MODELLING AND MISSING DATA SIMULATION FOR ANNUAL NATURAL GAS CONSUMPTION IN PROVINCES OF TURKEY BY USING PANEL DATA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

BERK AKYÜZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
STATISTICS

SEPTEMBER 2024

Approval of the thesis:

MODELLING AND MISSING DATA SIMULATION FOR ANNUAL NATURAL GAS CONSUMPTION IN PROVINCES OF TURKEY BY USING PANEL DATA

submitted by **BERK AKYÜZ** in partial fulfillment of the requirements for the degree of **Master of Science in Statistics Department, Middle East Technical University** by,

Prof. Dr. Naci Emre Altun
Dean, **Graduate School of Natural and Applied Sciences** _______________

Prof. Dr. Ayşen Dener Akkaya
Head of the Department, **Statistics** _______________

Prof. Dr. Özlem İlk Dağ
Supervisor, **Statistics, METU** _______________

Examining Committee Members:

Prof. Dr. Ceylan Talu Yozgatlıgil
Statistics, METU _______________

Prof. Dr. Özlem İlk Dağ
Statistics, METU _______________

Assist. Prof. Dr. Fatma Yerlikaya Özkurt
Industrial Engineering, Atılım University _______________

Date: 04.09.2024

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name Last name : Berk Akyüz

Signature :

ABSTRACT

MODELLING AND MISSING DATA SIMULATION FOR ANNUAL NATURAL GAS CONSUMPTION IN PROVINCES OF TURKEY BY USING PANEL DATA

Akyüz, Berk
Master of Science, Statistics
Supervisor : Prof. Dr. Özlem İlk Dağ

September 2024, 71 pages

Energy sources play a vital role in the strategic and economic development of countries. Accurate prediction of main energy sources, such as natural gas or electricity, is essential for effective energy planning, policy formulation, and investment decisions. This thesis focuses on modeling of annual natural gas consumption in provinces of Turkey using real-life panel data. Performances of fixed effect, random effect and transition models were compared and discussed. Additionally, this thesis addresses the challenges of missing data mechanisms (completely at random, at random, not at random) through simulation and evaluation of various imputation methods (last observation carried forward, mean imputation, multiple imputation). The performances of these methods were compared with complete data model results. According to the modeling results, the random effects model outperformed due to its ability to capture inter-province differentiation. The importance of figuring out which missingness mechanism is present was apparent in the missing data simulation study. Additionally, multiple imputation methods gave more accurate results compared to other imputation techniques.

Keywords: Fixed Effects Panel Data Model, Random Effects Panel Data Model, Transition Panel Data Model, Missing Data Mechanisms, Imputation Techniques

ÖZ

TÜRKİYE'DEKİ İLLERİN DOĞAL GAZ TÜKETİMLERİNİN PANEL VERİ KULLANILARAK MODELLENMESİ VE KAYIP VERİ SİMÜLASYON ÇALIŞMASI

Akyüz, Berk
Yüksek Lisans, İstatistik
Tez Yöneticisi: Prof. Dr. Özlem İlk Dağ

Eylül 2024, 71 sayfa

Enerji kaynakları, ülkelerin stratejik ve ekonomik gelişmelerinde hayati bir rol oynamaktadır. Etkili enerji planlaması, politika oluşturma ve ülke çapındaki yatırım kararları için enerji tüketiminin, özellikle doğal gaz ya da elektrik gibi ana kaynaklar için, doğru tahmin edilmesi çok önemlidir. Bu tez çalışması, Türkiye illerinin yıllık doğal gaz tüketiminin gerçek hayattaki panel veriler kullanılarak modellenmesine odaklanmaktadır. Sabit etki, rastgele etki ve dinamik modellerin performansları karşılaştırılmakta ve tartışılmaktadır. Ek olarak, bu tez, simülasyon çalışması üzerinden çeşitli kayıp veri giderme yöntemlerinin (son gözlem değerini ileri taşıma, ortalama atama, çoklu atama) değerlendirilmesini ve kayıp veri mekanizmalarının (tamamen rastgele, rastgele, rastgele değil) getirdiği zorlukları ele almaktadır. Bu yöntemlerin performansları, kayıp veri içermeyen modelin sonuçlarıyla karşılaştırılmıştır. Modelleme sonuçlarında, rastgele etki modeli, şehirler arası farklılaşmayı yakalayabilmesi sebebiyle daha iyi sonuç vermiştir. Simülasyon çalışmasındaz, kayıp veri mekanizmasının ne olduğunu anlamanın önemi ortaya çıkmıştır. Arıyca, çoklu değer atama yöntemleri daha doğru sonuçlar vermiştir.

Anahtar Kelimeler: Sabit Etkiler Panel Veri Modeli, Tesadüfi Etkiler Panel Veri Modeli, Dinamik Panel Veri Modeli, Kayıp Veri Mekanizmaları, Kayıp Veri Giderme Teknikleri

To My Family

for their endless love and support

ACKNOWLEDGMENTS

First and foremost, I am extremely grateful to my supervisor, Prof. Dr. Özlem İlk Dağ, for her continuous guidance and patience during my study. Without her encouragement, it would be impossible to complete my thesis study while working a full-time job.

I would also like to thank Dr. Cankut Kaan Bolat for his valuable support, suggestions, and close friendship from the beginning to the end of my master’s journey. This friendship has been one of the most valuable outcomes I have gained from the graduate program.

I also wish to express my appreciation to my classmate Mert Kaşif Ceylan for his sincere friendship and contributions to my academic achievements. In addition, I sincerely appreciate their guidance and friendship, dear research assistants İrem Tanrıverdi, Rana Arslan, Serenay Çakar, and Petek Aydemir who were always ready to help.

Ebru Aksoy, my manager at my previous job, deserves special thanks as well. She and my colleagues provided me with all the support I required in accordance with the academic calendar, despite the fact that I was working full-time.

Finally, I am eternally grateful to my family for being in my life. As always, they have continuously shown me love and backing throughout my master’s program.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ABBREVIATIONS

abs: absolute

AIC: Akaike Information Criterion

ANN: Artificial Neural Networks

ARIMA: Autoregressive Integrated Moving Average

EPDK: Enerji Piyasası Düzenleme Kurumu

GAZBİR: Türkiye Doğalgaz Dağıtıcıları Birliği

GDP: Gross Domestic Product

GEE: Generalized Estimating Equations

LOCF: Last Observation Carried Forward

LSE: Least Squares Estimation

MAPE: Mean Absolute Percentage Error

MAR: Missing at Random

MCAR: Missing Completely at Random

MI: Multiple Imputation

MLE: Maximum Likelihood Estimation

MNAR: Missing Not at Random

NGC: Natural Gas Consumption

REML: Restricted Maximum Likelihood

RF: Random Forest

RI: Random Indicator

RMSE: Root Mean Squared Error

SSE: Sum of Squared Errors

std: Standardized

tr: Transformed

VIF: Variance Inflation Factor

CHAPTER 1

INTRODUCTION

Increase in energy requirement is one of the most important consequences of growing world population and globalization. In order to achieve sustainable development, countries have been trying to find the best solutions in terms of reducing their energy dependencies and minimizing environmental impact. In this context, natural gas has emerged as a vital energy source for many countries in the world. Turkey, a country that is quickly industrializing, has been relying more on natural gas recently in an effort to lessen its dependency on imported coal and oil, which are expensive and have serious environmental consequences. Natural gas is significant for Turkey not only because it can meet the national energy needs, but also because it has the potential to support sustainable growth by lowering emissions, increasing energy efficiency, and encouraging the industrial development. In this perspective, natural gas is viewed as an essential part of Turkey's energy balance and a resource that will be crucial to the country's future growth.

When natural gas is considered, it is seen that Turkey is a highly dependent country even though recent discoveries have been occurred. According to the latest official report based on 2021 data, Turkey produces 0.67 % of the annual natural gas supply, with the remaining 99.33 % coming from various sources through imports from corporations in foreign countries which are Russia (52% of total import), Iran, Azerbajian, Algeria, and Nigeria (EPDK 2021). In addition to this information, it is asserted that the discovery of natural gas in Filyos will enable Turkey to meet approximately 30% of its national consumption with domestic resources in following years as a result of Sakarya Gas Field Development Project (TPAO 2022). However, it is a fact that Sakarya Gas Field Development is a long-term project. Turkey's energy needs and natural gas dependency continue to be a current and critical issue.

When 59.854 million Sm^3 annual natural gas consumption in Turkey is analyzed by sector, residential consumption is 27.7%, consumption for electrical generating is 34,81%, and industrial consumption is 25.54% when the sectoral distribution of 2021 is evaluated as a percentage (EPDK 2021). Thus, it can be concluded that conversion sector, industrial requirement and household usage are three main components of consumption. Considering Turkey's growing population and developing industrial sector, increases can be expected in natural gas consumption throughout the years. In addition, natural gas distribution is not available for all cities, especially for rural areas, in Turkey. By the end of 2021, 81% of the population would have access to natural gas, with an increase of 1.8 million people each year over the previous five years (GAZBİR, 2021). The population that has been reached and the number of active users will grow over time, taking into account that some natural gas users who have natural gas access are not active currently. Considering all of this information, modelling and interpreting Turkey's natural gas consumption is a strategic concern for policymakers, including politicians, investors or analysts in order to understand energy dynamics.

1.1 Objective of the Thesis

Especially in developed and developing countries, understanding energy requirement dynamics is a significant concern for decision makers. Considering this issue, annual natural gas consumption in Turkey is estimated by using panel data in province level. Examining earlier studies in the literature reveals that time series data has been used most frequently in various time scales. In addition, models in earlier studies have been developed by using national data of Turkey or specifically for a single city, which can be broken down into daily, monthly, or annual time periods. However, there is no previous study that implements panel data approach for modelling energy consumption in provinces of a country. The main objective of this study is to suggest the most suitable model for annual natural gas consumption in Turkey by using panel data consisting of provinces. In addition, this study also provides a comparative simulation of missing data methods and imputation techniques.

The modelling part includes fixed effects (marginal) model, random effects, and transition (dynamic) model for estimating natural gas consumption in Turkey. The missing data simulation part focuses on missingness mechanisms that cover missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) cases. As the imputation techniques, single and multiple imputation methods are applied and compared, which are last observation carried forward, mean imputations, and multivariate imputations.

Throughout the study, the R programming language is used in the modeling phase and missing data analysis (R Core Team, 2023).

1.2 Organization of the Thesis

There are five chapters in this thesis study. The goal of the thesis, its background and structure are briefly introduced in Chapter 1. The techniques for different statistical models used in literature are explained and summarized in Chapter 2. In Chapter 3, panel data approach is reviewed, and the methodologies for the fixed effects, random effects and transition (dynamic) models are explained. This chapter also focuses on missing data mechanisms and imputation techniques that are discussed in this thesis study. In chapter 4, the data prepared for this study is presented along with empirical results of model comparisons and missing data simulation part. Chapter 5 concludes with a summary of the study based on the findings. Moreover, limitations of this study and insights for further research are discussed.

CHAPTER 2

LITERATURE REVIEW

Studies on modelling natural gas consumption in Turkey are available in the literature. While the majority of these studies focus on modeling and forecasting the total consumption in Turkey over time series, some studies are carried out in specific provinces. It can be seen that the time series being used vary in these studies as daily, monthly, and yearly according to objective of authors. There are also studies that use hourly consumption data. While building statistical models with various methods, different relationships are examined by using explanatory variables such as population and economic growth.

In this section, models in articles in the literature on natural gas consumption are examined. In addition, information about missing data imputation is also provided in the literature.

2.1 Statistical Models for Natural Gas Consumption in Turkey

Energy-related studies are strategically important, particularly for nations that rely heavily on energy. For this reason, a variety of modeling studies, particularly for fundamental energy resources, are available in the literature for all energy sources. Due to the scope of this thesis study is mainly concentrated on natural gas consumption, models in literature are reviewed based on natural gas as the main interest.

The range of studies and topics covered by statistical modelling research on natural gas consumption in Turkey includes estimation, forecasting in different time periods, and the relationship between consumption and other variables mainly consisting of economic indicators. It can be stated that these studies generally focus on time series for Turkey or a specific province. Annual natural gas consumption modelling and forecasting studies in the literature have used various statistical techniques such as optimization algorithms (Toksari, 2012), autoregressive integrated moving average (ARIMA) models (Oztürk & Oztürk, 2018), grey prediction models (Boran, 2015), artificial neural networks (ANN) and support vector regression (Kayakuş, 2020). Furthermore, the relationship between natural gas consumption and other possible variables such as economic indicators is an important concern in the literature. Especially in hybrid model studies, which try to achieve better forecasting accuracy compared to simple linear regression, dependent variables such as gross national product, economic growth, and population considered in national level (Ozdemir et al., 2016). Similarly, in another study, same dependent variables were selected in hybrid genetic algorithm based on linear regression (Karadede et al., 2017). Another study that used multiple regression models to examine annual natural gas consumption stated that economic indicators were the most significant predictors of annual natural gas consumption, with the nonlinear multiple regression model standing as the most accurate model (Sen et al., 2019).

In addition to these studies, effects of natural gas usage on energy policies, environmental impact, and economic growth have been explored. The resarch emphasized how crucial it is to comprehend the variables affecting the use of renewable energy sources and how energy policies can help lessen the dependency on natural gas in Turkey (Ozcan et al., 2016). In terms of energy policies and environmental impact, natural gas is considered as a clean source of energy. While examining the effects of energy resources on the quality of the environment and economic growth, Eygu and Soğukpınar showed that natural gas consumption increases economic growth and affects environmental quality in a positive way (Eygu & Soğukpınar, 2023).

There are also studies on monthly or daily consumption in addition to studies that concentrate on annual natural gas consumption. For instance, based on the amount of natural gas imported and utilized for heating and electricity, Erdem and Kesen calculated Turkey's monthly natural gas consumption (Erdem & Kesen, 2020). Since the time interval in these studies was narrowed from annual to monthly or daily, different variables such as temperature values were used and province-based time series were considered. Taspinar focused on estimating short term residential natural gas consumption in Turkey by using time series methods with air temperature as the dependent variable (Taspinar, 2015). In previous study of Taspinar (2013), similar time series methods were applied to forecast short-term (daily) natural gas consumption in Sakarya by considering meteorological variables which were air temperature, cloud cover, humidity, atmospheric pressure, and wind speed. In this study, air temperature and cloud cover were found highly effective on daily natural gas consumption. In another daily time series study based on Istanbul, temperature values were found one the most effective variable on natural gas consumption (Demirel et al., 2012).

As well as monthly and daily modelling studies, in specific cases, a daily forecasting approach is also possible depending on the design of the study. Dombaycı (2010) considered a different approach, using hourly energy consumption and temperature data. Based on his assumptions, he designed a proposed model house in Denizli, where is a province in Aegean region of Turkey, taking into account the temperature and energy consumption values. For example, he assumed area of the model house and heat transfer coefficients of building materials. Hourly energy consumption values were calculated for the heating requirements of this house using various reference values based on his assumptions and measured hourly temperature data. The energy consumption value of the previous hour was used to estimate the hourly energy requirement value for next hour. Based on his accurate ANN model predictions, he concluded that the energy requirement of a house can be calculated in advance according to building materials and temperature that provides efficiency by reducing energy costs.

When the literature is examined in detail, studies on natural gas consumption in Turkey generally focus on time series. Depending on the type of data examined, annual, monthly, daily, and even hourly analyzes are available based on study design.

The most studies in the literature concentrate on forecasting natural gas consumption and understanding possible covariates. Annual consumption studies show that population and economic indicators, which are two factors that have a direct impact on natural gas consumption, stand out. Studies based on shorter time periods, like daily data, have demonstrated the significant influence of air temperature on natural gas consumption. Taking this information into consideration, it can be said that no study has been conducted with panel data approach. On the other hand, none of these studies in the literature include any discussions on missing data concern. Consequently, using panel data that includes all provinces of Turkey, the goal is to identify the optimal model for annual natural gas consumption. While providing a different perspective, flexibility and advantages of panel data are utilized.

2.2 Missing Data Simulation Studies

In this study, it is also aimed to provide analyses for potential missing data scenarios by discussing imputation techniques and missing data mechanisms that are commonly observed in panel data studies.

Simulation studies conducted with real-life data for the mean imputation and last observation carried forward (LOCF) methods showed that these single imputation techniques do not accurately reflect the nature of panel data and that these methods do not take into account the relationship with other variables even in the MCAR case (Baraldi et al., 2017). Similar results were obtained in a study that focused on the MAR case (He et al., 2011). Although mean imputation techniques, another common single imputation method, gave acceptable results for the MCAR case, they were still biased and could not fully reflect the panel data structure (Roberts et al., 2017).

In a missing data simulation study conducted with multilevel data and addressing the MAR mechanism, it was revealed that although the 2l.norm method gave encouraging results, it could not fully capture the variability in the data, and limited usefulness was stated for not at random cases (Audigier et al., 2018). In another study that uses “pan” package (Schafer & Zhao, 2014) in R software revealed that in the MNAR and MAR cases, the multiple imputation resulted in biased outputs while more unbiased and acceptable estimates were obtained in the MCAR case (Grund et al., 2015).

When the studies with real-life data are examined according to missingness cases, the MCAR and the MAR studies are common in the literature compared to the MNAR case. It might be due to the structure of real-life data or the fact that the random deletion is more common and easy to apply. As stated, when missingness is random, single imputation techniques such as mean imputation or LOCF can be utilized, but may give biased results. Instead of these techniques, multiple imputation techniques such as MICE methods that are more suitable for the structure of the panel data are recommended. On the other hand, the MNAR case can be seen as the most challenging scenario in the missing data studies. In the literature, many studies emphasized that the standard imputation techniques fail to deal with imputation in the MNAR mechanism in the multilevel data and recommended more advanced methods such as selection models and pattern-mixture models (Guevara Morel et al., 2022). Multiple imputation methods also provide an additional advantage. It might be hard to distinguish the MAR and MNAR cases by using specific statistical tests in a real-life dataset. In addition, although most of the imputation techniques are not appropriate for MNAR mechanism, it was found that the multiple imputation techniques are the least biased for MNAR pattern (Baraldi & Enders, 2017). Therefore, it is useful to use multiple imputation techniques in cases where the missing data mechanism in the data is unclear.

2.3 Contribution to the Literature

As it can be seen in the literature review, several techniques are used in different time scales to estimate the natural gas consumption in Turkey. Even though particular studies concentrate on specific provinces, the impact of all provinces or regions on national natural gas consumption has not been investigated yet. Moreover, there is no study conducted using panel data for Turkey. In order to model annual natural gas consumption in Turkey and understand regional effects, it is decided to apply different panel data analysis methods. The first contribution is to examine natural gas consumption across Turkey using panel data approach. It can be considered as an important study to benefit from panel data not only in natural gas consumption but also in several different fields. Secondly, there is a comparative simulation study about missing data which includes different missingness mechanisms and imputation techniques. The literature contains missing data simulation studies focusing on real-life data that generally use single-level data. In studies conducted with multilevel data, there are limited simulation studies for MCAR and MAR cases, but the MNAR case is not considered. In addition, for both modeling and missing data sections, there is no study based on energy consumption using an official real-life database on a country or province basis. In this context, it is expected that this thesis study will provide a new perspective and provide enlightening directions and ideas for future studies in similar fields.

CHAPTER 3

METHODOLOGY

The goal of this thesis study is, as previously mentioned, to recommend the most appropriate model of Turkey's annual natural gas consumption in province level. Various techniques have been used in order to estimate energy consumption in literature which are mainly focused on time series. In contrast to studies in literature, this study is conducted using panel data. In addition to the modelling phase, missing data mechanisms and imputation methods are also discussed. In this chapter, after a brief explanation of the panel data approach, methodological information about models, missing data mechanisms and imputation techniques will be provided.

3.1 Panel (longitudinal) data

Panel (longitudinal) analysis is a statistical method in which repeated measurements taken from the same subject over time are used. It is a common practice in different disciplines such as economics, medicine, biostatistics, and social sciences. Stock prices by specific firms followed over time, development and well-being data of children measured annually, monthly unemployment rates across states can be given as examples of panel data. In general, although the terms “panel data” and “longitudinal data” are used interchangeably, panel data is more frequently used in econometrics and social sciences (Diggle et al., 2002). The expression “panel data” is preferred in order to maintain the integrity of this thesis study.

3.1.1 Features and Advantages of Panel Data

As stated earlier, panel data is used in many different areas such as analyzes of consumer tendency surveys, blood sample tests or macroeconomic indicators.

Numerous benefits of panel data are a major reason in its broad range of applications. Below is a summary of some advantages of panel data.

Combining time-series and cross-sectional approaches: Combining time-series analysis and the cross-sectional approach is made possible by panel data, which make use of more observations and offer a more thorough understanding of the data (Sousa & Lopez, 2008). Both the change over time for a specific subject can be observed, that refers to time series data, and instantaneous analysis of the situation at a specific time point can be interpreted as implemented in cross-sectional data.

Integration of fixed and random effects: Researchers can take advantage of both approaches' strengths combining fixed and random effects in panel data analysis. This combined approach provides controlling time-invariant heterogeneity between subjects and unobserved subject specific effects (Ahapova, 2020). Therefore, differentiated subjects or groups can be identified, which means a more efficient model with higher interpretability.

Accurate predictions by pooling the data: In panel data, repeated measurements are collected from different subjects. By adding information about more and more subjects, it is possible to get a more accurate explanation of subjects' behaviors (Hsiao, 1989). It also increases the explanatory power of the model in terms of capturing differences between groups such as boys and girls in a study.

Stable and reliable estimates: Panel-data models offer a greater degree of freedom and are more stable and reliable when estimating over short time periods and cross-sections (Guan et al., 2021). To accurately infer conclusions from the data, stability and reliability are essential.

Understanding temporal changes: Panel data studies can capture temporal variation within data, generating insights into trends and changes over time such as seasonal patterns or specific fluctuations (Tawfik et al., 2011). This ability is necessary to detect how variables change and interact over time.

Use of lag values of variables: As well as fixed effects, observations in previous time points can be used for both dependent and independent variables in panel data structure that contributes to the reliability and robustness of the analysis (Mood, 2010). It also enables researchers to understand time-dependent relationships in the model.

Availability of structural and reliable data: Although it cannot be denied that panel data collection might be a costly process, several important resources have been accumulated over time in developed and developing countries (Hsiao, 2007). Official surveys created for a specific purpose, such as economic tendency surveys, have been creating a larger data pool over time. Likewise, with the digitalization of institutions, the national data resources of countries have been expanding accordingly. These data sources are safe and regulated when provided from official sources.

Balance and unbalanced panel data: In panel data, data is collected from the subjects included in the research at certain time points. If each individual is observed in all time points, it is called balanced panel. However, in an unbalanced dataset, subjects are observed a different number of times. Even though balanced datasets are preferred and unbalanced panel data may present challenges in analysis because of its nature, it crucial to understand the data and consider the required assumptions in modelling phase.

3.1.2 Normality Assumption

The assumption about normal distribution in population with a common variance and additive error terms should be met by statistical analysis. It should be also noted that exact normal distribution is not a strict requirement and cannot be achieved generally in a real-life dataset. However, theoretical normal distribution assumption should be satisfied, at least approximately, in order to make inferences about unknown parameters such as population mean, variance or regression coefficients.

Box-Cox transformation is employed to reduce variance and improve linearity (Box & Cox, 1964). It is especially helpful when working with data that deviates from the normal distribution which is a common case in real-life data. The basic equation of the Box-Cox transformation is given:

$$y(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log\ y, & \text{if } \lambda = 0 \end{cases}$$

where the original response variable is denoted by y, and the transformation type is determined by λ. Both positive and negative values of y can be handled by the transformation and λ should be chosen to obtain the best solution to approach normally distributed values. Common Box-Cox transformations can be seen in Table 3.1:

Table 3.1 Box-Cox Transformation

Lambda value (λ)	**Transformed data (Y')**
-3	$Y^{-3} = 1/Y^3$
-2	$Y^{-2} = 1/Y^2$
-1	$Y^{-1} = 1/Y^1$
-0.5	$Y^{-0.5} = 1/(\sqrt{Y})$
0	log(Y)
0.5	$Y^{0.5} = \sqrt{Y}$
1	$Y^1 = Y$
2	Y^2
3	Y^3

3.1.3 Fixed Effects (Marginal) Model

Fixed effect panel data models, also called marginal models, are a commonly used tool in panel data analysis, particularly in econometrics and social sciences. Main objective of fixed effect models is to understand the differences between groups and trends over time. For instance, child development differences over time between boys and girls can be explained by a fixed effects model in a study. Within-subject correlation is taken into consideration when modeling the dependent variable as a function of independent variables in fixed effects. These models are designed to account for individual-specific effects that are constant over time by controlling unobserved heterogeneity at the subject level (Xu et al., 2007). With these models, it may be easier to interpret how groups differ from each other and how the variables affect them over time, but not for individuals one by one.

The simple form of fixed effects model can be seen in equation below:

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \cdots + \beta_k X_{kit} + u_{it}$$

where $u_{it} \sim N(0, \sum)$

Y_{it} is the dependent variable where $i = 1,2, \ldots N$ corresponds subjects and $t = 1,2, \ldots T$ indicates time points.

β_k is the coefficient of X_{kit}

$X_{k_{it}}$ indicates independent variables for $i = 1,2, \ldots, N$ and time points $t = 1,2, \ldots, T$

u_{it} is the error term with a variance of Σ. Some structures for Σ might be autoregressive, Toeplitz, or Exchangeable structures.

3.1.4 Random Effects Model

The individual-specific effect in the random effects model is a random variable that has no correlation to the explanatory variables. This model might be a good option for handling unobserved heterogeneity because it assumes that the individual-specific effects are uncorrelated with the independent variables. For example, random effects models can be used when examining individual child developmental differences between children rather than comparing the development of girls and boys on a group basis. When there is a concern regarding the endogeneity of the explanatory variables in the fixed effects model, this model is especially helpful (Hassaballa, 2022). The fixed effects panel data model and the random effects panel data model are frequently contrasted. Subject-specific effects are handled by random effects model as random variables within the model. It is important to highlight the critical assumption that there is no correlation between these random variables and explanatory variables in the model. Furthermore, the type of data and the research question of study influence the decision between fixed effects and random effects models.

A random effects model can be built with only a random intercept or both random intercept and slope(s). Depending on the type of data and the requirements of the model, it can be decided that only a random intercept term is sufficient or additional random slope terms are required. AIC, BIC, and deviance can be considered as evaluation criteria that help to decide whether there is a need for random slope terms in a model or not. The simple form of random effects model can be seen in equation below:

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \cdots + \beta_k X_{kit} + \alpha_{0i} + \alpha_{1i} Z_{1it} + \cdots + \alpha_{\ell i} Z_{\ell it} + u_{it}$$

where $u_{it} \sim N(0, \sum)$

Y_{it} is the dependent variable where i = 1,2, …. N corresponds to subjects and t = 1,2,…. T indicates time points.

β is the coefficient of X_{it},

X_{kit} indicates independent variable k = 1,2,…, K for $i = 1,2, \dots, N$ and time points $t = 1,2, \dots, T$,

α_{0i} is the unknown intercept for subject i,

$\underline{\alpha_i} = (\alpha_{1i}, \dots, \alpha_{\ell i}) \sim N(\underline{0}, \underline{D})$ are random slope terms for subject i,

Z_{it} is a subset of X_{it},

u_{it} is the error term with a variance of Σ which is an identity matrix. Temporal correlation and inter-subject differentiation are considered in D matrix.

3.1.5 Transition (Dynamic) Model

In transition (dynamic) model, lagged values of dependent variable can be added to the model with other independent variables. These models are used to investigate the effects of time-varying factors on the dependent variable. For example, in a panel study where the development of children is modeled, if the height of the children in previous years is added to the model in addition to the characteristic features of the children, a dynamic model is obtained. In other words, children's development at a specific time point is tried to be explained by their development in previous years. Adding lagged values of variables turns the static panel data model into a dynamic one. Therefore, it can be stated that one of the main advantages of dynamic panel data models is their capacity to deal with endogeneity by using the dependent variable's lagged values as explanatory variables providing dynamic adjustment (Omolara et al., 2021).

The simple form of transition model can be seen in equation below:

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \cdots + \beta_k X_{kit} + \alpha_1 Y_{it-1} + \cdots + \alpha_m Y_{it-m} + u_{it}$$

where;

Y_{it} is the dependent variable where i = 1,2, …. N corresponds subjects and t = 1,2,…. T indicates time points.

β is the coefficient of X_{it},

X_{kit} indicates independent variable k = 1,2,…, K for $i = 1,2, \dots, N$ and time points $t = 1,2, \dots, T$,

α_m is the coefficient of lag variables,

$Y_{i,t-m}$ is the lag variable where $i = 1,2,\dots,N$ stands for entity, $t = 1,2,\dots,T$ corresponds to time and $m = 1$ for lag-1 model,
u_{it} is the error term with a variance of Σ, which is an identity matrix.

3.1.6 Model Comparison Metrics

In this part, the evaluation criteria used in this study are briefly explained. The main evaluation criteria considered in this study are Akaike Information Criterion (AIC), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Correlation (r).

- Akaike Information Criterion (AIC): $\mathrm{AIC} = 2\mathrm{K} - 2\ln(\mathrm{L})$

where K indicates the number of independent variables in the model and L is the log-likelihood.

It is a statistical technique that is used to how well a model fits the actual data. It is possible to compare several models and decide the best fit by using AIC that estimates the amount of information lost by a built model.

As shown in the formula of AIC, it becomes lower for model with high log-likelihood that means more information is captured and the model fits the data better. However, there is an additional penalty term that considers the higher parameter complexity. Therefore, there is a two-way effect from the number of parameters and captured information. The lower AIC is considered significantly better in the model comparison.

- Root Mean Squared Error (RMSE): $\sqrt{\frac{\Sigma(f_t-a_t)^2}{n}}$
- Mean Absolute Percentage Error (MAPE): $\frac{1}{n}\sum\frac{|a_t-f_t|}{a_t}$

where *f* represents fitted values from the models and *a* represents actual values from the original data.

- Correlation (r) value measures the linear relationships between fitted and actual values.

Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Correlation (r) are main performance indicators for a model that evaluates predicted and actual values. RMSE basically measures the standard deviation of the error terms. Similarly, MAPE calculates the average of absolute percentage differences between predicted values and the actual values. For both RMSE and MAPE, the lower values show more accurate predictions and fit the data better. As explained in correlation definition, when there is a high correlation between predicted and actual values, a better model performance is achieved.

- Percentage Bias: Percentage bias is a comparison metric which is calculated as follows:

$$\left(\frac{\hat{\beta}-\beta}{\beta}\right) * 100$$

where β is the true value of the slope parameter for population obtained from based marginal model with original data, and $\hat{\beta}$ is the estimated value of the coefficient obtained from marginal models with imputed data. Therefore, it can be said that the lower and closer to 0 percentage bias is desired.

3.1.7 Estimation Techniques

Generalized Estimation Equations (GEE): In marginal models and transition models in panel data analysis, generalized estimating equations (GEE) method is used as the estimation technique. The concept of GEE (Liang & Zeger, 1986) is a method to estimate the parameters of a generalized linear model (GLM) when working with correlated data, especially in panel data studies where observations within the same subject are not independent. GEE method can be used for both continuous and discrete response cases (Zeger & Liang, 1986). Diggle, Liang and Zeger (2002) explained that GEE is a superior technique to its competition with least squares estimation (LSE) and maximum likelihood estimation (MLE). One of the primary advantages of GEE method is its robustness to selection of the correlation structure by utilizing the “sandwich estimators”. Sandwich estimators, also referred to as robust standard error estimators, are used to acquire more reliable estimates of the standard errors of a regression model. The main aim of the sandwich estimators is to adjust the error structure of a model for the case of heteroskedasticity or a misspecified covariance (White, 1980). It provides more trustworthy inferences thanks to robust standard errors that could be valid even in cases where model’s assumptions regarding error variance are not truly specified. In other words, even if the working correlation structure is not correctly chosen, the standard errors and parameter estimates are consistent. As well as its benefits and extensive usage, it should be noted that there are some limitations and drawbacks to the GEE method. Even though its iterative nature provides consistent results, it might be computational challenges by working with large datasets because of intensive computation work (Fitzmaurice et al., 2011). On the other hand, a very small sample size may lead to less accurate results and lower statistical power. In addition, when there are completely at random missing cases in the dataset, the GEE method may give biased results. For such situations, extensions of the GEE might be considered.

In random effects model, restricted maximum likelihood estimation method (RMLE), first introduced by Patterson and Thomson (1971), can be used. It is a common method in estimating panel data models. Generally, maximum likelihood estimation (MLE) method constructs the model parameters and maximizes the likelihood function. RMLE method specifically focuses on providing more accurate estimates for model parameters, including fixed and random effects, by utilizing a restricted version of the likelihood function. As opposed to conventional MLE approach, RMLE is able to give estimates that are less biased while working with small sample sizes. Therefore, it is an important advantage that RMLE is particularly useful for small datasets or complex models (Laird & Ware, 1982). If the sample size is large enough, the difference between MLE and RMLE methods becomes unimportant.

3.2 Missing Data Mechanisms

In panel data analysis, missing data might be a serious problem that can significantly affect the reliability and validity of research findings. Panel data, which is the observation of subjects over time, frequently has missing observations due to several reasons including data collection errors, dropouts, and equipment failures. Missingness is frequently seen both in physical survey participation and in surveys collected online. In addition, some international or national data may also contain missing data for various reasons. For example, all information about relatively small countries or provinces cannot be followed naturally. These situations may affect the analysis negatively in terms of skewed estimates, lower statistical power, and less reliable results (Dong & Peng, 2013). Therefore, it is critical to understand the missing data mechanism in order to deal with it appropriately.

As mentioned above, understanding missing data mechanisms is an essential step to approach data as it can significantly impact the study results. Although some missing data types might be ignorable, some of them might have a pattern behind them. There are three main types of missing data mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR).

- **Missing Completely at Random (MCAR):** When there is no relationship between the probability of missing data and any observed or unobserved data, MCAR case occurs. In other words, the missingness is completely arbitrary and unrelated to other variables in the study. This indicates that the absence of data is entirely random and does not differ from the observed data.
- **Missing at Random (MAR):** If missingness is solely dependent on observed data and unrelated to unobserved data, there is MAR case. This mechanism suggests that the observed data can be used to predict the missing data, making the missingness of the data ignorable (Wang & Hall, 2009). In this situation, missing data can be predicted by using observed data which is already available.

- **Missing Not at Random (MNAR):** Even after considering the observed data, MNAR occurs when the missingness is related to unobserved data. In this case, it can not be ignored because absence of the data is not entirely random or based on observed data. It is not possible to use known information to predict missing part.

To clarify with an example in real-life, a clinical study about blood pressure can be considered. If the participants' blood pressure values are missing due to the measuring device error, there is MCAR case. Measuring device error is obviously not related to scope of the study. If participants left the study early because of severe side effects, there is MAR case. The reason for leaving the study is side effects. This situation is related to the study itself which depends on observed data. Finally, if the participants purposefully avoid taking samples because they feel uncomfortable, there is MNAR case. The participant's feeling of discomfort is not related to observed data in the study. In addition, it is not a situation by chance like measuring device error. Therefore, there is a case that is not at random.

It is noteworthy that distinguishing between MAR and MNAR might be difficult as the reasons may not be captured by using observed data (Beesley et al., 2009). Even though, missingness mechanisms are not generally testable directly, exploratory data analysis, basic statistical tests such as mean comparison between groups, prior knowledge about the study, and expert comments might be helpful to understand the reasons behind missingness.

3.3 Imputation Techniques

As missing data is a common issue in panel data analysis, imputation techniques may seriously affect the results of an analysis. Depending on the missing data mechanism, the impact of missing observations on the analysis may be more significant. Therefore, it should be clearly understood what type of missing data is observed and the imputation technique should be chosen carefully. As well as three different missingness mechanisms, imputation techniques can be broadly classified into single and multiple imputation methods.

3.3.1 Single Imputation Methods

As an approach to deal with missing observations, single imputation methods are widely used due to their ease of implementation. In a single imputation method, missing observations are replaced with single values which decrease the variability in the data. For instance, mean imputation involves substituting the mean of the available cases for a missing value on a particular variable. In other words, missing observations are directly replaced with a single average value. It is also similar in single regression imputation in which the information in complete observations is utilized to predict missing values. In the similar manner, there are many single imputation applications, such as replacing the most repeated value in the observed data or imputation based on the median value. In panel data, these methods might be modified among subjects or time points also.

In this study, subject mean imputation, occasional mean imputation, and last observation carried forward (LOCF) method, which are widely used among single imputation methods, are discussed.

3.3.1.1 Subject Mean Imputation

In subject mean imputation in panel data, missing observations are replaced with mean of the specific variable for that subject. As can be expected, using the relevant subject's mean value negatively affects the capturing trend over time and is very likely to disrupt the within-subject covariance structure. In addition, the relationships between variables are not considered in subject mean imputation. Therefore, it is obvious that the information based on other variables and subjects is not utilized.

3.3.1.2 Occasional Mean Imputation

In occasional mean imputation, which is another mean imputation method, missing data is replaced with the mean of the related time point. Obviously, this method also does not account for the impacts of other variables. Furthermore, in occasional mean imputation, the same imputation is applied for all subjects. In other words, there is absolutely no consideration given to subject-specific differences which are quite important to capture and preserve variability in the data.

3.3.1.3 Last Observation Carried Forward (LOCF)

In LOCF method, another common single imputation technique, replacement occurs with the last observed value of that variable for each subject without any calculations. Although this method is used especially in clinical trials or studies that require regular follow-up, it has some disadvantages. LOCF imputation can seriously change the mean or variance-covariance structure in the data and distort the trend over time. Additionally, this method is quite sensitive depending on where the missing observation is and what kind of missingness mechanism exists.

3.3.2 Multiple Imputation Techniques

In contrast to single imputation methods, multiple imputation techniques estimate several values for missing observations in order to obtain more accurate results. Using the distributions and correlations between the observed variables in the data set, multiple imputation generates plausible values to fill in the missing values (Rubin, 1987). These plausible values provide information about possible imputation scenarios. By this way, it is more potential to converge the accurate imputed values for missing data. Multiple imputation process can be divided into two main steps: repetitively obtaining several imputed data sets with replaced missing information and finalizing results by analyzing these imputed data sets. As seen from the definitions, multiple imputation methods can consider the missingness mechanism, relationships between variables, trends over time, and subject specific changes in the data. Another important point is that using multiple imputations in different missing data mechanisms provides an advantage in obtaining a more accurate result not only in MAR cases but also in nonignorable missing data (Demirtas, 2004). Especially when there is a reason behind the missingness, this can be captured with multiple imputation while single imputation methods do not have the processes to evaluate this. Even though the computational time of multiple imputation techniques might be seen as a disadvantage, especially with new methods and higher computational power, this becomes a rare issue (Li et al., 2016). Furthermore, for the purpose of improving imputation quality and choosing the best method, it is helpful to have a solid understanding of the algorithms employed in multiple imputation techniques.

Although there are several multiple imputation techniques, in this study, three different methods are analyzed which are based on mice (multivariate imputation by chained equation) package in R.

3.3.2.1 2l.norm

One of the imputation methods available in the mice package is the 2l.norm function that stands for “two level normal”. It is designed for two-level continuous data structures such as repeated measurements over time or subject within different groups. It is not applicable for a binary variable. As the name implies, it assumes the normality in the data. Therefore, imputation results of data sets that do not ensure normality assumption may be biased and inefficient. For more accurate imputation in hierarchical data structures, especially in panel data, the two-level imputation model implemented by 2l.norm method helps to capture both the variability within and between individuals (Van Buuren, 2018). In other words, both individual level variations and between group level variations are considered. Because it is an iterative imputation process using observed data and the requirement for estimating hierarchical parameters, computational time can be intensive, especially in large data sets.

3.3.2.2 Random Forest

Random forest method is a machine learning algorithm which builds multiple decision trees during training phase in order to obtain an output for each individual tree. It is possible to use random forest algorithms for both classification and regression purposes which can be seen as an advantage in terms of applicability in variety of datasets (Breiman, 2001). Output of the algorithm for classification tasks is the result that the majority of the trees return. In regression tasks, the average prediction obtained by each individual tree is achieved as the output. In the context of missing data imputation, the random forest algorithm works with observed values in the dataset to predict missing observations by using its tree-based approach. It should be also noted that there is no requirement for normality assumption in random forest method which makes it capable of modeling non-linear relationships between variables (Hastie, Tibshirani, & Friedman, 2009). It is particularly beneficial in panel data where relationships between variables can be non-linear and more complex. Even though it is a flexible and applicable method, the process behind the random forest algorithm might be hard to understand clearly. Because of the complex nature of the ensemble approach, the algorithm itself is frequently categorized as a "black-box" algorithm, making the imputation process more difficult to understand (Strobl et al., 2007). Therefore, interpretability can be seen as a disadvantage of this method. In addition, computational intensity should be carefully considered especially for large datasets or high number of trees. Random forest algorithm might be computationally expensive, especially in cases where the number of variables is large or the number of trees is set higher.

3.3.2.3 Random Indicator

Random indicator method is another technique in mice package that externally adds an indicator variable to the dataset for each column where there is a missing observation. The main function of these indicators is remarking whether the observation in the original dataset is missing or available (Van Buuren & Groothuis-Oudshoorn, 2011). When such an indicator is created, the mechanism of missing data in the dataset is also revealed. Considering missing data mechanism in the imputation process directly can help in reducing bias in the imputed values and improve the quality of the imputation (Collins et al., 2001). Since it takes into account the missingness mechanism, the random indicator method can be expected to be effective for non-ignorable missingness cases. Moreover, it is possible to use the random indicator method in different imputation models including linear or non-linear relationships which provides an advantage in terms of flexibility (Van Buuren, 2018). It can be also stated that there is no prior assumption about normality. Compared to random forest algorithm, creating an indicator for missing observations can be considered a simpler method than building multiple decision trees. Additionally, interpretation is easier because random indicator method uses more clearly defined indicators rather than a black-box algorithm.

CHAPTER 4

DATA DESCRIPTION AND EMPRICIAL RESULTS

Methodological background and aim of the study are explained in previous chapters. In chapter 4, details of the natural gas consumption dataset will be presented. The main highlights of the exploratory data analysis part will also be given. Then, applied panel data models and missing data simulation part will be explained in detail respectively in separate sections with a comparative perspective.

4.1 Data Description

In this study, real-life annual panel data of Turkey is analyzed by using official open-source databases with a time range from 2014 to 2021. The reason why the time period is chosen between 2014 and 2021 is that Turkey's province-based natural gas consumption data starts from 2014. Although the analyzes cover all provinces in Turkey, it should be noted that since natural gas consumption in Tunceli is not widespread yet, this province is not included in the analysis due to lack of consumption values. There is no missing data in the dataset after Tunceli is excluded from the study.

The three main uses of natural gas consumption can be classified into industrial usage, electricity generation and heating as explained in the official energy institution reports (EMRA, 2021). To interpret these relationships intuitively, increasing industrial activity may be a factor that increases energy supply and therefore natural gas consumption. As electricity production increases, the natural gas consumption used for this generation must also increase. In addition, as the need for heating increases in parallel with the need for sheltering, natural gas consumption increases in a city. In this context, models have been constructed with variables that may be relevant when modeling natural gas consumption. There are also some variables that are not included in the final models but are used at the beginning of the study such as temperature, inflation, and GDP. Although temperature values are closely related to natural gas consumption, especially for heating purposes, this relationship becomes more meaningful when working with daily data (Taşpınar, 2015). Considering daily life practices, it is obvious that daily, and even hourly, natural gas consumption is related to temperature. Official temperature data including average winter temperature, average summer temperature, heating degree days, and heating cooling days obtained from Turkish State Meteorological Service were included in the model at first. However, after the first evaluation in this study, temperature related variables were not utilized because they were insignificant in annual data modelling. Moreover, although there are many articles (D. Sen et al., 2019; Y. Karadede et al., 2017; Özdemir et al., 2016) showing the relationship between economic indicators, such as GDP and inflation, and energy consumption, these variables were not used directly in this study because they were not significant compared to more detailed economic activity indicators that were included to the study and directly affect the natural gas consumption. As a result of the evaluation, final variables in the study and their codes are as follows:

- Natural Gas Consumption (NGC)
- Electricity Consumption Industrial (ELCIND)
- Number of Manufacturing Enterprises per Person (MANUF_PP)
- Population (POP)
- Natural Gas Unit Price for Industry (IND_PRICE)
- Regions of Turkey (REG)
- Years (YEAR)
- Provinces (ID)

Natural gas consumption (in million Sm^3) is the response variable in the study and obtained from yearly Turkish Natural Gas Market Report that is published by Turkish Energy Market Regulatory Authority (EMRA). Industrial electricity consumption (in Kwh), number of manufacturing enterprises, and population data are obtained from Turkish Statistical Institute (TUIK). Industrial natural gas price data (TL/Sm^3) is requested and obtained from Turkish Energy Market Regulatory Authority (EMRA). Regions of Turkey indicates seven geographical regions of Turkey and it is used to understand the consumption characteristics of the regions. As mentioned previously, the time range of years variable is between 2014 to 2021.

4.2 Exploratory Data Analysis

Exploratory data analysis (EDA) is conducted to develop intuition about dataset and to summarize and visualize the key characteristics. Additionally, EDA section aims to provide useful information, such as relationships between variables or the distribution of a single variable, before moving on to the modeling part.

4.2.1 Descriptive Statistics

Descriptive statistics of all numerical variables in this study are indicated in Table 4.1. Values are reported by using raw data which does not include any transformation.

Table 4.1 Descriptive Statistics

Desc. Stats.	NGC	POP	ELCIND	MANUF	IND_PRICE
Min Value	0.11	78550	3.0	217.0	0.903
1st Quartile	61.81	290618	398.2	897.2	0.957
Median	145.51	535817	966.5	1942.5	1.130
Mean	620.35	1015373	1465.4	5503.9	1.401
3rd Quartile	526.37	1062722	1841.0	4864.0	1.801
Max Value	8742.26	15840900	9582.0	143648.0	2.250

Based on Table 4.1, it can be clearly seen that the min and the max value of natural gas consumption (NGC) are far from the mean value. It is an indication that the provinces of Turkey have quite different consumption patterns from each other. This is quite normal because bigger provinces are regions where more energy is consumed for many reasons, while energy consumption is low in small and rural areas. In addition, mean of NGC is greater than median value which is a sign of right skewed distribution. In order to satisfy the normality assumption in response variable, a proper transformation might be required. Similar pattern and right skewed distribution can be captured in the population (POP), the industrial electricity consumption (ELCIND), and the number of manufacturing enterprises (MANUF) variables. As an additional note, it should be stated that the MANUF variable shows the total number of manufacturing enterprises in a province. It is considered that the number of manufacturing enterprises per person (MANUF_PP) might be more suitable for analysis and interpretation. Therefore, the MANUF_PP variable is used in the modelling part.

4.2.2 Normality and Box-Cox Transformation

Before moving to details of exploratory variables and modelling part, it is important and required to check normality in the response variable.

In order to understand the distribution of the response variable visually, the histogram in Figure 4.1 can be used.

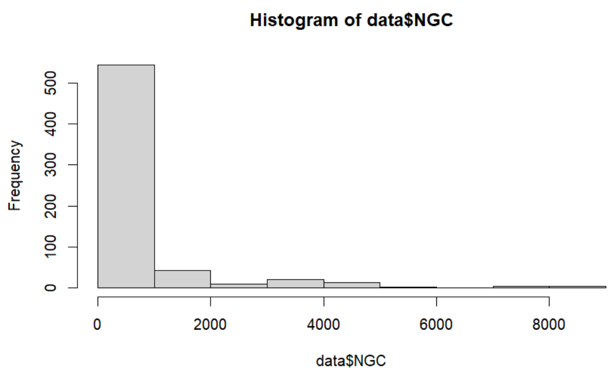

Figure 4.1 Histogram of Natural Gas Consumption (NGC)

The Figure 4.1 shows that there is a highly right skewed distribution. The reason behind this distribution is that especially in big cities of Turkey, consumption values are quite higher than other cities. Therefore, Box-Cox transformation is applied to obtain more normally distributed response variables. When deciding on Box-Cox method among transformation methods, the comparison in Table 4.2 is taken as basis. When the Pearson P test statistics are compared, it is seen that Box-Cox transformation gives the best result among the transformations in order to achieve more normally distributed data.

Table 4.2 Pearson P Test Statistics For Normality

Transformation Methods	**Pearson P / df**
double reversed log_b(x+a)	149.625
center+scale	115.481
sqrt(x+a)	21.557
log_b(x+a)	4.401
Yeo-Johnson	3.799
arcsinh(x)	3.693
Box-Cox	**3.620**

When Box-Cox transformation is applied, the λ is chosen 0.06 which is close to log transformation. The value of λ is determined as 0.06 by trying to optimize the log likelihood value. Due to the right skewed distribution in response variable, it is an expected result that optimal λ value is close to logarithmic transformation.

After applying Box-Cox transformation, the results are checked with the new histogram and q-q plot. As can be seen in the figures, although a completely normal distribution is not achieved, a more appropriate result is obtained. The reason for departure from the straight line in the tails is the low and high values that may be considered as outliers.

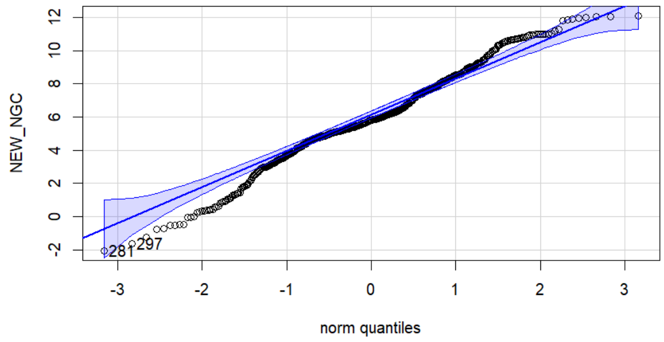

Figure 4.2 QQ Plot of Natural Gas Consumption (NGC) After Transformation

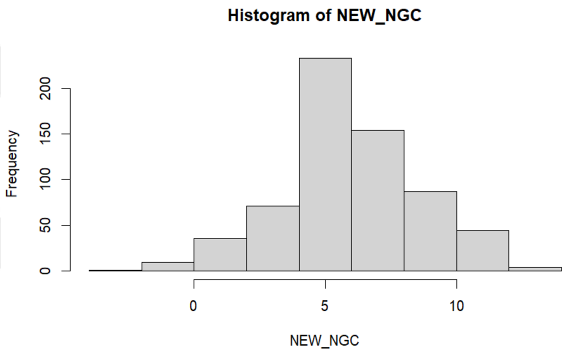

Figure 4.3 Histogram of Natural Gas Consumption (NGC) After Transformation

The Shapiro Wilk normality test results in Table 4.3 also indicate that after the transformation, completely normally distributed data is not obtained. However, the change in the p-value shows that more normally distributed data is obtained compared to the original data.

Table 4.3 Shapiro-Wilk Normality Test

Shapiro-Wilk Test	**W**	**p-value**
original data	0.505	$2.2*10^{-16}$
transformed data	0.986	$1.1*10^{-5}$

4.2.3 Correlation Structure of Response

To understand the correlation structure of response variable between years, the correlation matrix in Figure 4.4 can be used.

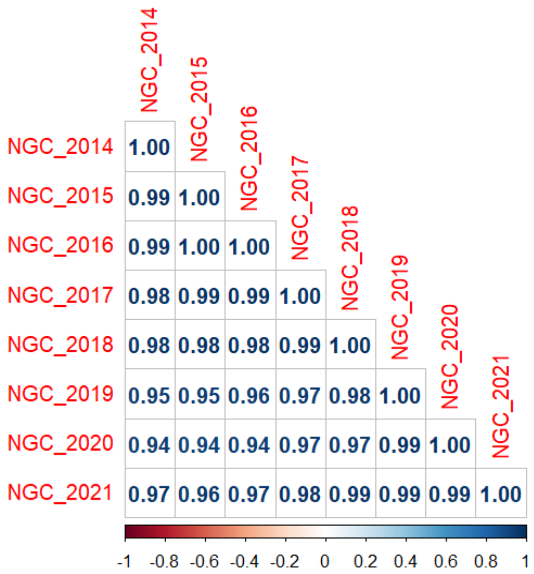

Figure 4.4 Correlation Matrix of Response (NGC) Between Years

The correlation matrix (Figure 4.4) clearly presents that there is a strong correlation in natural gas consumption between years which is above 0.90 for each year. It can be concluded from this table that natural gas is not an energy source that its consumption fluctuates and changes rapidly over the years. Same pattern can be captured from the Draftsman display (Figure 4.5) which is a helpful illustration, especially in balanced data, in order to understand the correlation structure (Dawson et al., 1997)

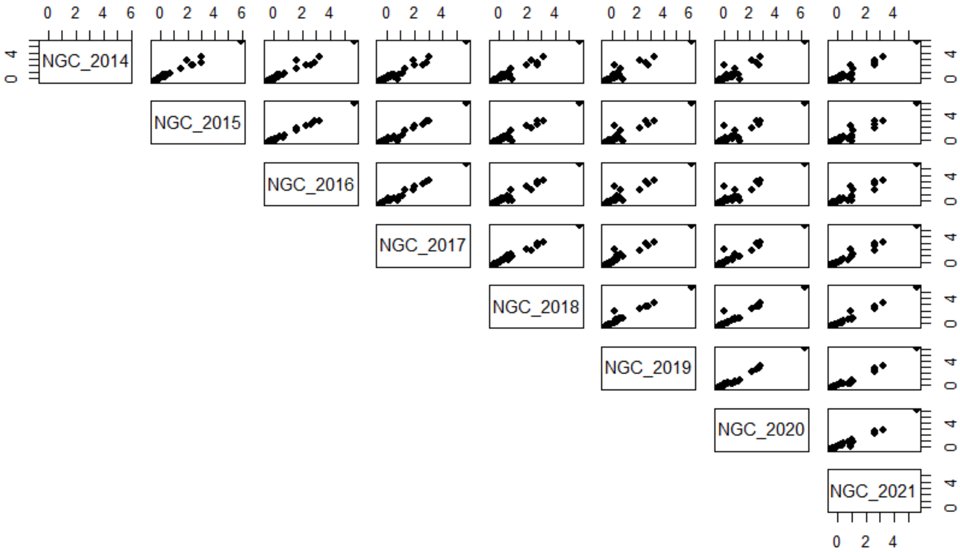

Figure 4.5 Draftsman Display of Response (NGC)

The strong correlation can be visually captured from the Draftsman display in Figure 4.5. Especially for one year ago correlations, which are indicated on the longest diagonal, the highest correlations are observed as the most linear lines. Strong one year ago correlations show that there is a strong relationship between the previous year's natural gas consumption and the current year's consumption.

4.2.4 Correlation between Response and Covariates

After examining the correlation between the response variable itself, it makes sense to examine the correlations between the response and numerical covariates. This makes it possible to understand whether there is a sensible association between the variables included in the study and the response. In this context, scatterplots (Figure 4.6, Figure 4.7, and Figure 4.8) are displayed over the years.

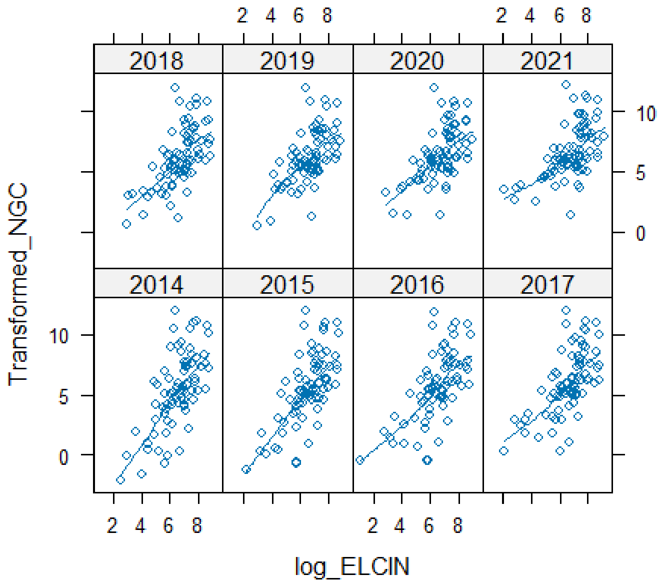

Figure 4.6 Scatter Plot of Transformed NGC and log Values of Industrial Electricity Consumption

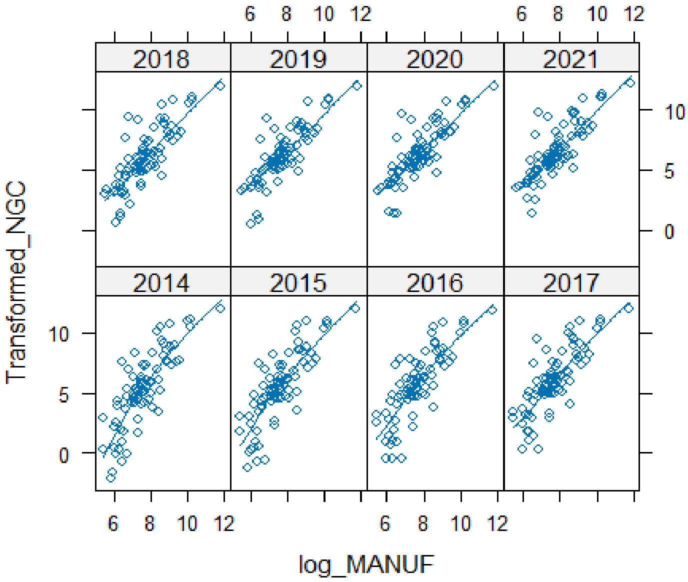

Figure 4.7 Scatter Plot of Transformed NGC and log Values of Number of Manufacturing Enterprises

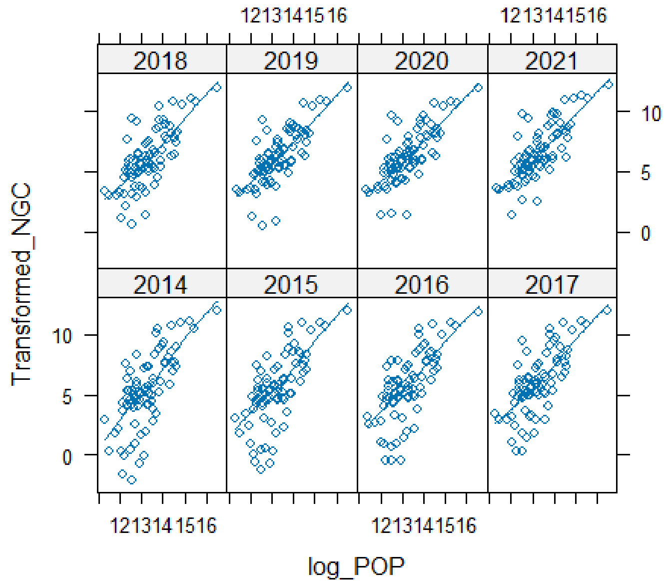

Figure 4.8 Scatter Plot of Transformed NGC and log Values of Population

In order to get linear relationships, log values are used in the plots and modelling part. From all the plots, it can be seen that there is a positive linear relationship between covariates and response in all years included in the study. For instance, an increase in the number of manufactural enterprises in a city can lead to an increase in the demand for natural gas energy in that city. Similarly, as the number of people living in a city increases, it would be expected that natural gas consumption increases, especially for heating purposes. As explained previously, industrial electricity consumption, number of manufactural enterprises, and population are related to main usages of natural gas. Therefore, it can be stated that these variables included in the study are related to natural gas consumption.

4.3 Panel Data Models

In the modelling part, three different panel data models are created respectively: fixed effects (marginal) model, transition (dynamic) model, and random effects model.

4.3.1 Empirical Results and Comparison of Modelling Part

Distinguishing features of each model are highlighted and overall model performances are compared in order to obtain the best estimation of natural gas consumption with specific variables which are industrial electricity consumption, number of manufacturing enterprises, population, natural gas unit price for industry, and years. The main objective is to find the optimum model in estimating natural gas consumption with certain variables, as well as to discuss the findings that arise in the modeling processes.

4.3.2 Fixed Effects (Marginal) Model

The first model in the modelling part is fixed effects model with following structure:

$$fitted\_NGC^{tr}{}_{it} = \beta_0 + \beta_1 YEAR_{it} + \beta_2 \log{(MANUF\_PP)_{it}} + \beta_3 \log{(POP)_{it}} + \beta_4 \log{(ELCIND)_{it}} + \beta_5 \mathrm{IND\_PRICE}_{it} + u_{it}$$

where:

fitted_NGC^{tr} is box-cox transformed values of natural gas consumption with a lambda value of 0.06. In addition, *MANUF_PP*, *POP*, and *ELCIND* variables are added to the model in log scale,

$i = 1,2, \ldots, 80$,

$u_{it} \sim N(0, \sum)$, and

t = 8 time points from 2014 to 2021.

It is decided which correlation structure to be chosen when establishing the marginal model by evaluating the correlation relationship of the response over the years. As indicated in the exploratory data analysis part, there is a similar correlation pattern over years. Therefore, correlation structure is determined as “exchangeable” in the marginal model.

As explained previously, the main requirements of natural gas consumption can be classified into three categories: household use, industrial use and use for electricity generation. In parallel with this classification, population variables related to industrial production and electricity use are significant variables in the fixed effect model as shown below (Table 4.4).

Table 4.4 Results of The Fixed Effects Model

Coefficients	Estimate	Roboust S.E.	Roboust Z value
intercept	1.662	3.939	0.411
YEAR	0.125	0.030	4.119
log(MANUF_PP)	2.633	0.424	6.209
log(POP)	1.303	0.157	8.257
IND_PRICE	-0.214	0.067	-3.182
log(ELCIND)	0.220	0.125	1.749

Estimated Scale Parameter: 1.995
Number of iterations: 7

According to Table 4.4, the fixed effects model equation can be written as follows;

$$fitted\_NGC^{tr}{}_{it} = 1.622 + 0.125YEAR_{it} + 2.633\log{(MANUF\_PP)_{it}} + 1.303\log{(POP)_{it}} + 0.220\log{(ELCIND)_{it}} - 0.214\mathrm{IND\_PRICE}_{it}$$

It should be noted that Box-Cox transformation is applied to NGC variable with λ which is equal to 0.06. Therefore, Box-Cox transformation in response variable and log transformations in exploratory variables need to be considered in interpretation of each model. Regarding the factors and their relative effects on natural gas consumption, following remarks can be stated;

A year increase causes a $0.125^{(16.6)}$ increase on NGC when all other variables are held constant.

A unit increase in log(MANUF_PP) variable causes a $2.633^{(16.6)}$ increase on NGC when all other variables are held constant.

A unit increase in log(POP) variable causes a $1.303^{(16.6)}$ increase on NGC when all other variables are held constant.

A unit increase in IND_PRICE variable causes a $0.214^{(16.6)}$ decrease on NGC when all other variables are held constant. The negative relationship between IND_PRICE and NGC is logical because increases in industrial natural gas prices negatively affect natural gas consumption associated with industrial production.

A unit increase in log(ELCIND) variable causes a $0.220^{(16.6)}$ increase on NGC when all other variables are held constant. The increase in industrial electricity consumption can be related to the increase in industrial activity. At the same time, the requirement for electricity production is also increasing. Therefore, it is a logical comment that there is an enhancing effect on natural gas consumption.

In addition to the effects of variables on natural gas consumption, intuitive associations can be considered between explanatory variables, however there is no multicollinearity in the model. Multicollinearity is checked by examining the variance inflation factor (VIF). According to Table 4.5, it can be concluded that the model does not have multicollinearity because the VIF values are less than 10 which is considered as the upper limit (O’Brien, 2007).

Table 4.5 Variance Inflation Factor (VIF) Values of The Fixed Effects Model

YEAR	log(MANUF_PP)	log(POP)	IND_PRICE	log(ELC_IND)
5.271	1.793	1.149	5.270	1.628

When a fixed model is used, all subjects are evaluated collectively. It is possible to determine whether particular groups vary or do not by using dummy variables. First, “Marmara Region” is created as a dummy variable to observe regional impact. The Marmara region is anticipated to have more intense distribution in terms of population and industrial activity compared to other regions. In this direction, the dummy variable that distinguishes the provinces in Marmara region from others is added to the fixed effects model. Following equation is obtained according to Table 4.6;

$$fitted\_NGC^{tr}{}_{it} = 0.663 + 0.129YEAR + 2.384\log{(MANUF\_PP)_{it}} + 1.263\log{(POP)_{it}} + 0.200\log{(ELCIND)_{it}} - 0.203\mathrm{IND_{PRICE}}_{it} + 1.456\mathrm{MAR}$$

(MAR: 1 = Marmara region provinces, 0 = others)

Table 4.6 Results of The Fixed Effects Model with Dummy Variable for Marmara Region

Coefficients	Estimate	Roboust S.E.	Roboust Z value
intercept	0.663	3.738	0.177
YEAR	0.129	0.030	4.299
log(MANUF_PP)	2.384	0.399	5.975
log(POP)	1.263	0.161	7.840
IND_PRICE	-0.203	0.067	-3.028
log(ELCIND)	0.200	0.121	1.640
MAR	1.456	0.422	3.448

Estimated Scale Parameter: 1.737
Number of iterations: 7

Although the model coefficients are close to the first fixed effects model, it is noteworthy that a positive coefficient is added for the Marmara region. In other words, when a province in the Marmara region is investigated, a positive effect of +1.456 is added to the natural gas consumption. It is quite normal for the Marmara region to be separated by an additional positive coefficient (+1.456) because natural gas consumption is expected to be higher than other regions in terms of both industrial activity and population impact. Revealing this difference in an analysis for natural gas consumption may be valuable in evaluating national consumption in general.

Similarly, dummy variables for big cities are created and added to the fixed effects model. First, a dummy variable is added for Istanbul, Bursa, Kocaeli, Ankara and Izmir, but since Istanbul and Bursa are not significant, these cities are removed from the model. With significant big cities, following equation is obtained based on Table 4.7;

$$fitted\_NGC^{tr}{}_{it} = 2.495 + 0.127YEAR_{it} + 2.577\log{(MANUF\_PP)_{it}} + 1.208\log{(POP)_{it}} + 0.221\log{(ELCIND)_{it}} - 0.213\mathrm{IND\_PRICE}_{it} + 2.308\mathrm{KOC} + 1.019\mathrm{IZM} + 1.430\mathrm{ANK}$$

(KOC: 1 = Kocaeli, IZM: 1 = İzmir, ANK: 1 = Ankara, 0 = others)

Table 4.7 Results of The Fixed Effects Model with Dummy Variable for Big Cities (Kocaeli, İzmir, Ankara)

Coefficients	Estimate	Roboust S.E.	Roboust Z value
intercept	2.495	3.971	0.628
YEAR	0.127	0.030	4.186
log(MANUF_PP)	2.577	0.419	6.139
log(POP)	1.208	0.165	7.315
IND_PRICE	-0.213	0.067	-3.155
log(ELCIND)	0.221	0.124	1.775
DUM_KOC	2.308	0.295	7.815
DUM_IZM	1.019	0.321	3.168
DUM_ANK	1.430	0.342	4.175

Estimated Scale Parameter: 1.918
Number of iterations: 7

As seen from equation, additional positive coefficients are added to the model that indicates significantly higher consumption for big cities. Like the Marmara region, big cities have higher industrial activity and population compared to other cities which lead to higher natural gas consumption.

4.3.3 Random Effects Model

In the fixed effects model, province specific impacts could not be captured in previous sections due to the nature of the model. In random effects models, it is possible to understand subject specific effects and changes. In this study, it enables analysts to observe the different characteristics of each province.

As the second model in the modelling part, two random effects models can be built with following structures:

- Random intercept only:

$$fitted\_NGC^{tr}{}_{it} = \beta_0 + \beta_1 YEAR_{it} + \beta_2 \log{(MANUF\_PP)_{it}} + \beta_3 \log{(POP)_{it}} + \beta_4 \log{(ELCIND)_{it}} + \beta_5 \text{IND\_PRICE}_{it} \ + \ \alpha_{0i} + u_{it}$$

where:

fitted_NGC^tr is box-cox transformed values of natural gas consumption with a lambda value of 0.06. In addition, *MANUF_PP*, *POP*, and *ELCIND* variables are added to the model in log transformed.

i = 1,2, …, 80,

$u_{it} \sim N(0, \sum)$ in which $\sum$ is an identity matrix.

t = 8 time points from 2014 to 2021 , and

α_0 = random intercept term

- Random intercept and slope:

$$fitted\_NGC^{tr}{}_{it} = \beta_0 + \beta_1 YEAR_{it} + \beta_2 \log{(MANUF\_PP)_{it}} + \beta_3 \log{(POP)_{it}} + \beta_4 \log{(ELCIND)_{it}} + \beta_5 \text{IND\_PRICE}_{it} \ + \ \alpha_{0i} + \ \alpha_{1i} YEAR_{it} + u_{it}$$

where:

fitted_NGC^{tr} is box-cox transformed values of natural gas consumption with a lambda value of 0.06. In addition, *MANUF_PP*, *POP*, and *ELCIND* variables are added to the model in log transformed.

$i = 1,2, \ldots, 80$,

$u_{it} \sim N(0, \sum)$ in which $\sum$ is an identity matrix.

t = 8 time points from 2014 to 2021 , and

$\alpha_{0,1}$ = random intercept and random slope terms.

Two random effects models mentioned above are built to model natural gas consumption. By comparing AIC and deviance values as evaluation criteria, it is determined which of these models give better results. The random effects model that combined the random intercept and slope together result with better outcomes as seen in Table 4.8. In other words, a distinct intercept and slope impact are observed for each province in Turkey.

Table 4.8 Comparison Between Random Effects Models

Comparison between random effects models	**AIC**	**Deviance**
with random intercept only	1385.84	1369.84
with random intercept and slope	**913.82**	**893.82**

Therefore, random intercept and slope will be used together for the random effects model. Similar to other models, the random effect model output with the same variables as in other models can be seen in Table 4.9. Intercept and slope values for the first 10 provinces according to ID is given in Table 4.10.

Table 4.9 Results of The Random Effects Model

Coefficients	Estimate	Std. Error	t value
intercept	-8.057	2.836	-2.841
YEAR	0.149	0.025	5.845
log(MANUF_PP)	0.995	0.280	3.543
log(POP)	1.391	0.147	9.466
IND_PRICE	0.136	0.050	-2.705
log(ELCIND)	0.091	0.053	1.709

Scaled Residuals:				
Min	1Q	Median	3Q	Max
-4.717	-0.282	-0.015	0.270	4.440

Random Effects:	Name	Variance	Std. Dev.	Corr
Groups	(intercept)	4.488	2.118	
ID	YEAR	0.039	0.199	-0.83
Residual		0.089	0.298	

Table 4.10 Intercept and Slope Values for The First 10 Provinces

ID	**(intercept)**	**YEAR**
1	-0.329	-0.056
2	-1.123	0.0005
3	-0.892	0.006
4	-2.528	0.244
5	0.310	-0.045
6	1.936	-0.123
7	1.174	-0.188
8	-2.978	-0.018
9	-1.299	-0.012
10	2.116	-0.0149

According to Table 4.9 and Table 4.10, the random effects model equation can be written differently for each province. The equation is presented below for Ankara (ID=6) as an example, with the random effects terms marked as bold. Each province in the model has different random effects terms that represent the province specific natural gas consumption patterns.

$$fitted\_NGC^{tr}{}_{it} = (-8.057 + \mathbf{1.936}) + (0.149 - \mathbf{0.123})YEAR_{it} + 0.995\log(MANUF\_PP)_{it} + 1.391\log(POP)_{it} + 0.091\log(ELCIND)_{it} - 0.136\text{IND\_PRICE}_{it}$$

Based on the random effects model for Ankara, following province specific remarks can be stated;

A year increase causes 0.026(16.6) increase on NGC when all other variables are held constant.

A unit increase in log(MANUF/PP) variable causes a 0.995(16.6) increase on NGC when all other variables are held constant.

A unit increase in log(POP) variable causes a 1.391(16.6) increase on NGC when all other variables are held constant.

A unit increase in IND_PRICE variable causes a 0.136(16.6) decrease on NGC when all other variables are held constant.

A unit increase in log(ELC_IND) variable causes a 0.091(16.6) increase on NGC when all other variables are held constant.

4.3.4 Transition (Dynamic) Model

The third model in the modelling part is the transition model with following structure:

$$fitted\_NGC^{tr}{}_{it} = \beta_0 + \alpha_1 NGClag_1 + \beta_1 YEAR_{it} + \beta_2 \log(MANUF\_PP)_{it} + \beta_3 \log(POP)_{it} + \beta_4 \log(ELCIND)_{it} + \beta_5 \text{IND\_PRICE}_{it} + u_{it}$$

where:

fitted_NGC^{tr} is box-cox transformed values of natural gas consumption with a lambda value of 0.06. In addition, *MANUF_PP*, *POP*, and *ELCIND* variables are added to the model in log transformed.

$i = 1,2, \ldots, 80,$

$u_{it} \sim N(0, \Sigma)$ in which Σ is an identity matrix and,

t = 8 time points from 2014 to 2021

In the dynamic model, as can be seen from the equation, natural gas consumption one year ago is added to the model as an explanatory variable. It is decided which lag values should be added to the model by comparing the standard errors for different lag values. As well as sum of square error, absolute sum of square error is also compared in case of a possible outlier effect. As can be seen from Table 4.11, the lowest standard errors are achieved with lag 1 values. In other words, there is no need to add natural gas consumption 2 or 3 years ago as an explanatory variable to the model. In addition, considering the correlation structure of the response, which is emphasized in the exploratory data analysis section, that there is a high correlation between consumption values of current year and 1 year ago. This high correlation provides sufficient explanation, thus adding further lag values are not necessary.

Table 4.11 Comparison between Transition Models with Different Lag Values

Lag term	SSE	abs_SSE
LAG_3	1.469	0.891
LAG_2	1.352	0.842
LAG_1	**1.169**	**0.767**

Along with 1 year lag values, the variables in transition model are similar to fixed effects model as seen in Table 4.12.

Table 4.12 Results of The Transition Model

Coefficients	Estimate	Roboust S.E.	Roboust Z value
intercept	-10.215	3.089	-3.306
NGC_LAG_1	0.336	0.051	6.592
YEAR	-0.297	0.054	-5.471
log(MANUF_PP)	0.637	0.266	2.394
log(POP)	1.081	0.142	7.601
IND_PRICE	0.975	0.149	6.529
log(ELCIND)	0.542	0.112	4.813

Estimated Scale Parameter: 1.169
Number of iterations: 1

According to Table 4.12, the transition model equation can be written as follows;

$$fitted\_NGC^{tr}{}_{it} = -10.215 + 0.336 NGC\_LAG\_1 + -0.297 YEAR_{it} + 0.637\log(MANUF\_PP)_{it} + 1.081\log(POP)_{it} + 0.542\log(ELCIND)_{it} + 0.975\text{IND\_PRICE}_{it}$$

Based on transition model, following remarks can be stated;

A year increase causes -0.297[(16.6)] decrease on NGC when all other variables are held constant. When interpreting the year variable, it should be taken into account that the previous year's natural gas consumption will also change.

One unit increase in natural gas consumption a year ago results in 0.336[(16.6)] increase in current year natural gas consumption when all other variables are held constant. Higher natural gas consumption in the previous year is a factor that increases the consumption in the current year.

A unit increase in log(MANUF/PP) variable causes a 0.637[(16.6)] increase on NGC when all other variables are held constant.

A unit increase in log(POP) variable causes a 1.081[(16.6)] increase on NGC when all other variables are held constant.

A unit increase in IND_PRICE variable causes a $0.975^{(16.6)}$ increase on NGC when all other variables are held constant.

A unit increase in log(ELC_IND) variable causes a $0.542^{(16.6)}$ increase on NGC when all other variables are held constant.

Similar to the fixed effects model, multicollinearity is checked when evaluating the model coefficients. The results in Table 4.13 indicate that there is no multicollinearity between variables since VIF values are below the upper limit 10.

Table 4.13 Variance Inflation Factor (VIF) Values of The Transition Model

YEAR	NGC_LAG_1	log(MANUF_PP)	log(POP)	IND_PRICE	log(ELC_IND)
1.824	7.827	1.646	2.926	6.324	1.905

4.3.5 Model Comparison Results

In the modelling part, fixed effects (marginal) model, random effects model, and transition (dynamic) model are implemented and interpreted according to their own features. Within the section dedicated to model comparison, the models are compared based on their performance on modelling natural gas consumption. As criteria for comparison, both visual and numerical factors are assessed. At first, the fitted values and the actual values are compared visually. Then, root mean square error (RMSE), mean absolute percentage error (MAPE), and correlation between fitted and actual values are calculated in order to compare the performances of the models.

In visual comparison of model performance, fitted values and actual values can be seen separately for each model. The random effects model exhibits performance that is most similar to the actual values, as can be captured from Figure 4.10. Generally, there is not a remarkable difference between fitted and actual values, despite the fact that the deviation is slightly higher in the values below the average. After random effects model, the dynamic model appears to have superior performance over the fixed effects model according to Figure 4.9 and Figure 4.11.

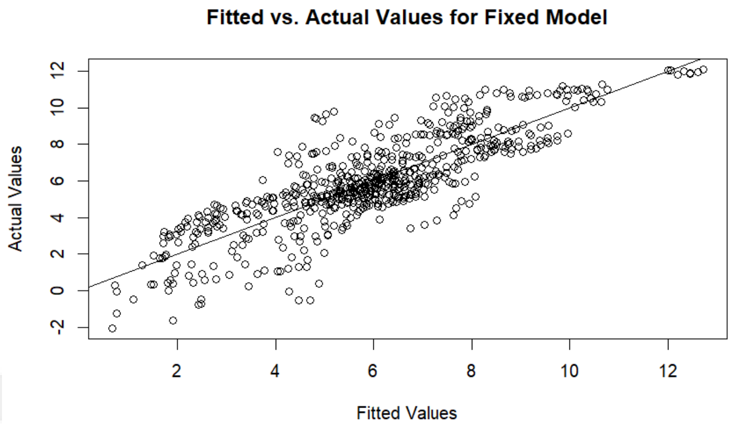

Figure 4.9 Fitted vs Actual Values for Fixed Effects Model

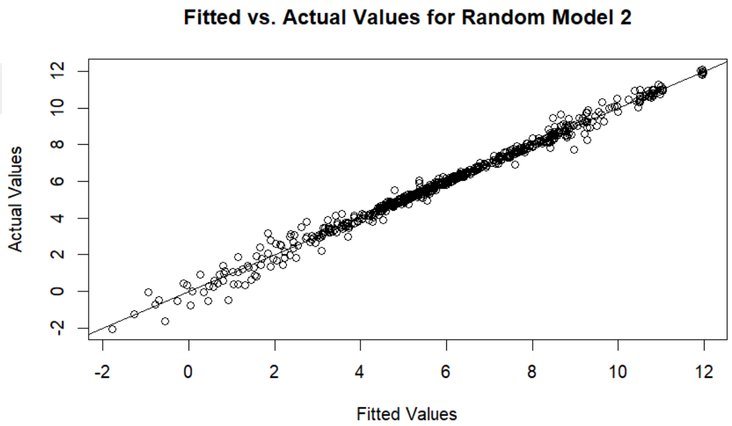

Figure 4.10 Fitted vs Actual Values for Random Effects Model

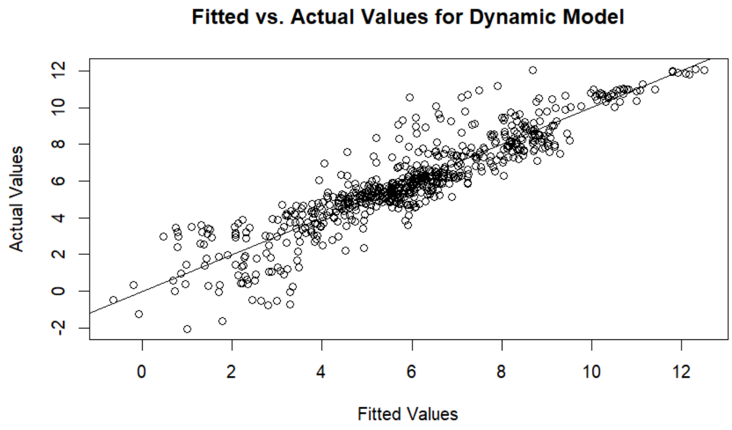

Figure 4.11 Fitted vs Actual Values for Transition (Dynamic) Model

In addition to the comments made from the scatter plots, numerical values should also be compared between the models. As stated, RMSE, MAPE and correlation metrics are calculated.

The accuracy performance outcomes in Table 4.14 validate the visual results, demonstrating that the random effects model performs better based on all criteria. Notable aspects of the random effects model can be highlighted as low error rates and very strong correlation between fitted and actual natural gas consumption values. The transition model, that is the second-best model in terms of performance, has 90% correlation which can be categorized as strong. The performance of the fixed effects model is relatively low compared to the other models.

Table 4.14 Comparison Between Models

Model \ Criteria	RMSE	MAPE	Correlation (r)
Fixed Effects	1.406	64.308	0.836
Random Effects	**0.260**	**10.661**	**0.995**
Transition Model	1.075	46.606	0.907

To summarize the comparison part, according to all metrics, random effects model with random intercept and slope outperformed. This result can be also captured from fitted and actual values plots visually. The primary result of this outcome may be the distinctive characteristics that Turkey's provinces have. Individual effects, or province specific effects, are therefore quite significant and need to be taken into account. Particularly, when industrialized cities and more crowded cities consume substantially more natural gas consumption than other provinces. In other words, the consumption patterns vary among the provinces. Therefore, the analyst is able to capture this variation using a random effects model with random intercept and slope. This also explains why the random effects model performs better when both random slope and random intercept term are added to the model. Furthermore, random effects model has an additional advantage which is its interpretation flexibility. In the fixed effects model and the transition model, a pooled equation covering all provinces is used for interpretation. However, with different intercept and slope parameters for each province, more exact interpretation for a specific year and more accurate observation of the trend over time can be captured by using random effects model. Due to its superior performance and additional benefits, the random effects model with random intercept and slope is the best model in terms of modelling natural gas consumption in Turkey by using panel data.

4.4 Missing Data Simulation Part

After the modelling part, the possible effects of missing data in different missingness mechanisms and the performance of imputation techniques are examined in this simulation part. As previously mentioned, there is no missing data in the original data set. Therefore, a simulation study is applied where the missing data is repeatedly generated using various assumptions and imputed by applying different methods. The following sections provide a step-by-step explanation of the simulation design and its details. At the end, the comparison of imputation techniques in different missingness scenarios will be discussed.

4.4.1 Simulation Design

The simulation design can be summarized step by step in Figure 4.13. The main purpose of the missing data simulation study is to evaluate and discuss the performance of various imputation techniques by generating missing data based on different assumptions.

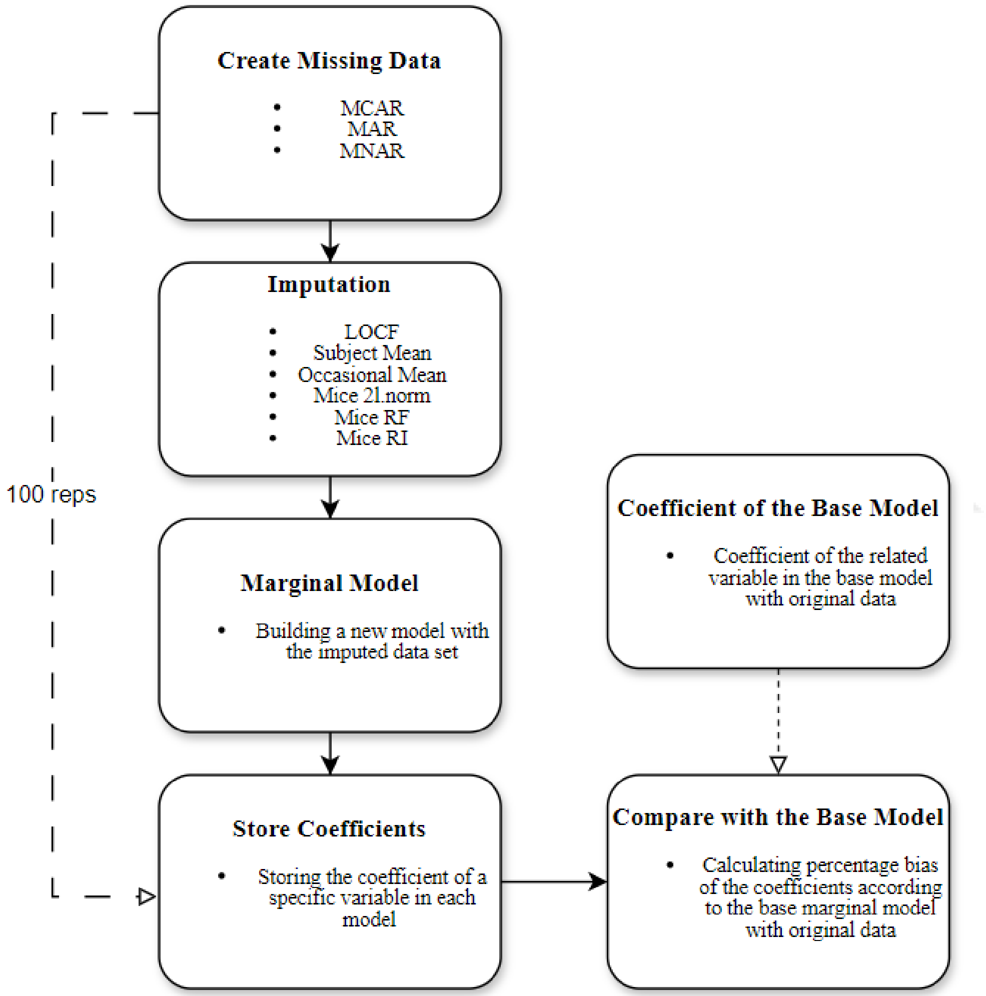

Figure 4.12 Missing Data Simulation Design

4.4.2 Missing Data Generation

Missing data in the simulation study is created under three different assumptions: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). The main purpose of analyzing these three different structures is to evaluate all missing data mechanisms. As the name implies, in MCAR case, the natural gas consumption values are randomly deleted to generate missing data in the response variable. Missing observations for MAR and MNAR cases need to be created by using a related variable in the data set. Therefore, the population variable is selected, and the missing values in the response variable are repeatedly deleted in connection with the population variable. However, it should be noted that MNAR case could not be regenerated in every repetition due to the nature structure of the algorithm. In MNAR mechanism, the same observations are eliminated from the data set after each iteration because the missing data is produced using a specific logic. Instead of creating different missing value sets depending on a specific variable as in the MAR assumption, a one-shot deletion is performed in the MNAR assumption. Since this logic does not change in each repetition, the deleted observations could not differ, and only one data set with MNAR mechanism could be achieved. To ensure consistency in the analyses, this single data set is used for MNAR case.

4.4.3 Missing Data Imputation

After generating missing data under three different mechanisms, following imputation methods are applied respectively: last observation carried forward (LOCF), subject mean imputation, occasional mean imputation, mice 2l.norm method, mice random forest method, and mice random indicator method. Imputation techniques in this study can be classified into two categories: single imputation methods, which are LOCF and mean imputations, and multiple imputation methods that consist of mice (multivariate imputation by chained equations) methods.

4.4.4 Model Building with Imputed Data

Figure 4.13 explains how the missing data is initially generated and then imputed using various methods. Following these procedures, new marginal models are built using the obtained imputed data. For each missing data mechanism and imputation technique, this process is carried out 100 times. In order to compare new models with the marginal model created using the original data, the coefficient of the population variable in the model is stored during these repetitions.

4.4.5 Comparison of Imputation Techniques under Different Missingness Mechanisms

The coefficient of the population variable is chosen as the control variable and comparisons are made by using this variable. The coefficient of the population variable in the marginal model established with the original data is the base point which is 1.303 in this study. The percentage bias is used as evaluation criteria (Burton et al., 2006).

The stored coefficients, the base model coefficient, and the percentage bias for each missingness case are given in the comparison tables (Table 4.13, Table 4.14 and Table 4.15) in the next section. The closer the population variable coefficient in the models built with imputed data to the base point (1.303), the more successful imputation it is. The tables also include computation times of 100 repetitions which might be considered as an additional characteristic of imputation techniques.

4.4.5.1 Performance of Single Imputation Techniques

In the comparison section, single imputation techniques are first discussed. The techniques covered here are subject mean imputation, occasional mean imputation, and last observation carried forward method. The comparison tables of the simulation study with single imputation techniques are given in Table 4.15, Table 4.16 and Table 4.17.

Table 4.15 Subject Mean Imputation Results

time	**Subject Mean**	coefficient imputed model	coefficient base model	**error**
6,242 (6 sec)	MCAR	1.391	1.303	7%
rep	MAR	1.367	1.303	5%
100	MNAR	1.223	1.303	-6%

Table 4.16 Occasional Mean Imputation Results

time	**Occasional Mean**	coefficient imputed model	coefficient base model	**error**
60,412 (1 min)	MCAR	1.056	1.303	-19%
rep	MAR	0.753	1.303	-42%
100	MNAR	0.820	1.303	-37%

Table 4.17 LOCF Imputation Results

time	**LOCF**	coefficient imputed model	coefficient base model	**error**
4,466 (4 sec)	MCAR	1.458	1.303	12%
rep	MAR	1.421	1.303	9%
100	MNAR	0.632	1.303	-51%

Among single imputation methods, subject mean imputation gives better results than other methods. Error rates of 5-7% are obtained with different missingness mechanisms as seen from Table 4.15. It is emphasized in the previous sections that the characteristics of the provinces in the study are very different from each other. Because it is less affected by differentiation between provinces, smaller error rates result from subject mean imputation compared to other single imputation methods regardless of the missing data mechanism. In other words, average natural gas consumption value within a province might be reasonable for imputation in this study. On the other hand, in another type of mean imputation method, which is occasional mean imputation, it is observed that the error rates are significantly higher as indicated in Table 4.16. Especially for MAR and MNAR cases, error rates are 42% and 37% respectively which are quite misleading. Even under MCAR case, there is a 19% deviation from the original model. According to these results, it can be concluded that occasional mean imputation yields a misleading imputation in each case because it is more influenced by subject-specific differences, particularly if the missing data is not random. Furthermore, subject mean imputation outperforms occasional mean imputation in terms of computational time when two mean imputation techniques are compared. Finally in single imputation techniques, for LOCF, the type of missing data mechanism has a crucial impact. In LOCF imputation, the error rates for MCAR and MAR are 12% and 9% respectively (Table 4.17). However, in the case where missing data is not at random, the LOCF method results in the highest bias among single imputation techniques that is 51%. Consequently, in MNAR case, the LOCF approach is more misleading despite being simple to use and computationally efficient.

4.4.5.2 Performance of Multiple Imputation Techniques

Secondly, in the comparison section, the results of multiple imputation methods are given. The three methods included in this study are based on the Multivariate Imputation by Chained Equations (Mice) package in R: Mice Random Indicator, Mice 2l.norm, and Mice rf. The comparison tables of the simulation study where multiple imputation techniques are applied are given in Table 4.18, Table 4.19 and Table 4.20.

Table 4.18 Mice RI Imputation Results

time	Mice ri	coefficient imputed model	coefficient base model	error
199,782 (3 min)	MCAR	1.357	1.303	4%
rep	MAR	1.304	1.303	0.1%
100	MNAR	1.341	1.303	3%

Table 4.19 Mice 2l.norm Imputation Results

time	Mice 2l.norm	coefficient imputed model	coefficient base model	error
2406,633 (40 min)	MCAR	1.394	1.303	7%
rep	MAR	1.414	1.303	9%
100	MNAR	1.141	1.303	-12%

Table 4.20 Mice RF Imputation Results

time	Mice rf	coefficient imputed model	coefficient base model	error
142,631 (2 min)	MCAR	1.449	1.303	11%
rep	MAR	1.372	1.303	5%
100	MNAR	1.080	1.303	-17%

Regarding multiple imputation techniques, it can be generally stated that more consistent and accurate results are obtained with longer computational times compared to single imputation techniques. When multiple imputation techniques are compared to each other, Mice ri outperforms with its lower bias. Notably, under each missingness mechanism, mice ri method gives successful results that are very close to original data as seen from Table 4.18. Both random and not at random cases can be satisfactorily handled with this method. Mice 2l.norm approach is another multiple imputation method whose computational time is significantly longer. The performance is examined in terms of missing data mechanisms in Table 4.19, and it is found that the MNAR case has a slightly higher bias (12%) than the random cases (7% and 9%). Finally, the results of Mice rf method is given in Table 4.20. For the MNAR case, the percentage bias is at 17%, and according to this result, it shows lower performance than other multiple imputation methods. While the error rate is 5% in the MAR case, the error rate for the MCAR case is 11%.

CHAPTER 5

CONCLUSION

Energy consumption modelling is a critical and extensively researched topic in the literature. Understanding the national energy requirements and related variables are crucial for policy makers, particularly from a macroeconomic perspective. In light of evolving energy consumption patterns, fundamental energy sources have become more important for nations that rely heavily on energy. This thesis study focuses on panel data modelling to examine the natural gas consumption patterns of provinces in Turkey. The first part of the study examines which panel data models are more effective at modelling natural gas consumption. The second part includes a missing data simulation with real-life data. The missing data simulation part aims to investigate how various imputation methods result in various missing data mechanisms. Numerous studies that use real-life data and model energy consumption in literature concentrate on time series mainly. Using the benefits and advantages of panel data, this thesis study focuses on modelling natural gas consumption in Turkey at province level. Moreover, three different missingness assumptions (MCAR, MAR, MNAR) are taken into consideration in the simulation part. In this simulation, the performance of single and multiple imputation methods is compared under each assumption. Covering each missing data scenario and working on different imputation techniques by using panel data in a comparative approach are valuable. In this context, it is aimed to present a different perspective as a pioneering study.

In the modeling phase, which is the first part of the study, natural gas consumption in Turkey for the years 2014-2021 is modeled using fixed effects (marginal model), random effects and transition (dynamic) model. While discussing the different features of the models, model performances are evaluated according to RMSE, MAPE and correlation metrics. As a result of these analyses, it is concluded that the random effects model with random intercept and slope give the best results. The reason for this may be that subject-specific effects are captured better in the random effects model. In addition to capturing subject-specific effects, the high interpretation power of the random effect model is seen as a distinct advantage in order to accurately interpret the behavior of different provinces.

In the missing data simulation study, three different missing data mechanisms are considered: MCAR, MAR, and MNAR assumptions. Since there is no missing data in the original data set, artificial missing data is created for each mechanism. According to the simulation design, the repeatedly created missing data are imputed with six different methods which are subject mean imputation, occasional mean imputation, last observation carried forward, mice random indicator, mice 2l.norm, and mice random forest. The basic approach in the simulation is based on how much the coefficients in a marginal model built with original data and the marginal model built with imputed data differ from each other. In this context, the percentage bias in the "population" coefficient, which is the selected control variable, is chosen as the performance criterion. A lower bias indicates that the imputation converges to the original data, that means it is more appropriate. One of the important outcomes of the simulation study is that understanding the missing data mechanism is crucial when deciding which imputation method to be applied. The success rate obtained from imputation methods varies depending on the mechanism of missing data. As another result, it can be said that single imputation techniques are unreliable although acceptable error rates can be obtained in random missingness. Especially in non-random missingness, single imputation techniques might be misleading. Multiple imputation techniques provide lower bias even in not random missingness due to evaluating the relationships between missing variable and observed data. The technique that gives the best results in this study is the mice random indicator method. In summary of the simulation study, after understanding the cause and mechanism of missing data, choosing multiple imputation methods using the power of existing data gives more reliable results.

5.1 Limitations

Some of the limitations included in the modeling and missing data simulation part are as follows:

- Because of the nature of the dataset itself, the number of subjects is 80 (one province of Turkey is removed due to the lack of data). Since the study covers provinces in Turkey, it is not possible to increase the number of subjects.
- The time scope of the study can not be expanded, as 2014 is the first year in which province-based natural gas consumption data is released.
- In the modeling part, the data can not be split into train and test sets (80%-20%) for validation due to the small number of provinces and their diverse features within the dataset.
- While the missing data is generated in the simulation part, the MCAR and MAR cases yield different data sets in each iteration, whereas the MNAR mechanism is provided with a single data set. MNAR mechanisms that give different results in each application could be obtained with a larger data set and different missing data algorithms.

5.2 Further Research

Several aspects of this study can be expanded and developed, including the modeling phase and the application of missing data simulation part. For instance, the relationship between different energy sources, such as electricity or coal, and natural gas consumption might be investigated in the modeling phase. With this investigation, it may be possible to observe the patterns of different energy sources and their impacts on Turkey's aggregate natural gas consumption. Furthermore, for future research, a macroeconomic approach to this study is suggested by modeling at the country level. For example, it is possible to build different panel data models that consider macroeconomic variables and examine the OECD countries or the European Union. In such a study, while the accessibility of data may be simple by using national databases, it is important to note that a small number of observations may lead to estimation issues.

For future research, it is also suggested that different machine learning algorithms can be applied to model energy consumption. Even though the number of observations might be a drawback for machine learning applications, the impact of a small number of observations can be discussed, and the performance of models can be compared with marginal, dynamic, and transition models.

For the missing data simulation section, it is suggested to combine more sophisticated imputation methods, such as selection models or pattern mixture models, with the single and multiple imputation techniques. Therefore, a simulation with real-world data can be applied to assess how these techniques perform, especially for non-ignorable missing data.

REFERENCES

EPDK (2021), Energy Market Regulatory Authority, Turkish Natural Gas Market Report 2021, retrieved from: https://www.epdk.gov.tr/

TPAO (2022), Turkish Petroleum Corporation, Sakarya Gas Field Development Project Non-Technical Summary Report 2022, retrieved from: https://www.tpao.gov.tr/

GAZBİR (2021), The Association of Natural Gas Distributors of Turkey, Natural Gas Distribution Industry Report 2021, retrieved from: https://www.gazbir.org.tr/

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, retrieved from: https://www.R-project.org/

Toksari, M. (2012). Ant colony optimization based on multi-searching to estimate the natural gas demand: case of turkey. *Energy Exploration & Exploitation*, 30(2), 223-238.

Öztürk, S. and Ozturk, F. (2018). Forecasting energy consumption of turkey by arima model. *Journal of Asian Scientific Research*, 8(2), 52-60.

F. E. Boran (2015) Forecasting Natural Gas Consumption in Turkey Using Grey Prediction, *Energy Sources Part B: Economics, Planning, and Policy*, 10:2, 208 213.

Kayakuş, M. (2020). The estimation of turkey's energy demand through artificial neural networks and support vector regression methods. *Alphanumeric Journal*, 8(2), 227-236.

Ozdemir, G., Aydemir, E., Olgun, M. O., & Mulbay, Z. (2016). Forecasting of Turkey Natural Gas Demand Using a Hybrid Algorithm. *Energy Sources Part B: Economics, Planning, and Policy*, 11(4).

Karadede, Y., Ozdemir, G., & Aydemir, E. (2017). Breeder hybrid algorithm approach for natural gas demand forecasting model. *Energy*, 141, 1269–1284.

Sen, D., Günay, M. E., & Tunç, K. M. M. (2019). Forecasting annual natural gas consumption using socio-economic indicators for making future policies. *Energy*, 173, 1106–1118.

Özcan, M., Yildirim, M. B., & Öztürk, S. (2016). Generation expansion planning scenarios to reduce natural gas dependency of turkey. *Energy Exploration &Amp; Exploitation*, 34(2), 244-261.

Eygu, H., Soğukpınar, F. (2023). Investigation of the relationship between renewable energy, natural gas, and coal consumption with economic growth in Turkey: evidence from augmented ARDL approach. *Environmental Science and Pollution Research*, 30, 58213–58225.

Erdem, O.E., & Kesen, S.E. (2020). Estimation of Turkey's Natural Gas Consumption by Machine Learning Techniques. *Gazi University Journal of Science*, 33(1), 120-133.

Taşpınar, F. (2015). Time Series Models for Air Pollution Modelling Considering the Shift to Natural Gas in a Turkish City. *CLEAN - Soil, Air, Water*, 43(7), 980–988.

Taşpınar, F., Çelebi, N., & Tutkun, N. (2013). Forecasting of daily natural gas consumption on regional basis in Turkey using various computational methods. *Energy and Buildings*, 56, 23–31.

Demirel, Ö. F., Zaim, S., Çalişkan, A., & Özuyar, P. (2012). Forecasting natural gas consumption in Istanbul using neural networks and multivariate time series methods. *Turkish Journal of Electrical Engineering and Computer Sciences*, 20(5), 695-711.

Dombaycı, Ö. A. (2010). The prediction of heating energy consumption in a model house by using artificial neural networks in Denizli–Turkey. *Advances in Engineering Software*, 41(2), 141-147.

Baraldi, A. N., & Enders, C. K. (2017). An introduction to modern missing data analyses. *Journal of School Psychology*, 48, 5–37.

Roberts, M. B., Sullivan, M. C., & Winchester, S. B. (2017). Examining solutions to missing data in longitudinal nursing research. *Journal for Specialists in Pediatric Nursing*, 22(2), 10.1111/jspn.12179.

He, Y., Yucel, R., & Raghunathan, T. E. (2011). A functional multiple imputation approach to incomplete longitudinal data. *Statistics in Medicine*, 30, 1137–1156.

Audigier, V., White, I. R., Jolani, S., Debray, T. P. A., Quartagno, M., Carpenter, J., van Buuren, S., & Resche-Rigon, M. (2018). Multiple Imputation for Multilevel Data with Continuous and Binary Variables. *Statistical Science*, 33(2), 160–183.

Schafer, J. L., & Zhao, J. H. (2023). pan: Multiple imputation for multivariate panel or clustered data (R Package Version 1.9), retrieved from: https://cran.r-project.org/web/packages/pan/pan.pdf

Grund, S., Lüdtke, O., & Robitzsch, A. (2015). Multiple imputation of missing covariate values in multilevel models with random slopes: a cautionary note. *Behavior Research Methods*, 48(2), 640–649.

Guevara Morel, G., et al. (2022). Dealing with missing data in real-world data: A scoping review of simulation studies. *Journal of Statistical Software.*

Diggle, P., Liang, K.Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data* (2nd Edition). Oxford University Press.

Sousa, J. and Lopez, V. (2008). Erp impact on number of employees and personnel costs in small and medium sized enterprises - a panel data appro*ach. Proceedings of the Tenth International Conference on Enterprise Information Systems.* 1(1).

Ahapova, V. (2020). Media and economic growth: country groups difference via panel data. *European Journal of Business Management and Research*, 5(5).

Hsiao C, Luke Chan MW, Mountain DC, Tsui, KY (1989) Modeling Ontario regional electricity system demand using a mixed fixed and random coefficients approach. *Regional Science and Urban Economics,* 19 (4), 567–587.

Guan, D., Comite, U., Sial, M., Salman, A., Zhang, B., Gunnlaugsson, S., … & Mentel, G. (2021). The impact of renewable energy sources on financial development, and economic growth: the empirical evidence from an emerging economy. *Energies*, 14(23), 8033.

Tawfik, A., Sciarini, P., & Horber, E. (2011). Putting voter turnout in a longitudinal and contextual perspective: an analysis of actual participation data. *International Political Science Review*, 33(3), 352-371.

Mood, C. (2010). Neighborhood social influence and welfare receipt in Sweden: a panel data analysis. *Social Forces*, 88(3), 1331-1356.

Hsiao, C. (2007). Panel data analysis—advantages and challenges. *TEST*, 16(1), 1–22.

G. E. P. Box and D. R. Cox (1964). An Analysis of Transformations, *Journal of the Royal Statistical Society. Series B (Methodological)*, 26 (2), 211-252.

Xu, H., Lee, S., & Eom, T. (2007). Introduction to panel data analysis: Concepts and Practices. *Handbook of Research Methods in Public Administration*, 571-587.

Hassaballa, H. (2022). Studying the relationship between women and the environment in developing countries. *Journal of Chinese Economic and Foreign Trade Studies*, 15(3), 298-315.

Omolara, A., Ayobami, A., Omowumi, A., & Olatunde, O. (2021). Hierarchical bayesian estimation of random cofficient dynamic panel data model. *The International Journal of Science & Technoledge*, 9(4).

Diggle, P., Liang, K.Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data* (2nd Edition). Oxford University Press.

Liang, K.Y and Zeger, S.L. (1986), Longitudinal Data Analysis Using Generalized Linear Models, *Biometrics*, 73(1), 13-22.

Zeger, S. L., & Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, *42*(1), 121-130.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817-838.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied longitudinal analysis*. John Wiley & Sons.

Patterson and Thomson (1971) Recovery of inter-block information when block sizes are unequal, *Biometrika*, 58 (3), 545-554.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics, 38*(4), 963-974.

Dong, Y. and Peng, C. (2013). Principled missing data methods for researchers. Springerplus, 2(1).

Wang, C. and Hall, C. (2009). Correction of bias from non-random missing longitudinal data using auxiliary information. *Statistics in Medicine*, 29(6), 671-679.

Beesley, L. J., Bondarenko, I., Elliott, M. R., Kurian, A. W., & Taylor, J. M. G. (2021). Multiple imputation with missing data indicators. *Statistical methods in medical research*, 30(12), 2685–2700.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons Inc., New York.

Demirtas, Hakan (2004). Modeling Incomplete Longitudinal Data. *Journal of Modern Applied Statistical Methods*, 3(2), 305-321.

Li, P., Stuart, E. A., & Allison, D. B. (2015). Multiple Imputation. *JAMA*, 314(18), 1966.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics,* 8, 25.

Van Buuren, S. & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3).

Van Buuren, S. (2018). *Flexible Imputation of Missing Data* (2nd ed.). Chapman and Hall/CRC.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*(4), 330-351.

EMRA. (2021). Republic of Turkey Energy Market Regulatory Authority, Strategy Development Department, *Turkish Natural Gas Market Report,* Ankara.

Taşpınar, F. (2015). Time Series Models for Air Pollution Modelling Considering the Shift to Natural Gas in a Turkish City, *CLEAN – Soil, Air, Water , 43 (7), 980–988.*

Sen, D., Günay, M. E., and Tunç K. M. M. (2019). Forecasting annual natural gas consumption using socio-economic indicators for making future policies, *Energy,* 173, 1106-1118.

Karadede Y., Özdemir G. and Aydemir E. (2017). Breeder Hybrid Algorithm Approach for Natural Gas Demand Forecasting Model, *Energy,* 141, 1269-1284.

Özdemir G., Aydemir E., Olgun M. and Mulbay Z. (2016). Forecasting of Turkey natural gas demand using a hybrid algorithm, *Energy Sources, Part B: Economics, Planning, and Policy,* 11 (4), 295-302.

Kathryn S. Dawson , Chris Gennings and Walter H. Carter (1997). Two Graphical Techniques Useful in Detecting Correlation Structure in Repeated Measures Data, *The American Statistician,* 51 (3), 275-283.

O'brien, R. M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity*, 41(5), 673–690.

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24), 4279–4292.