that only statistically significant predictors were used. After choosing the significat predictors, the data was split into training and testing set and ARMA models for each state were created. Their primary limitation was their inability to forecast large peaks due to certain unknown factors. While ARMA models served as a solid baseline for time series forecasting their inability to simultaneously model interdependencies across multiple states and account for potential cross-sectional influences lead us to the next phase of the analysis: implementing dynamic panel data models. These models will allow the full exploitation of the panel structure of the data, incorporating time-invariant characteristics and capturing the dynamics aross both temporal and cross-sectional dimensions.

https://doi.org/10.26219/heal.aueb.9465

4.4 Pooled Ordinary Least Squares (Pooled OLS) Models

4.4.1 Inference

To beging the panel analysis the Pooled Ordinary Least Squares (Pooled OLS) model is applied. This model treats the panel data as a simple cross-sectional dataset by assuming a common intercept and slope for all states across all time periods. It does not account for any time-invariant heterogeneity across entities (states) or over time. First a dynamic panel data model that aggregates data from all states and includes up to four lags for each variable, encompassing both dependent and independent ones is constructed. Best way to initiate the model selection process was with the full model, employing a bidirectional stepwise approach that utilizes both the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) for model optimization.

BIC tends to penalize the number of variables more heavily than AIC, leading to a simpler model. This is because BIC includes a stronger penalty term for the number of parameters in the model, which helps to avoid overfitting by selecting a model with fewer parameters unless the addition of more variables substantially improves the model fit. On the other hand, the AIC prioritizes model fit over simplicity, often favoring models with a larger number of predictors in order to maximize the proportion of explained variance (as reflected by a higher R-squared). This dual approach allows the evaluation of the trade-offs between model simplicity and explanatory power, ensuring that the final model is both parsimonious and robust. Choosing to continue with the model that resulted from the BIC stepwise because it is much simpler with a small decrease in R-Squared:

	Variables	R-Squared
AIC	52	64%
BIC	9	53%

Also in the AIC model most variables are not statistically significant. Meanwhile in BIC one all of the coefficients are highly significant (p-value > 0.05). As expected most of the lags of the dependent variables (Yit-Cement Imports) are significant with negative coefficients, suggesting that past values have an inverse relationship with the current values. Approximately 53% of the variability in the dependent variable is explained by the model, there is room for model improvement. We also run a VIF – Multicollinearity check and we found that 2 of our variables have really high multicollinearity(>10) so after removing both of them there is no issue of multicollinearity,(<3) and the trade-off is a 2% increase in R-Squared. Unfortunately this model does not adequately address the assumptions necessary for valid inference. Normality and Homoscedasticity assumptions do not hold and that can significantly affect the reliability and validity of the models outputs.

4.4.2 Predictions

Next the independent variables that the BIC stepwise algorithm resulted are used and by splitting the dataset into a training and a testing set predictions are made. The Pooled OLS regression will be fit on the training data (first 80% of each state's time series) and the trained model will be used to predict values on the test set (final 20% of each state's time series). The RMSE(Root mean squared error) and MAE(Mean average error) metrics to assess prediction accuracy were calculated. The model has the following form:

$$y_{it} = \alpha + \rho_1 y_{i,t-1} + \rho_2 y_{i,t-2} + \rho_3 y_{i,t-3} + \rho_4 y_{i,t-4} + \sum_{j=1}^{k} \beta_j x_{j,i,t-1_j} + \epsilon_{it}$$

 https://doi.org/10.26219/heal.aueb.9465

Where:

- y_{it}: Dependent variable (e.g., cement imports) for entity i at time t
- $y_{i,t-1}$: Lagged values of the dependent variable (lags 1 to 4)
- $x_{j,it-\ell_j}$: Lagged explanatory variable j for entity i, with individual lag length ℓ_j
- ρ_l: Coefficients for the lagged dependent variables
- β_j: Coefficients for lagged independent variables
- α: Common intercept across all entities
- ϵ_{it}: Error term

And the metrics are as follows:

State	RMSE	MAE
California (1)	5.58	3.84
Florida (2)	5.63	3.08
North Carolina (3)	7.27	5.09
Texas (4)	4.74	2.73

The variables chosen by the BIC stepwise are the following

Variable	Description	Lag	Interpretation
Yit_lag1	Cement Imports	1	Short-term persistence
Yit_lag2	Cement Imports	2	Short-run dynamics/delayed adjustment
Yit_lag3	Cement Imports	3	Medium-term memory
Yit_lag4	Cement Imports	4	Long-run memory
x7lag1	Unemployment	1	Lower unemployment boost demand
x8lag1	Employment	1	Construction sector activity
x12lag3	Real Estate, Rental & Leasing Earnings (Thousands of $)	3	Income signal related to real estate returns
x13lag3	All-Transactions House Price Index	3	Property marker valuation
x16lag2	Wages and Salaries	2	Proxy for input costs

https://doi.org/10.26219/heal.aueb.9465

Bellow there is a plot with the results: 1: California, 2: Florida, 3: North Carolina, 4: Texas

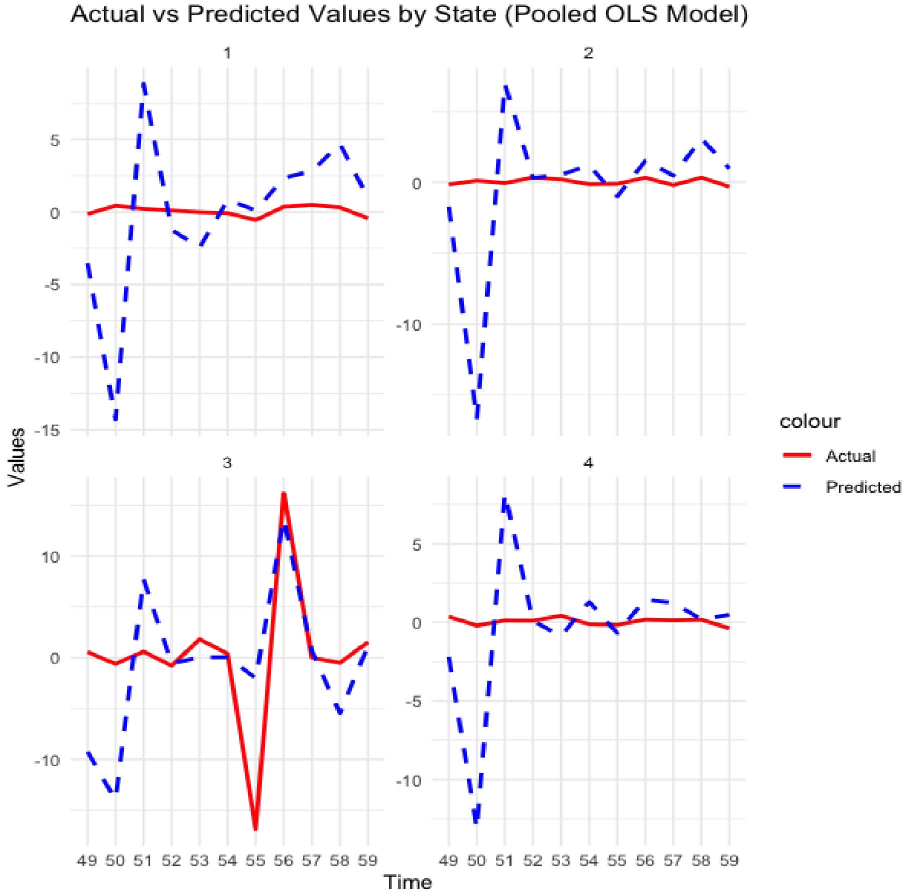

Figure 13 Pooled OLS Prediction Model - Using BIC Variable Selection Method

As observed, the Pooled OLS model struggles to capture fluctuations and the dynamic nature of the time series data as it ignores the panel structure, creates one regression line for all observations and uses the same intercept and same slope for all coefficients, for all states at all times. The large deviations at certain points suggest that treating all states homogeneously fails to account for state-specific effects, which may lead to biased estimates. An algorithm was created to select independent variables based on the best predictive RMSE by adding variables one by one. The variables chosen by the algorithm where the following:

Variable	Description	Lag	Interpretation
Yit_lag1	Cement Imports	1	Short-term persistence; negative effect expected
Yit_lag2	Cement Imports	2	Recent trend continuation
Yit_lag3	Cement Imports	3	Medium-term dynamic effect
x1lag3	Homeownership Rates by State	3	Construction activity signal (slightly delayed impact)
x1lag4	Homeownership Rates by State	4	Long-term housing-driven demand
x2	New Private Housing Units Authorized by Building Permits	0	Real-time measure of construction planning activity
x4lag3	Construction All Employees	3	Workforce size reflects delayed building capacity effect

https://doi.org/10.26219/heal.aueb.9465

Variable	Description	Lag	Interpretation
x5lag3	Construction of Buildings: Employment	3	Building-specific employment; reflects capacity and labor cost
x6	Building Materials Retail Employment	0	Retail-side demand and distribution
x9lag2	Rental Vacancy Rates	2	Inverse demand signal: higher vacancies reduce import need
x11lag4	Homeownership Rates by State	4	Delayed investment effect from real estate profitability
x15lag3	Wages and Salaries by Place of Work	3	Broader labor cost effect; may reduce construction margins
x17lag2	Construction Earnings	2	Sector output impact on construction inputs
x18	All industry total	0	Macro-level growth indicator; drives overall demand
x21	Real estate and rental leasing	0	Sector-specific GDP; direct driver of cement imports
x21lag3	Real estate and rental leasing	3	Lagged construction growth; trailing effect on demand
x21lag4	Real estate and rental leasing	4	Long-run sector performance effect

The resulting model predictions are shown bellow:

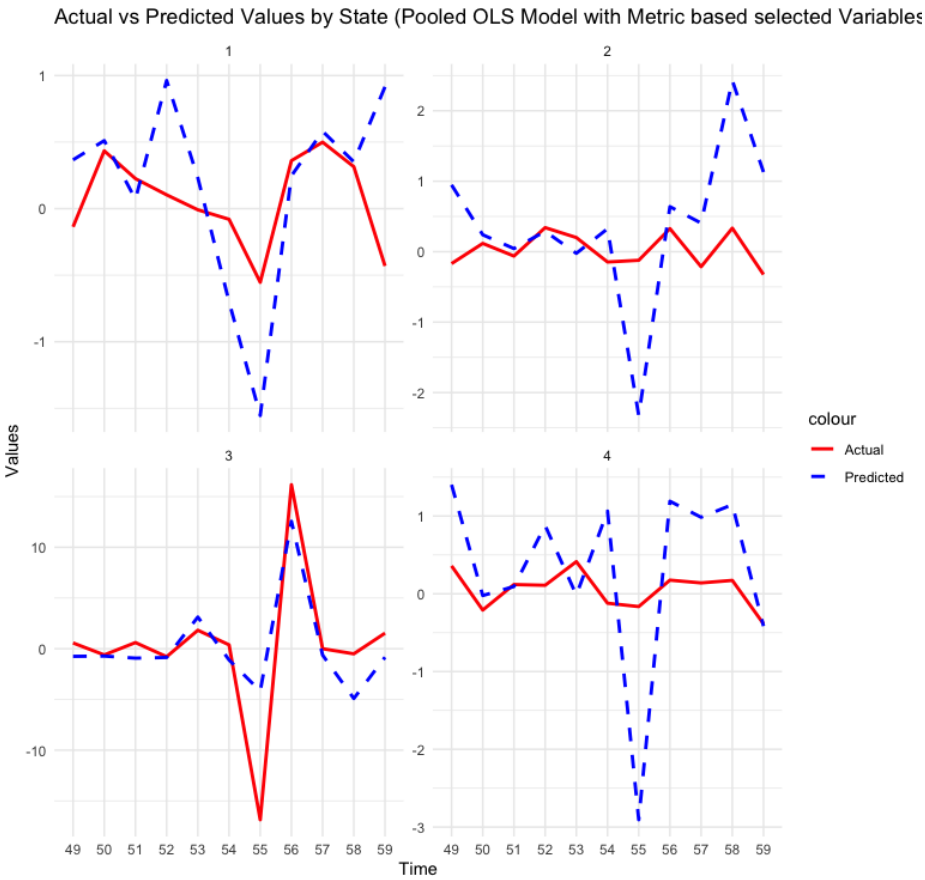

Figure 14 Pooled OLS predictions with RMSE-based variables selected

https://doi.org/10.26219/heal.aueb.9465

Table 4 Pooled OLS prediction metrics using RMSE selected variables method

State	RMSE	MAE
California (1)	0.625	0.457
Florida (2)	1.10	0.798
North Carolina (3)	4.34	2.69
Texas (3)	1.11	0.840

The pooled OLS model with variables selected based by the RMSE optimization shows mixed predictive performance across states. The prediction plot reveals that while the model captures the general trends there are notable deviations. The predictions for Florida and Texas appear relatively stable but still diverge at certain peak and drought points. North Carolina has the highest RMSE (4.34) suggesting the the model struggles significantly to capture the variability in this state but in the plot the predictions do relatively good if we exclude the large dip that could not be predicted by the model. California has the lower RMSE (0.625) indicating a more reliable prediction. Overal the predictions performance is better than the previous BIC Pooled OLS model but still do not fully capture the nature of the data.

These deviations from the actual data are assumed to exist due to unobserved individual-specific effects. To address this issue, we will proceed with the Fixed Effects Model (FE), which allows us to control for unobserved heterogeneity across states by removing time-invariant differences and work from there to get some better inference and predictions.

4.5 Fixed Effects Model

4.5.1 Inference

A fixed effects panel data model was applied to capture individual-specific characteristics across the four U.S. states under study (California, Florida, North Carolina and Texas), taking into account both the lagged dependent variable and selected independent variables. By creating an algorithm that begins with the null model and adds variables one by one, checks the BIC value and accepts or rejects the added variables results in strong and statistically significant temporal dependencies, particularly at lags one and two, highlighting considerable inertia in cement import dynamics. Specifically the algorithm accepted variables **'Real Estate, Rental and Leasing Earnings, Thousands of Dollars, Quarterly, Seasonally Adjusted Annual Rate'** at Lag 3 and **'New Private Housing Units Authorized by Building Permits: 1-Unit Structures, Units, Seasonally Adjusted'** at Lag 1. The model overall achieves an R-squared of approximately 50%, indicating a moderate explanatory capability for variations in cement imports.

Using standard OLS with a dynamic panel model can lead to Bias, spiecifically Nickel Bias that can lead to incorrect inference about coefficients. In this case due to the relatively large T(55) and small N(4 states) the bias becomes significantly less severe as the correlation between lagged dependent variables and error terms reduces as T increases (Nickell 1981).

To assess the validity of the dynamic fixed-effects panel model, key regression assumptions such as homoscedasticity, normality, and absence of autocorrelation were evaluated through diagnostic testing on the residuals. The residual diagnostics-including the Breush-Pagan test, Shapiro-Wilk normality test, and visual inspection via residual plot- indicated violations of several standard regression assumptions. To address these violations and ensure robust inference, robust standard errors (HC3 type) were implemented (Davidson and MacKinnon 1993). This adjustment will allow for accurate hypothesis testing despite deviations from classical assumptions.

 https://doi.org/10.26219/heal.aueb.9465

The Fixed effects model has the following form:

$$y_{it} = \alpha_i + \rho_1 y_{i,t-1} + \rho_2 y_{i,t-2} + \rho_3 y_{i,t-3} + \rho_4 y_{i,t-4} + \sum_{j=1}^{k} \beta_j x_{j,i,t-l_j} + \epsilon_{it}$$

Where:

- y_{it}: Dependent variable (e.g., cement imports) for entity i at time t
- α_i: **State-specific intercept** (captures time-invariant heterogeneity)
- ρ_l: Coefficients for the lagged dependent variables
- β_j: Coefficients for lagged independent variables
- $y_{i,t-1}$: Lagged values of the dependent variable (lags 1 to 4)
- $x_{j,it-\ell_j}$: Lagged explanatory variable j for entity i, with individual lag length ℓ_j
- ϵ_{it}: Error term
-

4.5.2 Predictions

Using the variables chosen using the BIC stepwise selection method and running predictions on the testing data using an 80-20 split produces the following:

Variable	Description	Lag	Interpretation
Yit_lag1	Cement Imports	1	Short-term persistence
Yit_lag2	Cement Imports	2	Short-run dynamic adjustment
Yit_lag3	Cement Imports	3	Medium-term memory; part of cyclical pattern
Yit_lag4	Cement Imports	4	Long-run inertia in demand
x12lag3	Real Estate, Rental & Leasing Earnings (Thousands of $)	3	Lagged profitability in real estate; stronger earnings raise demand
x3lag1	New Private Housing Units Authorized by Building Permits	1	Retail-side labor demand; reflects near-term construction activity

https://doi.org/10.26219/heal.aueb.9465

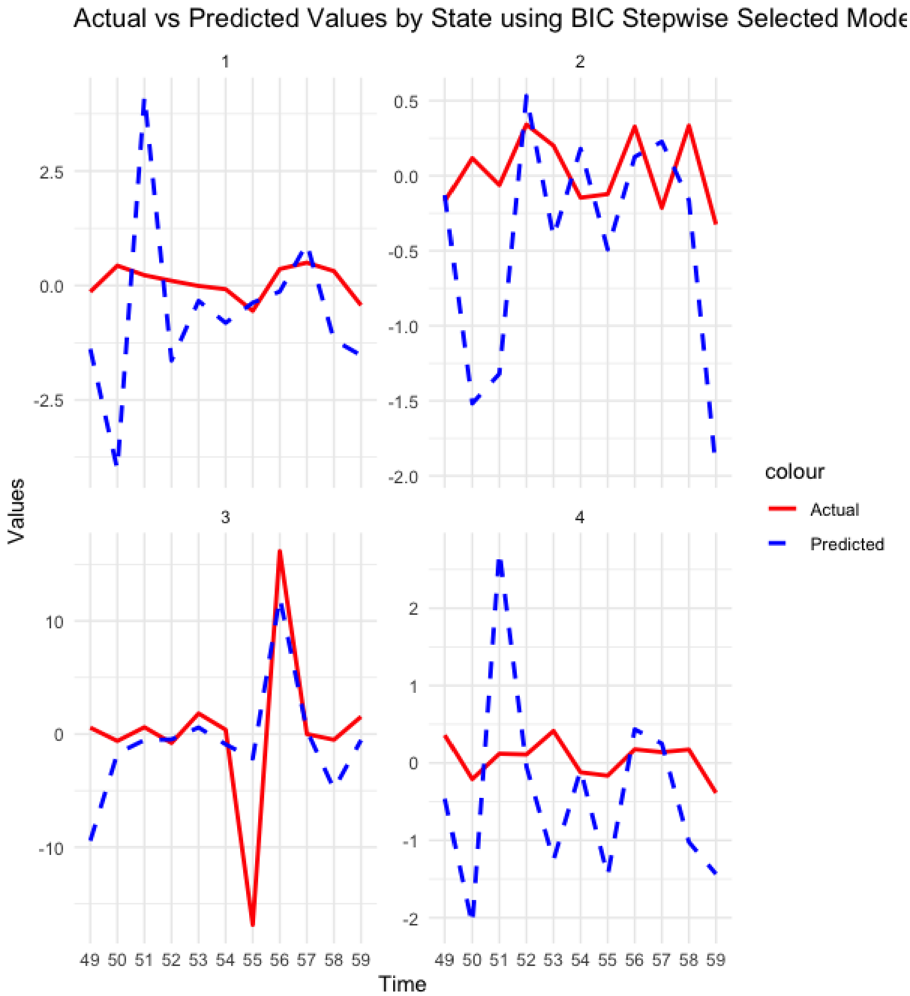

Figure 15 BIC Stepwise selected model Predictions by State

Table 5 BIC Stepwise model prediction metrics

State	RMSE	MAE
California (1)	2.0068	1.4638
Florida (2)	0.8557	0.6544
North Carolina (3)	5.7369	3.6938
Texas (4)	1.2846	1.0075

From the Figure and Table above the BIC model captures the general trend of the data in some states but in other it truggles to replicate fluctuations accurately. North Carolina (3) exhibits significant deviations between actual and predicted values, which aligns with its high RMSE (5.73) indicating poor predictive accuracy. Florida (2) has the lowest RMSE (0.85) suggesting that the model fits this state better than others. Again, the algorithm that adds variables one by one and checks the RMSE or the MAE and 'accepts' or 'rejects' each variable until the lowest metric available is met is used. The variables chosen are the following:

https://doi.org/10.26219/heal.aueb.9465

Variable	Description	Lag
Yit_lag1	Cement Imports	1
Yit_lag2	Cement Imports	2
Yit_lag3	Cement Imports	3
x1lag3	Homeownership Rates by State	3
x1lag4	Homeownership Rates by State	4
x2	New Private Housing Units Authorized by Building Permits,	0
x2lag3	New Private Housing Units Authorized by Building Permits,	3
x3lag3	New Private Housing Units Authorized by Building Permits: 1-Unit Structures	3
x4lag3	Construction All Employees	3
x4lag4	Construction All Employees	4
x5lag3	Construction of Buildings Employment	3
x5lag4	Construction of Buildings Employment	4
x6	Building Material Retail Employment	0
x7lag4	Unemployment Rate	4
x9lag2	Rental Vacancy Rate	2
x15lag3	Construction Wages and Salaries	3
x16	Wages and Salaries by Place of Work	0
x17lag2	Construction Earnings	2
x18lag4	All industry total	4
x19lag3	Private Industries	3
x21	Real estate and rental and leasing	0
x21lag1	Real estate and rental and leasing	1
x21lag4	Real estate and rental and leasing	4
x23	Real GDP: Real Estate Sector	0

 https://doi.org/10.26219/heal.aueb.9465

And the predictions are the following:

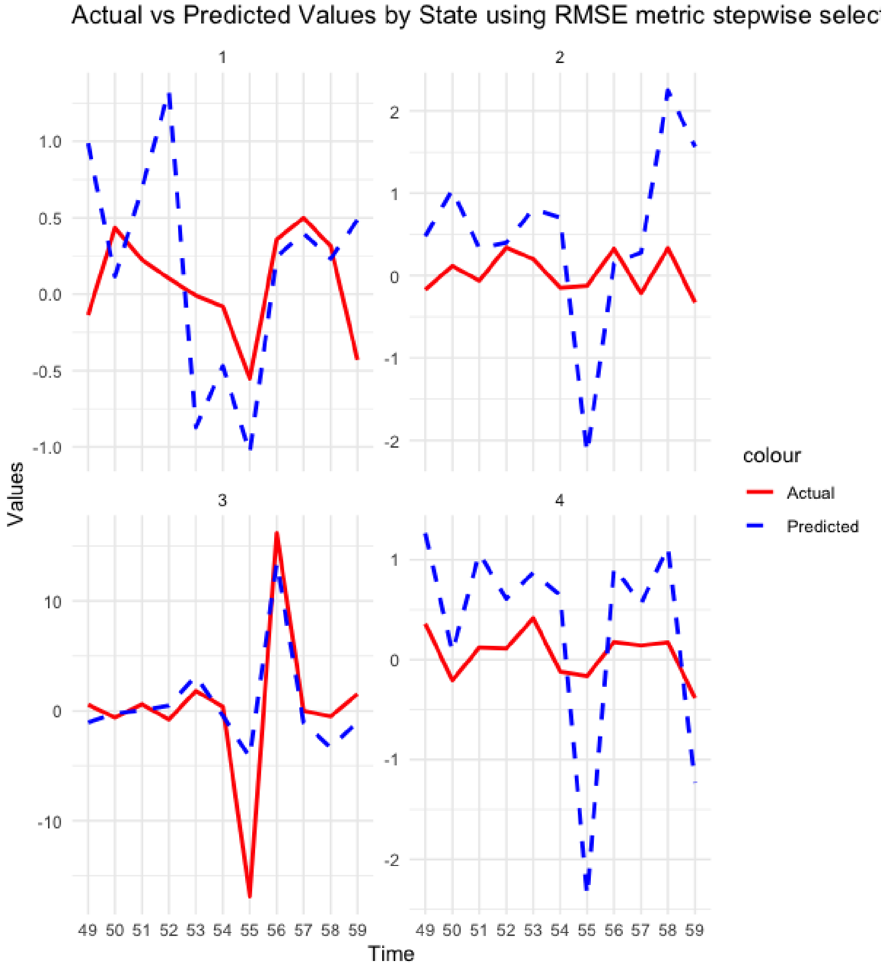

Figure 16 RMSE Stepwise selected model Predictions by State

Table 6 RMSE-MAE Stepwise model prediction metrics

State	RMSE	MAE
California (1)	0.681	0.554
Florida (2)	1.140	0.909
North Carolina (3)	4.140	2.520
Texas (4)	0.956	0.820

The RMSE-based stepwise selection method aimed to improve the predictive accuracy of the fixed effects model by iteratively adding variables that minimized the square error (RMSE) during model selection. The predictions moderately capture trends but exhibit substantial deviations, particularly in North Carolina (3) where there are large fluctuations. That can also be seen in the metrics table which we have the highest prediction errors. Comparatively with the BIC selection method we can see that the BIC-model had a smoother prediction line across all states but failed in dynamic fluctuations. The RMSE-model showed slightly better allignment with the trends in California (1) and Florida (2) but again struggled with North Carolina (3). In Texas (4) both models had similarly poor performance. To conclude though the RMSE-model had lower metric values across most states compared to the BIC-model. This suggests that minimizing RMSE during variable selection resulted in a model better suited for short-term predictions. After testing for Fixed effects model vs Pooled OLS the conclusion that the within-state variation is low and the fixed effects do not add much value is reached. To

https://doi.org/10.26219/heal.aueb.9465

conclude there are no better predictions using FE than using Pooled OLS model. Moving on to Seemingly Unrelated Regression.

4.6 Seemingly Unrelated Regression (SUR)

4.6.1 Introduction – Theory of SUR

In Fixed effects models each cross-sectional unit (states) is assumed to operate independently. However, economic and market forces often have dependencies among states, meaning that one state's imports could affect or be affected by anothes state's cement imports. Seemingly Unrelated Regression (SUR) is a technique that allows these cross-state dependencies to be taken into account by estimating separate regression equations for each state while though permiting their error terms to be correlated. SUR is applied to produce more efficient parameter estimates than by running separate OLS regressions for each state. For the problem with 4 states we will have the following multiple regression equations system:

$$Y_{1,t} = \alpha_1 + \sum_{l=1}^{L_1} \rho_l^{(1)} Y_{1,t-l} + \sum_{j=1}^{k_1} \beta_j^{(1)} x_{j,t-l_j^{(1)}} + \varepsilon_{1,t},$$

$$Y_{2,t} = \alpha_2 + \sum_{l=1}^{L_2} \rho_l^{(2)} Y_{2,t-l} + \sum_{j=1}^{k_2} \beta_j^{(2)} x_{j,t-l_j^{(2)}} + \varepsilon_{2,t},$$

$$Y_{3,t} = \alpha_3 + \sum_{l=1}^{L_3} \rho_l^{(3)} Y_{3,t-l} + \sum_{j=1}^{k_3} \beta_j^{(3)} x_{j,t-l_j^{(3)}} + \varepsilon_{3,t},$$

$$Y_{4,t} = \alpha_4 + \sum_{l=1}^{4} \rho_l^{(4)} Y_{4,t-l} + \sum_{j=1}^{k_4} \beta_j^{(4)} x_{j,t-l_j^{(4)}} + \varepsilon_{4,t},$$

where:

- $Y_{s,t}$: The Cement Imports for state s at time t
- α_s : State-specific intercept, baseline level unique to state s
- $\sum_{l=1}^{L_s} \rho_l^{(s)} Y_{s,t-l}$: The sum of lagged dependent variables (Lags 1 to 4(s)) for state s with coefficients $\rho_l^{(s)}$
- $\sum_{j=1}^{k_s} \beta_j^{(s)} x_{j,t-l_j^{(s)}}$: The sum of independent variables for state s, each possibly lagged by $l_j^{(s)}$ with corresponding coefficient $\beta_j^{(s)}$
- $x_{j,t-l_j^{(1)}}$: The j-th explanatory variable, lagged by $l_j^{(s)}$periods, in state s
- $\varepsilon_{s,t}$: The error term for state s at time t, errors are assumed to be contemporaneously [2]correlated across states

We should note that unlike running separate regressions for each state, SUR assumes that:

$$Cov\left(\varepsilon_t^{(s)}, \varepsilon_t^{(r)}\right) \neq 0 \quad \text{for } s \neq r$$

[2] It means that the error terms $\varepsilon_{s,t}$ from the different state equations at the same time t are not independent — they tend to move together

https://doi.org/10.26219/heal.aueb.9465

The Seemingly Unrelated Regressions (SUR) methodology is implemented. Each U.S. state (California, Florida, North Carolina and Texas) is modeled with its own regression equation to account for region-specific dynamics in cement imports. While the predictors and responses differ across states it is reasonable to expect that economic shocks, policy changes or other factors in one state may influence others. The traditional separate OLS models fail to capture these interdependencies. By applying SUR, jointly estimating the state-specific equations while accounting for cross-equation correlations in the error terms will lead to improvements in the efficiency of coefficient estimates and provide more accurate representation of the interconnected nature of state-level construction material trends if these exist.

4.6.2 Inference

For the inference part of the study, a variable selection methods based on the Bayesian Information Criterion (BIC) and Adjusted R-squared (Adj. R^2) is applied, both of which prioritize model parsimony while accounting for the trade-off between fit and complexity. The BIC penalizes the inclusion of unnecessary variables more strongly, helping to identify a minimal yet informative set of predictors. Adjusted R^2, on the other hand, evaluates how well the model explains the variance in the dependent variable after adjusting for the number of regressors. This inference-oriented SUR implementation offers an alternative to the prediction-driven RMSE-based selection that will be implemented later on and enables a comparative evaluation of trade-offs between forecast accuracy and model explainability.
Although the stepwise procedure selected a broader set of variables for each state, the primary interest lies in statistical inference. Therefore, highlighting and interpreting only the variables that are statistically significant at the 5% level is important.
Bellow is the table for the results of the algorithm:

Table 7. SUR Model Summary Table - Adjusted R-Squared

SUR Model Summary Table (Adjusted R-Squared-Based Variable Selection)

Equation	R-Squared	Adj R-Squared	RMSE	Significant Variables
California EQ1	0.20563	0.017136	4.02864	x12lag3*, x10*
Florida EQ2	0.396672	0.25351	3.51094	Yit_lag1***, Yit_lag2***, Yit_lag3*, Yit_lag4**, x12lag3**
North Carolina EQ3	0.34403	0.174382	3.69233	Yit_lag1***, Yit_lag2***, x12lag3*
Texas EQ4	0.324483	0.178121	3.68396	Yit_lag1***, Yit_lag2***, x12lag3*

The table presents the SUR model results using Adjusted R-Squared variable selection for each U.S. state. Among the included predictors, very few variables are statistically significant (at the 5% level), highlighting the challenge of finding consistent explanatory power across multiple panels. Notably, **Real Estate, Rental and Leasing Earnings, Thousands of Dollars,**

https://doi.org/10.26219/heal.aueb.9465

Quarterly, Seasonally Adjusted Annual Rate (x12lag3) appears significant across several states, suggesting it could be a key driver of short-term changes in imports.

4.6.3 Prediction

The algorithm for variable selection using the Root Mean Squared Error (RMSE) as a performance criterion is also used, aiming to enhance prediction accuracy solely. This model was constructed independently for each state and then estimated jointly using the Seemingly Unrelated Regressions (SUR) framework. Unlike traditional inference-focused modeling approaches, which aim for statistical significance, this method focuses on out-of-sample predictive performance. Below we present the results of this RMSE-optimized model.

Table 8. SUR Model Summary Table - RMSE

SUR Model Results Based on RMSE-Oriented Variable Selection

State	R-squared	Adjusted R-squared	RMSE	Significant Variables (p < 0.05)
California	0.0004	-0.0576	4.1790	x12lag3, x10
Florida	0.3967	0.2535	3.5109	Yit_lag1, Yit_lag2, Yit_lag3, Yit_lag4, x12lag3
North Carolina	0.3440	0.1744	3.6923	Yit_lag1, Yit_lag2, x12lag3
Texas	0.3245	0.1781	3.6840	Yit_lag1, Yit_lag2

As shown in the table, although Adjusted R-squared values remain relatively low, this is expected in prediction focused models. The main goal of the algorithm was to minimize RMSE which was achieved across the four state-level equations but without noticable improvement. The prediction accuracy can be seen on the next plot:

https://doi.org/10.26219/heal.aueb.9465

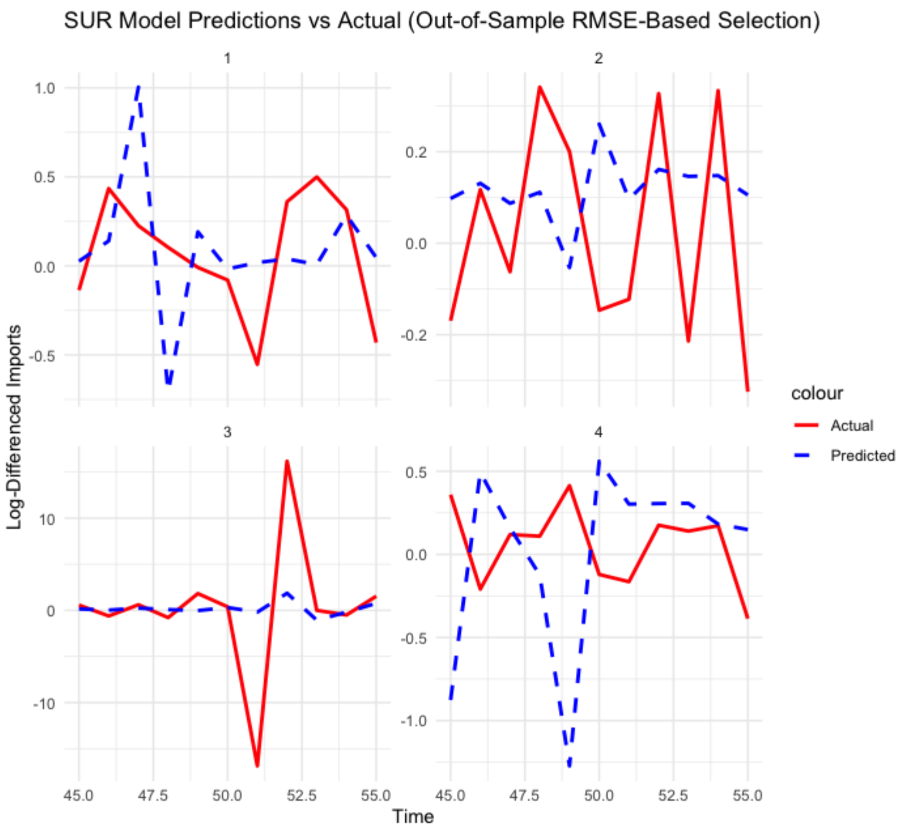

Figure 17. SUR Model Predictions using RMSE-based selection method

https://doi.org/10.26219/heal.aueb.9465

https://doi.org/10.26219/heal.aueb.9465

5 Chapter 5: Conclusion

5.1 Summary

This thesis focused on modeling and forecasting the percentage changes in cement imports in the United States, using panel data for four states: California, Florida, North Carolina, and Texas. The time series were transformed using log-differencing to achieve stationarity, and multiple forecasting methodologies were applied: dynamic panel data models (Fixed Effects, Pooled OLS), ARMA models with exogenous variables (ARIMAX) where predictors were selected using Lasso regression and Seemingly Unrelated Regressions method.

5.2 Method Comparison

When comparing results, it was found that the ARIMAX models—applied separately for each state—produced significantly lower forecasting errors (RMSE, MAE) than the panel models. The use of Lasso regression for variable selection helped reduce overfitting and simplified model specification. In contrast, the Fixed Effects and SUR models provided valuable insights into inter-variable relationships but did not perform well in terms of forecasting accuracy. Notably, North Carolina demonstrated poor model fit, possibly due to high variability or insufficiently informative predictors.

5.3 Conclusions and Applications

The analysis suggests that for predictive purposes, individualized ARIMAX models with state-specific exogenous variables outperform pooled panel data approaches. This makes the method more flexible and capable of capturing regional dynamics in cement import behavior. The findings could support commercial and construction firms in planning imports more efficiently and managing costs.

5.4 Suggestions for Future Research

Future work could explore more robust variable selection methods such as Stability Selection, or experiment with system-based ARIMAX variants (e.g., VARX or SUR models with ARIMA errors) to account for spatial dependencies. Moreover, incorporating a larger time frame and additional external macroeconomic indicators, for example energy price data could enhance the explanatory power of the models.

https://doi.org/10.26219/heal.aueb.9465

https://doi.org/10.26219/heal.aueb.9465

6 Bibliography

2024. *FEDERAL RESERVE BANK of ST.LOUIS.* 1 Nov. https://fred.stlouisfed.org/series/TTLCONS.

TTLCONS, FRED. 2024. *Federal Reserve of St.Louis.* 1 Nov. https://fred.stlouisfed.org/series/TTLCONS.

VAPGDPC, FRED. 2024. *Federal Reserve of St.Louis.* 1 Nov. https://fred.stlouisfed.org/series/VAPGDPC.

Vrontos. 2023. *Modeling and Predicting U.S. Construction Materials in the COVID era using Econometric Models and Machine Learning Techniques.* Athens: Ioannis D.Vrontos, Ioannis Galanis, MariosAngelis.

Miller, Osborne, Sheu, Sileo. 2023. "Technology and Market Power: The United States Cement Industry, 1974-2019*."

Kaur, Sanaz Tayefeh Hashemi1· Omid Mahdi Ebadati · Harleen. 2020. "Cost estimation and prediction in construction projects: a systematic review on machine learning techniques."

OEC. 2022. *OEC.* https://oec.world/en/profile/hs/cement.

Hsiao, Cheng. 2014. *Analysis of Panel Data.* 3rd Edition. Econometric Society Monographs, Cambridge.

Powell, James L. n.d. *Panel Data Models.* University of California, Berkeley.

Baltagi, Badi H. 2005. *Econometric Analysis of Panel Data.* John Wiley & Sons, Ltd.

Wooldridge, Jeffrey M. 2012. *Intoductory Econometrics: A Modern Approach.* 5th Edition. United States, Michigan State University: SOUTH-WESTERN CENGAGE Learning.

Nickell, Stephen. 1981. "Biases in dynamic models with fixed effects." *Econometrica* (Ec) 49 (6).

Arellano, Manuel, and Stephen Bond. 1991. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations ." *OXFORD JOURNALS* 58 (2).

Blundell, Richard, and Stephen Bond. 1998. "Initial conditions and moment restrictions in dynamic panel data models." *JOURNAL OF Econometrics* 87.

Elmousalami, Haytham H. 2019. "Artificial Intelligence and Parametric Construction Cost Estimate Modeling: State-of-The-Art Review." *Journal of Construction Engineering and Management.*

Hashemi, Sanaz Tayefeh, Omid Mahdi Ebadati, and Harleen Kaur. 2020. *Cost estimation and prediction in construction projects: a systematic review on machine learning techniques.* Springer Nature Switzerland .

Fengchang, Jiang, John Awaitey, and Haiyan Xie. 2022. *Analysis of Construction Cost and Investment Planning Using Time Series Data.* Sustainability .

Cheung, S.O., S. Thomas Ng, R. Marting Skitmore, and L.Y. Wong. 2000. *Prediction of Tender Price Index Directional Changes.* Hong Kong: Construction Management and Economics.

Ilbeigi, M., S.M.ASCE, B. Ashuri, M.ASCE, and A. Joukar. 2017. *Time-Series Analysis for Forecasting Asphalt-Cement Price.* American Society of Civil Engineers.

Afolabi, Adedeji Olushola, and Oluwamayowa Abimbola. 2022. *Application of machine learning in cement price prediction through a web-based system.* Department of Building Technology, College of Science and Technology, Covenant University, Ota, Nigeria: Adedeji Olushola Afolabi, Oluwamayowa Abimbola.

Kholodilin, Konstantin Arkadievich, Boriss Siliverstovs, and Stefan Kooths. 2008. "A Dynamic Panel Data Approach to the Forecasting of the GDP of German Länder." *Spatial Economic Analysis* 3 (2).

https://doi.org/10.26219/heal.aueb.9465

Matei, Iuliana. 2024. "Assessing the impact of pandemic measures on economic growth in a globalizing world: a non-linear panel analysis." *Applied Economics* 57:59.

Ben-Ahmed, Kais, Imen Ayadi, and Salah Ben Hamad. 2022. "COVID-19 impact on digital companies' stock return: A dynamic data analysis." *Finance Research Letters* 46.

Jian, Fengchang, John Awaitey, and Haiyan Xie. 2021. "Analysis of Construction Cost and Investment Planning Using Time Series Data." *sustainability* 14.

Vrontos, Ioannis D., Ioannis Galanis, and Marios Angelus. 2023. "Modeling and Predicting U.S. Construction Materials in the COVID era using Econometric Models and Machine Learning Techniques." *StonesNews.eu.*

Vrontos, Ioannis D., John Galakis, Ekaterini Panopoulou, and Spyridon D. Vrontos. 2024. "Forecasting GDP growth: The economic impact of COVID- 19 pandemic(2024a)." *Journal of Forecasting* 43.

Vrontos, Ioannis D., John Galakis, Ekaterini Panopoulou, and Spyridon D. Vrontos. 2024. "Modeling the Economic Impact of the COVID-19 Pandemic Using Dynamic Panel Models and Seemingly Unrelated Regressions(2024b)." *econometrics* 12.

Davidson, Russel, and James G. MacKinnon. 1993. *Estimation and Inference in Econometrics.* New York: Oxford University Press.

Nelson, Mark C., Masao Ogaki, and Donggyu Sul. 2003. "Dynamic Seemingly Unrelated Cointegrating Regression." *TECHNICAL WORKING PAPER SERIES* (National Bureau of Economic Reseearch).

Labra, Romilio, and Celia Torrecillas. 2018. *Estimating dynamic Panel data. A practical approach to perform long panels.* Revista Colombiana de Estadistica.

Taeil, Kim, Park Sunghwa, Kim Hanna, and Kwon Janghan. 2023. *Estimating the Impact of COVID-19 on International Trade: Cases of Major Countries Using the SUR Model.* sustainability.

https://doi.org/10.26219/heal.aueb.9465

7 Appendix

7.1 Variables & Transformation:

Variables used:		Transformation
2523 Portland Cement Aluminous Cement Slag Cement Etc (Kg)	Yit	Δ/ln
New Privately Owned Housing Units Authorized	x1	Δ/ln
New Private Housing Units Authorized by Building Permits, Units, Seasonally Adjusted	x2	Δ/ln
New Private Housing Units Authorized by Building Permits: 1-Unit Structures, Units, Seasonally Adjusted	x3	Δ/ln
Construction All Employees, In Thousands	x4	Δ/ln
All Employees: Construction: Construction of Buildings, Thousands of Persons, Seasonally Adjusted	x5	Δ/ln
All Employees: Retail Trade: Building Material and Garden Equipment and Supplies Dealers, Thousands of Persons, Seasonally Adjusted	x6	Δ/ln
Unemployment	x7	Δ/ln
Employment	x8	Δ/ln
Rental Vacancy Rates by State	x9	Δ/ln
Homeowner Vacancy Rates by State	x10	Δ/ln
Homeownership Rates by State	x11	Δ/ln
Real Estate, Rental and Leasing Earnings, Thousands of Dollars, Quarterly, Seasonally Adjusted Annual Rate	x12	Δ/ln
All-Transactions House Price Index for California, Index 1980:Q1=100, Quarterly, Not Seasonally Adjusted	x13	Δ/ln
Total Personal Income, Millions of Dollars, Quarterly, Seasonally Adjusted Annual Rate	x14	Δ/ln
Construction Wages and Salaries, Thousands of Dollars, Quarterly, Seasonally Adjusted Annual Rate	x15	Δ/ln
Wages and salaries by place of work (Thousands of dollars)	x16	Δ/ln
Construction Earnings, Thousands of Dollars, Quarterly, Seasonally Adjusted Annual Rate	x17	Δ/ln
All industry total	x18	Δ/ln
Private industries	x19	Δ/ln
Construction	x20	Δ/ln
Real estate and rental and leasing	x21	Δ/ln
Chain-Type Quantity Index for Real GDP: Construction (NAICS 23), Index 2012=100, Quarterly, Seasonally Adjusted	x22	Δ/ln
Chain-Type Quantity Index for Real GDP: All Industry Total, Index 2012=100, Quarterly, Seasonally Adjusted	x23	Δ/ln
Weather and Climate Billion-Dollar Disasters to affect	x24	Δ/ln

https://doi.org/10.26219/heal.aueb.9465