Freight generation and freight trip generation modeling

Doctor’s Thesis

FREIGHT GENERATION AND FREIGHT TRIP GENERATION MODELING

March 2020

Graduate School of Marine Science and Technology
Tokyo University of Marine Science and Technology
Course of Applied Marine Systems Engineering

LIDASAN AL HANZ SEIJI BASA

TABLE OF CONTENTS

Chapter 1 INTRODUCTION

1.1 Background

The movement of goods or freight transport is vital to the economic activities of a region. Not only is it essential for the everyday lives of citizens as a means of distribution of consumer goods and services but it is invaluable in the procurement process of raw materials for the manufacturing of consumer goods or other intermediate goods for other purposes, and equipment for the further production of other goods and services. It essentially enables businesses to operate and provide the necessary goods and services for an economy to grow. Freight transport is also what enabled trade across cities and countries, which lead to the existence of a globalized economy. Because of differences in labor and raw material costs between regions, countries, or even cities within a country, producing a single good may have had multiple freight transportation associated with it. Freight transport also plays a vital role in transporting goods, especially relief goods and medicines, to areas affected and stricken by disasters for humanitarian logistics purposes. Trade and freight transport are drivers of economic growth and development of a nation, and without freight transportation, goods and services will not be accessible to citizens who determine the demand for goods and services.

The importance of freight transportation to the local and global economy, to people's everyday lives, and to humanitarian logistics highlights its significance in the overall transportation system of a region, which includes private and public passenger transportation. This poses challenges from the viewpoint of having an integrated transport policy. Compared to passenger transport, there are many aspects to be considered in terms of integration in freight transport such as the involvement of many different companies in production and supply of a particular product, the object of study (whether vehicle or goods), and the effort that goes into the management of information flow during logistics management (Allen et al., 2010). Thus, bringing about greater integration in freight transport can mean different things, and is likely to be a relatively complex process that needs the involvement of many different organizations (Allen et al., 2010). This can be seen from the various key stakeholders in freight transportation, namely, shippers, freight carriers, residents, and administrators (Taniguchi et al., 2001). The inter-relations of the different stakeholders in freight transportation are shown in Figure 1.1.

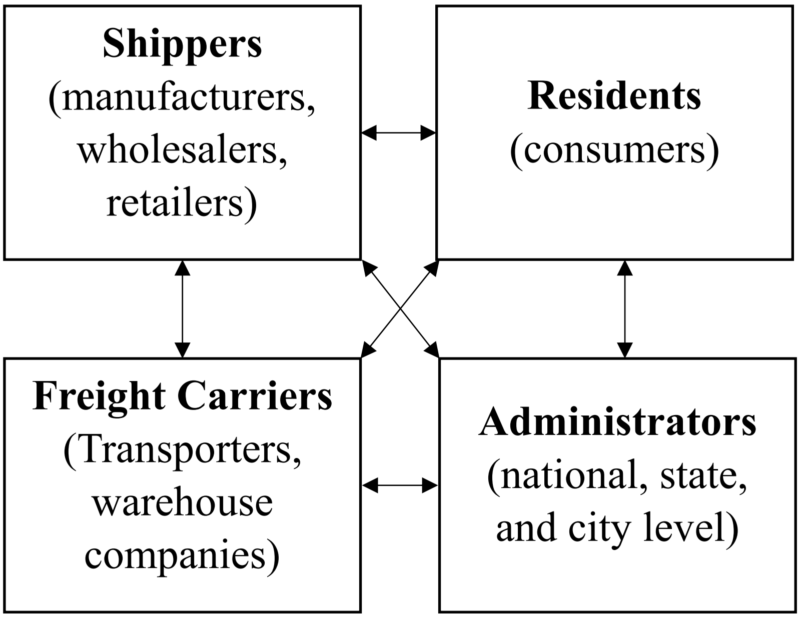

Shippers
(manufacturers, wholesalers, retailers)
Residents
(consumers)
Freight Carriers
(Transporters, warehouse companies)
Administrators
(national, state, and city level)

Figure 1.1 Stakeholders in freight transportation (Taniguchi et al., 2001)

Shippers are the customers for freight carriers who either send goods to other companies or persons or receive goods from them (Taniguchi et al., 2001). Freight carriers are those directly related to transporting the goods from origin to destination where there can be intermediate origins and destinations which are identified as transshipment points. Freight carriers typically attempt to minimize costs and maximize profits as well as maintain a certain level of service due to pressure from customers (Taniguchi et al., 2001). Residents are the people who live, work, and shop in the city (Taniguchi et al., 2001) are the end-consumers that are the source of the demand for goods and services. Administrators are the ones that attempt to enhance the economic development of a city or region (Taniguchi et al., 2001). Administrators or the public sector (government) are the ones essentially in charge of maintaining the balance between the interests of different stakeholders in the freight transport system. They also aim to provide better quality-of-life through investments that will make it easier for other stakeholders to function.

The above leads to the need for a better understanding of the different aspects of freight transportation to be able to predict their respective outcomes for more informed policies and decision making. This is because stakeholders in freight transportation have their own specific objectives and tend to behave differently from each other (Taniguchi et al., 2001) and will either want to optimize operations, lower costs, increase utility and profits, lower the externalities to the environment, and improve services and quality of life. This has led to a surge of studies focusing on freight transportation, particularly in modeling freight transportation in terms of jointly optimizing inputs, outputs, profits, costs, scheduling, and routes choices. Often, studies related to freight transportation are tied to private of publicly funded projects for practical purposes like conducting forecasts through prediction models. These prediction models are for freight-related indices and measures of production of goods and services, which are a precursor to measures of economic performance and growth, and evaluations of investments and infrastructure projects.

While there has been a surge of studies focusing on freight transportation, it still generally follows the traditional four-step modeling approach that is well established in passenger transportation modeling, and transportation planning literature in general. The four-step model consists of trip generation, trip distribution, modal split, and traffic assignment. The output in the trip generation step is usually the number of trips generated (Rodrigue et al., 2016), and in the context of freight transportation, an additional layer is considered in the form of freight volume generated. The output in trip distribution is usually a flow matrix between spatial units (Rodrigue et al., 2016). Modal split disaggregates movement between origins and destinations by modes (Rodrigue et al., 2016), and there can be more than one mode used for a specific shipment, especially when there are intermediate locations (such as warehouses and distribution centers) or transshipment points. Traffic assignment loads the estimated trips onto the transportation network (Rodrigue et al., 2016). In line with this, Tavasszy and de Jong (2014) distinguished three primary layers or markets of freight systems: (1) the commodity market, (2) the inventory logistics services market, and (3) the transport logistics services market. The commodity market is where the consumption and production of goods occur and are primarily driven by producers and consumers. The producers are on the receiving end of goods flows, goods such as raw materials and other inputs for manufacturing, and consumers are on the sending end of freight when it comes to waste or return shipments (Tavasszy and de Jong, 2014). Consumers and households shape the final demand for goods, and their decision includes, by analogy, the residential location, their consumption patterns, and the way they deal with waste, (Tavasszy and de Jong, 2014). The inventory services market deals mostly with keeping logistics costs low (by bundling shipments and transport flows) and maintaining high levels of service with proximity to markets (Tavasszy and de Jong, 2014). The intermediate inventories where transshipment happens changes the spatial pattern of trade where new origins and destinations for transport are created (Tavasszy and de Jong, 2014) and is where consolidation or deconsolidation of shipments happen which adds another factor in the decision-making process in freight transportation. The transport logistics services market is concerned with the transport mode of freight which can be road, rail, water, and air or a combination thereof. The transport mode of freight is the most discussed point of intervention for freight transport policies and each mode of transport offers different specialized means of transport appropriate to different type of goods and shipment sizes (Tavasszy and de Jong, 2014). When all modes of transport are available, decision makers intensively optimize transport (Tavasszy and de Jong, 2014) to either decrease cost and travel time, increase profit, or a compromise while maintaining a high level of service when feasible.

The decisions leading to freight transport are not independent (Tavasszy and de Jong, 2014), as is clear from the different stakeholders and the different corresponding markets (which reflects the four-step model for passenger transport) in freight transport. In contrast to passenger transport, where a decision-maker is usually a single person, freight transport involves a multitude of decision-makers, which lead to freight transport and, thus, freight vehicle trips.

1.2 Problem Identification

Freight models have a tendency to overfit the data by increasing the number of independent variables in the model to improve model fit as well as to not consider unobserved factors, particularly the spatial effects in the freight system. While increasing the number of predictor variables will increase a model’s fit to data, an excellent model fit does not necessarily mean that the model is suitable for forecasting because overfitted models tend to perform poorly in prediction. On the other hand, unobserved variables are neglected, particularly the spatial dependence of some variables in freight transportation, which lead to

1.3 Objectives of the Study

This thesis aims to tackle the problems of overfitting of freight volume generation and freight trip generation, the dependence of freight trip generation to the location of logistics facilities, and the issues of unobserved spatial dependencies in freight trip generation. The specific objectives of this thesis are as follows:

1) To recommend an alternative method to deal with the overfitting issues in freight transportation models;

2) To be able to consider unobserved factors particularly the spatial dependencies in freight trip generation through modeling of spatial autocorrelation

1.4 Significance of the Study

The studies conducted in this thesis are significant in proposing improved methods for applied modeling of freight volume and freight transport generation. The study will show how to deal with overfitting and unobserved effects (spatial dependence) by using penalized regression methods and spatial regression.

1.5 Scope

While the freight transport system is complex and involves different stakeholders and markets and different layers or steps in modeling, which are interrelated, this thesis will primarily focus on the generation of freight volume and freight trips.

1.6 Thesis Composition

Chapter 1 introduces the study; Chapter 2 presents the literature review of freight generation and freight trip generation models; Chapter 3 aims to solve the overfitting problem in freight volume generation by applying a Bayesian varying (random) intercept model for national freight volume in Japan. Chapter 4 presents a method to consider location choice and location variables of logistics facilities to freight trip generation. Chapter 5 introduces sparse regression methods as an alternative to the classical regression method to deal with overfitting for freight trip generation. Chapter 6 shows how unobserved spatial variables can be considered for modeling freight trip generation through regression that considers spatial autocorrelation. Chapter 7 summarizes, concludes, and discusses the implications.

Figure 1.2 below shows the thematic relationship of the respective chapters in the study, particularly those directly dealing with freight volume and freight trip generation modeling. As stated in the problem identification and objectives of the study, the two main themes in this study is to deal with the overfitting of freight generation and freight trip generation models, and to consider the spatial dependencies in freight trip generation.

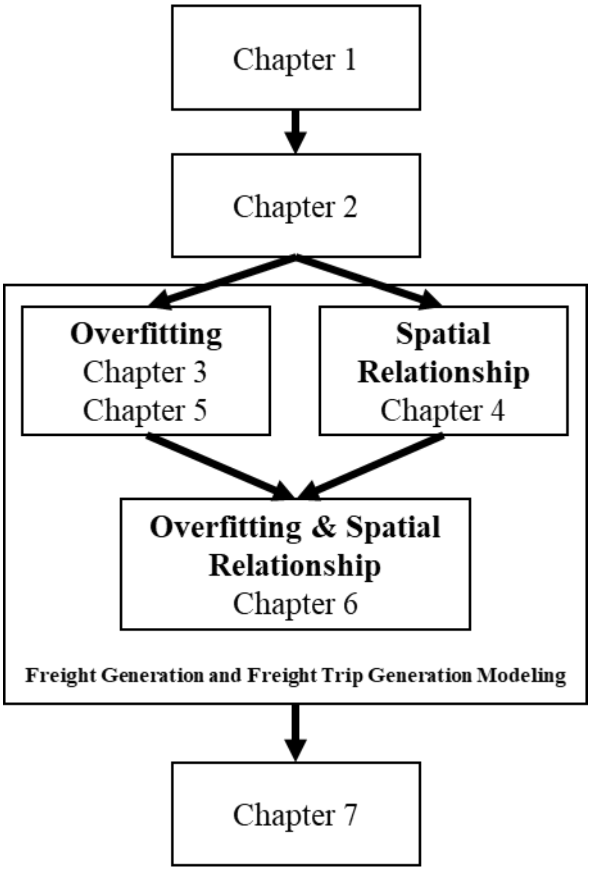


Figure 1.2 Thematic relationship of the chapters

Chapter 3, modeling the national freight volume generation of Japan, and Chapter 5, sparse regression as a method for trip generation modeling in Kanto, Japan specifically deals with overfitting by introducing a varying-intercepts model to improve out-of-sample prediction, and by conducting variable selection through sparse regression, respectively. On the other hand, Chapter 4, logistics facility allocation, size, and freight trip generation of Tokyo Metropolitan Area introduces methodologies that takes into account the numerous zero values in the data which is a representation of the spatial dependency of the allocation of freight facilities and truck trip generation. Chapter 6, truck trip generation modeling considering spatial auto-correlation in Kanto and Kansai, Japan using spatial regression, both considers the overfitting issue and spatial dependencies in the data by first conducting a variable selection to narrow down the number of independent variables relevant to modeling truck trip generation, then proceeding to model the spatial dependencies in the data through spatial regression. Although the methodology of considering the spatial dependencies in the freight trip generation data differs between chapter 4 and chapter 6, it is the objective of this study to introduce methods that will be appropriate depending on the different characteristics of freight trip generation data.

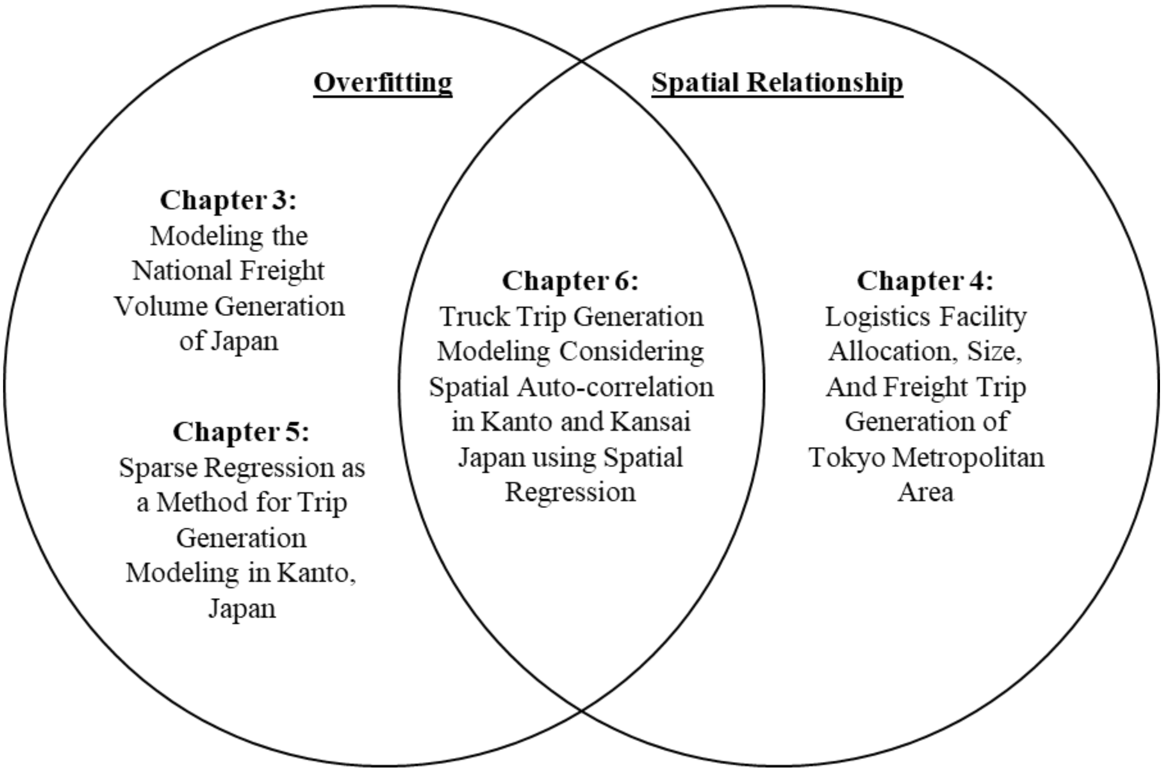

Overfitting
Spatial Relationship
Chapter 3:
Modeling the National Freight Volume Generation of Japan
Chapter 5:
Sparse Regression as a Method for Trip Generation Modeling in Kanto, Japan
Chapter 6:
Truck Trip Generation Modeling Considering Spatial Auto-correlation in Kanto and Kansai Japan using Spatial Regression
Chapter 4:
Logistics Facility Allocation, Size, And Freight Trip Generation of Tokyo Metropolitan Area

Figure 1.3 Venn diagram of the relative thematic relationship of the freight and freight trip generation models

Figure 1.3 highlights the relative differences of the chapters with respect to the main themes of the study. While chapter 3 and chapter 5 both deals with the overfitting of freight volume generation and freight trip generation models, respectively, chapter 4 deals with the spatial dependencies in the data. On the other hand, chapter 6 is the intersection between overfitting and the spatial relationships as it shows how the two issues are tackled by utilizing a method for dealing with overfitting introduced in chapter 5 and applying the said method to aid the application of a model that considers the spatial dependency of freight trip generation. And while the different chapters are demonstration of the different models and how they deal with the issues of overfitting and spatial dependencies, they give insights to how different data and their context are handled through modeling. As most cases of freight transportation modeling are based on observational data, that is, data that are collected not for the sole purpose of modeling but for statistical summaries, it is important that different approaches to model the freight transportation data for forecasting purposes be considered and explored.

Chapter 2 LITERATURE REVIEW OF FREIGHT GENERATION AND FREIGHT TRIP GENERATION MODELS

2.1 Overview of Freight Transport Modeling

Tavasszy & de Jong (2014) gave an overview of the different decisions and markets of the freight transport system as well as the different freight modeling approaches from a theoretical and practical perspective. Before analyzing the different actors in a freight system, we must first determine the decision-makers, what decisions they make, and how those decisions affect the freight transport system (Tavasszy and de Jong, 2014). This gives a clear overview of the system in general. Decisions in transport policy are generally strategic (5 to 10 years), tactical (months to years), or operational (days to months) in the timeframe. Strategic decisions involve major investments and cannot be reviewed frequently; tactical decisions are related to smaller investments and are reviewed frequently but still has a lag time; and operational decisions are those that can be taken at discretion and have a short review period in the planning and management cycles (Tavasszy & de Jong, 2014).

In conjunction with the timeframes, there are three main layers or markets of freight systems: the exchange of goods (commodity market), the inventory networks (inventory services market), and the transport organization (transport logistics services market) (Tavasszy & de Jong, 2014). The exchange of goods market revolves around producers and consumers; inventory networks are spatial forms of organization of inventories that provide storage, consolidation and/or deconsolidation of flows at intermediate locations in between production and consumption areas which aims to keep logistics costs low and maintain high service levels; and transport organizations deal with the choice of modality or mode of transportation as well as the shipment size or a joint decision on both (Tavasszy & de Jong, 2014) and to an extent, the network assignment and route choice. There can be direct and mutual dependence between decisions in freight transport and, although it is desirable to strive for a comprehensive and integrated model, in practice, sub-models are synthesized through integrative theory and where empirically feasible (Tavasszy & de Jong, 2014). The framework for the type of decisions and markets in freight transport systems leads to different models that reflect the market or layer in the freight transport system and the specific decisions they aim to describe. Table 1 summarizes the different model types. See Tavasszy & de Jong (2014) for more discussion.

Table 2.1. Model types based on decisions and markets (Tavasszy & de Jong, 2014).

Market	Partial Models	Disciplinary Focus
Production/consumption and trade	• Production/consumption: Input/output models	• Input/output economics
	• Trade: Gravity models	• Engineering
	• Combined: Spatial computable general equilibrium models and derivatives	• Economic geography
	• Freight generation models	• Econometrics
Inventory logistics	• Shipment size choice	• Operations research
	• Inventory chain models	• Discrete choice theory
Transport logistics	• Mode choice models	• Discrete choice theory
	• Freight to trip conversion models	• Engineering
	• Mode and route choice: supernetworks	• Network modeling

2.2 Freight Generation and Freight Transport Generation

Freight volume generation or freight generation (FG) is defined as the amount of cargo generated, specifically the amount of cargo produced or consumed, while freight trip generation (FTG) is the freight traffic required to transport cargo or the number of freight trips generated (Holguin-Veras et al., 2014; Holguín-Veras et al., 2011). Holguín-Veras et al. (2011) argue that FG and FTG must be treated as separate concepts because FTG is the output of logistic decisions, while FG is determined by the economics of productions and consumption. The freight transport modeling literature has established that the size of establishments is a determining factor of FG, i.e., as the size of businesses increases, the FG also increases. However, although it is expected that FG increases with business size, FTG does not necessarily increase proportionately; increases in FG can be accommodated by smaller increases in shipment size that may not necessarily have an impact on FTG, which could lead to changes in vehicle or mode and even a decrease in FTG (Holguín-Veras et al., 2011). The fact that FG and FTG are different phenomena in the overall context of freight transport modeling (Holguín-Veras et al., 2011) requires FG and FTG to be treated separately.

Comprehensive reviews on the state-of-the-art and recent developments in freight transport modeling were conducted by different groups that determined research gaps and opportunities in the literature (Chow et al., 2010; de Jong et al., 2013; Tavasszy et al., 2012). Chow et al.

(2010) reviewed the recent advances in freight forecasting models and the data requirements for model development in the USA and concluded the need for including dynamic shipper-carrier interactions in freight transport modeling and that the development of hybrid models such as the integration of regional logistics models with urban truck touring models will result in new problems consistency in the results of multiple models. de Jong et al. (2013) reviewed the recent developments in freight transport modeling in Europe and determined that the introduction of logistics has been the main improvement of the models. However, de Jong et al. (2013) assessed that most practical freight transport models still lack logistics choice making. This is mostly because of constraints in the readily available data that are suitable for modeling logistic decisions and its relationship to FG and FTG. The development of logistics models and their integration in FG and FTG models in Europe has only been possible due to the availability of data beyond those compulsory aggregate freight transport statistics (de Jong et al., 2013). In a similar vein, Tavasszy et al. (2012) reviewed the developments in freight modeling regarding the state-of-the-art in representing logistics focusing on service and cost drivers of changes in logistics networks and how theses affect freight transport. In agreement with the findings of de Jong et al. (2013) concerning the lack of data befitting of including logistics in freight transport models, Tavasszy et al. (2012) also noted that freight modeling requires data on the various logistics infrastructures, the quality of costs of logistics services, and transport flows.

The relationship between land-use and transportation is very-well established and a well-researched area (Geurs & van Wee, 2004; James et al. 1972; Wegener, 2004; Newman & Kenworthy, 1996) that it is already standard to consider land-use-transport (LUT) interactions in city/town planning and regional planning. However, there is a lack of research tackling the interaction between logistics land-use and transportation, especially when utilizing models for policy analysis. Only a few have attempted to do so due to the complexities of the logistics sector (Hesse, 2002; Hesse, 2004; Wagner, 2010).

Previous research on logistics facilities distribution focuses on decisions where to locate. For instance, evaluation criteria for the location selection of city logistics centers were formulated by combing economic, environmental, and social sustainability indicators through a fuzzy multi-attribute group decision-making method (Rao et al., 2015). Lindsey et al. (2014) used an econometric approach to evaluate longitudinal data of metropolitan markets wherein a methodology was developed to rank 20 metropolitan markets from 1997 to 2007 based on their potential for industrial space using macroeconomic, demographic, and freight flows as input variables. In relation to econometric modeling, Woudsma et al. (2008) applied a spatial-temporal modeling approach to quantify the effects of transportation system performance on the patterns of logistics land-use. Woudsma et al. (2015) investigated logistics sprawl and its relation to locating and identifying facilities. Sakai et al. (2016) presented the historical transition of logistics facilities in TMA from 1980 to 2003, which revealed that the asset pricing bubble in Japan during the period of 1986 to 1991 was a significant factor in the decentralization of logistics facilities into the suburbs. Iwakata et al. (2015) highlighted the importance of accessibility to interchanges and expressways for mega distribution centers in TMA. Hong (2007) found that the location of foreign logistics firms in Chinese cities depended on transport conditions in terms of the roadway, railway, and waterway, as well as market size, labor quality, agglomeration economies, and government incentives.

Freight trip generation, on the other hand, has been modeled through various methods in the past. The most basic of which is through the use of trip generation rates (Kulpa, 2014; Sorratini & Smith 2000), which determines the number of truck trips generated per unit of the independent variable (e.g., number of trips per number of employed persons). Multiple linear regression has also been used in numerous papers modeling truck trip generation either to develop generation rates or directly forecast truck trip generation (Tadi & Balbach, 1994; Holguín-Veras et al. 2002; El-maghraby, 2000; Sorratini & Smith, 2000; Kulpa, 2014). Truck trip generation models fall under the vehicle-based models as opposed to commodity-based models in road freight transport trip generation modeling (Kulpa, 2014). While these methods have been the standard in urban-transport planning, the modeling of a truck trip generation focused on certain types of land-use or facility one at a time (Tadi & Balbach, 1994; Holguín-Veras et al. 2002). This is uniquely flawed when considering mixed land-use patterns, particularly at the regional level.

There is a disconnect between research on logistics land-use and other land-use classifications and actual truck trip generation wherein past research considers one aspect independent of the others and vice-versa. Therefore, in line with one of the objectives of this study to consider unobserved effects to freight trip generation, a methodology of linking land-use and transport in the context of freight transport (vehicle-based) is presented as well as simultaneously account for allocation and size of logistics facilities and truck trip generation considering all land-use classifications available in the data.

Chapter 3 MODELING THE NATIONAL FREIGHT VOLUME GENERATION OF JAPAN

3.1 Introduction

There are different approaches to modeling freight volume generation or freight generation (FG), and it differs depending on the resolution of analysis, i.e., if the unit of analysis is at the international level, national level, or the urban/local level. The methods applied in modeling FG vary depending on the available data, or the data that can be feasibly collected for the different levels of units of analyses. This means that there are various methods of modeling freight volume generation, and this is evident from the abundance of methods in the literature where authors present the novel methodology they used. However, most of these methodologies are dependent on availability or the collected data. A common approach to improving model fit is to include as many independent variables from the available data. This naturally improves the fit of the freight model to data. However, a model with a good fit to data does not necessarily mean it has a good predictive capacity. Here lies the problem because models are developed in freight transportation primarily for prediction and forecasting and mostly for policy evaluation. In this chapter, an approach to modeling FG using the national freight volume generation of Japan is presented with the aim of dealing with the overfitting issue in freight transportation models.

3.2 Data Abstract

The national freight volume data used in this chapter has 1,504 samples. There are 6 variables in the data, namely, year, prefecture, the population (in 1,000s), gross regional product (GRP), type of goods, and weight in tons of goods generated. There are 4 years considered, namely, 2000, 2005, 2010, and 2015. There are 47 prefectures in Japan, and the population in (in 1,000s) and GRP in millions are considered for each prefecture for each of the 4 years. There are eight types of goods considered: agriculture, wood, mining, machine, chemical, small machinery, miscellaneous industry goods, and special goods. A sample of the data used is shown in *Table 3.1*.

Table 3.1 Sample of national freight volume data

year	num	prefecture	pop.1000	GRP.mill	goods	ton
2000	1	Hokkaido	5683	20471299	agriculture	13062354
2000	1	Hokkaido	5683	20471299	wood	2100076
2000	1	Hokkaido	5683	20471299	mine	47400509
2000	1	Hokkaido	5683	20471299	machine	8638957
2000	1	Hokkaido	5683	20471299	chemical	68047304
2000	1	Hokkaido	5683	20471299	Smachine	13457707
2000	1	Hokkaido	5683	20471299	miscind	3209660
2000	1	Hokkaido	5683	20471299	special	8445845

3.3 Framework of Analysis

The effects of overfitting are shown using different measures of model fit and predictive accuracy. A widely used measure of model fit is the r-squared, which measures the proportion of the variance in the data captured by the model. Including more variables would lead to a better fit because the model will learn more from the data. However, this does not necessarily mean that the model is suitable for the prediction of data outside of the sample used for learning. Thus, we compare side-by-side the r-squared and an established method of measuring the predictive accuracy of regression models under a Bayesian framework. We use the Bayesian framework because it can easily allow the estimation of more complex models, such as those with varying-intercepts.

3.4 Including Varying Effects in Modeling Freight Volume Generation

One approach to improve predictive accuracy is to allow for varying effects in the regression model. Here varying effects are considered in the form of varying-intercepts in regression models. While varying-intercepts can be considered in the classical regression method, the Bayesian method of estimation allows the varying-intercepts to be correlated through a hyperparameter, which represents the mean effect of each varying intercept. The advantage is that we can estimate this from the data and allow intercepts to covary. A total of 24 national freight FG models were estimated, as shown in the national FG models Fit1 to Fit24 below. The dependent variable is the log of the total volume of freight generated, $\log(y_i)$, and the independent variables are the population and GRP of the prefectures, the dummy for the prefecture, the dummy for the year, and the dummy for the goods type. The independent variables population and GRP are standardized to have a mean equal to zero and a standard deviation of one. Fit1 to Fit3 is the conventional method of incorporating dummy variables for each prefecture, year, and goods type. The difference among Fit1 to Fit3 is that, in addition to the dummy variables for each prefecture, for each year, and for each goods type, Fit1 consists of both population and GRP, Fit2 consists of only the independent variable population, and Fit3 consists of only the independent variable GRP. The rest of the national FG models estimated are of the varying effects type of models where the varying effects are represented as varying-intercepts. Fit4 to Fit24 are all varying-intercepts models, and the difference is what variables are considered as varying-intercepts and whether population and/or GRP is included. Fit4 to Fit6 are varying-intercepts models where prefecture is the varying-intercept variable. Fit7 to Fit9 are varying-intercepts models where year is the varying-intercept variable. Fit10 to Fit12 are varying-intercepts models where goods type is the varying-intercept variable. Fit13 to Fit15 are varying-intercepts models where prefecture and year are the varying-intercept variables. Fit16 to Fit18 are varying-intercepts models where prefecture and goods are the varying-intercept variables. Fit19 to Fit21 are varying-intercepts models where year and goods are the varying-intercept variables. The last set of national FG models, Fit22 to Fit24, are varying-intercepts models where prefecture, year and goods are the varying-intercept variables.

Fit1:$\log(y_i) = \alpha + \alpha_{pref1} + \cdots + \alpha_{pref46} + \alpha_{year1} + \cdots + \alpha_{year3} + \alpha_{goods1} \cdots \alpha_{goods7} + \beta_{pop}x_{pop} + \beta_{GRP}x_{grp}$

Fit2:$\log(y_i) = \alpha + \alpha_{pref1} + \cdots + \alpha_{pref46} + \alpha_{year1} + \cdots + \alpha_{year3} + \alpha_{goods1} \cdots \alpha_{goods7} + \beta_{pop}x_{pop}$

Fit3:$\log(y_i) = \alpha + \alpha_{pref1} + \cdots + \alpha_{pref46} + \alpha_{year1} + \cdots + \alpha_{year3} + \alpha_{goods1} \cdots \alpha_{goods7} + \beta_{GRP}x_{grp}$

Fit4: $\log(y_i) = \alpha + \alpha_{pref} + \beta_{pop}x_{pop} + \beta_{GRP}x_{grp}$

Fit5: $\log(y_i) = \alpha + \alpha_{pref} + \beta_{pop}x_{pop}$

Fit6: $\log(y_i) = \alpha + \alpha_{pref} \qquad\qquad + \beta_{GRP}x_{grp}$

Fit7: $\log(y_i) = \alpha + \alpha_{year} + \beta_{pop}x_{pop} + \beta_{GRP}x_{grp}$

Fit8: $\log(y_i) = \alpha + \alpha_{year} + \beta_{pop}x_{pop}$

Fit9: $\log(y_i) = \alpha + \alpha_{year} \qquad\qquad + \beta_{GRP}x_{grp}$

Fit10: $\log(y_i) = \alpha + \alpha_{goods} + \beta_{pop}x_{pop} + \beta_{GRP}x_{grp}$

Fit11: $\log(y_i) = \alpha + \alpha_{goods} + \beta_{pop}x_{pop}$

Fit12: $\log(y_i) = \alpha + \alpha_{goods} \qquad\qquad + \beta_{GRP}x_{grp}$

Fit13: $\log(y_i) = \alpha + \alpha_{pref} + \alpha_{year} + \beta_{pop}x_{pop} + \beta_{GRP}x_{grp}$

Fit14: $\log(y_i) = \alpha + \alpha_{pref} + \alpha_{year} + \beta_{pop}x_{pop}$

Fit15: $\log(y_i) = \alpha + \alpha_{pref} + \alpha_{year} \qquad\qquad + \beta_{GRP}x_{grp}$

Fit16: $\log(y_i) = \alpha + \alpha_{pref} + \alpha_{goods} + \beta_{pop}x_{pop} + \beta_{GRP}x_{grp}$

Fit17: $\log(y_i) = \alpha + \alpha_{pref} + \alpha_{goods} + \beta_{pop}x_{pop}$

Fit18: $\log(y_i) = \alpha + \alpha_{pref} + \alpha_{goods} \qquad\qquad + \beta_{GRP}x_{grp}$

Fit19: $\log(y_i) = \alpha + \alpha_{year} + \alpha_{goods} + \beta_{pop}x_{pop} + \beta_{GRP}x_{grp}$

Fit20: $\log(y_i) = \alpha + \alpha_{year} + \alpha_{goods} + \beta_{pop}x_{pop}$

Fit21: $\log(y_i) = \alpha + \alpha_{year} + \alpha_{goods} \qquad\qquad + \beta_{GRP}x_{grp}$

Fit22: $\log(y_i) = \alpha + \alpha_{pref} + \alpha_{year} + \alpha_{goods} + \beta_{pop}x_{pop} + \beta_{GRP}x_{grp}$

Fit23: $\log(y_i) = \alpha + \alpha_{pref} + \alpha_{year} + \alpha_{goods} + \beta_{pop}x_{pop}$

Fit24: $\log(y_i) = \alpha + \alpha_{pref} + \alpha_{year} + \alpha_{goods} \qquad\qquad + \beta_{GRP}x_{grp}$

3.5 Bayesian Inference Framework of Parameter Estimation

The theoretical foundation for parameter estimation under the Bayesian Inference framework is the Bayes' theorem. Bayes' theorem and probability theory is used to derive the distribution of parameters of a statistical model given the data, called the posterior distribution $p(\theta|y)$. To utilize Bayes' theorem to estimate the posterior distribution, the beliefs about the parameters θ, before taking into account the data, must be specified using a probability distribution called the prior distribution $p(\theta)$ of the parameters θ. Also, a probability model must be chosen for the data y given the parameters θ to complete the factors necessary for Bayes' theorem to derive the posterior distribution of parameters $p(\theta|y)$. The posterior distribution of parameters $p(\theta|y)$ using Bayes' theorem is derived as follows.

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

The numerator is the product of the likelihood $p(y|\theta)$ and the prior distribution of parameters $p(\theta)$, while the denominator is the sampling distribution $p(y)$, which is also called the "evidence" or "average likelihood" (McElreath, 2018). It is the average probability of the data where the probability is taken for all possible values of the parameters from its prior distribution, as shown below.

$$p(y) = \int p(y|\theta)p(\theta)d\theta$$

There are different methods of calculating the posterior distribution of parameters. However, in the advent significant improvement in computing power and storage capacity of computers, the most powerful methods to estimate the posterior distribution of parameters have become simulation methods such as Markov chain Monte Carlo (MCMC) methods because of its capacity to estimate complex model specifications such as hierarchical or multilevel models. This also means that simple model specifications will not pose any problem for MCMC methods.

The national freight volume generation models in this chapter apply MCMC to estimate the posterior distribution of parameters conditional on the national freight volume generation data of Japan. The estimation of the parameters through the posterior distribution was conducted using the probabilistic modeling language "Stan" (Carpenter et al., 2017) through "brms" (Bürkner, 2018, 2017), the interface R package in the R programming language (R Development Core Team, 2018). Stan applies MCMC algorithms known as Hamiltonian Monte Carlo (HMC) (Neal, 2011) to simulate the posterior distribution of parameters and thus estimate the model parameters conditional on the data. The advantage of using the HMC algorithm is that it is more efficient because it does not need many samples to describe the posterior distribution, it requires less computation time, and outperforms other algorithms when models become more complex (McElreath, 2018).

3.6 Measures of Model Fit and Predictive Accuracy

Various measures of model fit and predictive accuracy were used to evaluate the classical regression model and the varying-intercepts model. The model fit of classical regression models are traditionally evaluated using the coefficient of determination, R^2, which measures the proportion of variance in the data explained by the model. The classical R^2 is defined as follows:

$$classical\,R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

where the numerator of the fraction above is the Residual Sum of Squares (RSS), and the denominator is the Total Sum of Squares (TTS). However, a problem arises under the Bayesian treatment of regression models where it is possible for the classical formula classical R^2 to be greater than 1 (Gelman et al., 2019; Tjur, 2009). Hence, instead of using the classical R^2 to evaluate the fit of the nationl FG models, a generalization of the classical R^2 under the Bayesian framework that is based on variance-decomposition is used as follows (Gelman et al., 2019):

$$Bayesian\ R^2 = \frac{V_{n=1}^{N} y_n^{pred_s}}{V_{n=1}^{N} y_n^{pred_s} + var_{res}^{s}}$$

where $V_{n=1}^{N} y_n^{pred_s}$ is the variance of the modeled predictive means and var_{res}^{s} is the residual variance. In the Bayesian framework, instead of point estimates of the parameters $\hat{\theta}$, a posterior distribution of parameters $\theta^s, s = 1, \dots, S$, conditional on the data is estimated. Hence the subscript s in the factors of the $Bayesian\ R^2$.

It is not enough to measure the model fit of freight transport models because, ultimately, freight transport models are estimated for the purpose of evaluating freight transport solutions and forecasting future FG and FTG. The predictive performance of freight transport models on future data must be evaluated to determine the appropriate model for policy evaluation and forecasting impacts and externalities. There are two approaches to measuring the predictive accuracy of regression models: information criteria and cross-validation (McElreath, 2018). However, only cross-validation will be used to evaluate the national FG models. The established measure to compare the predictive accuracy of different models is called the log-probability score $S(q) = \Sigma_i \log(q_i)$, which was derived from information entropy (McElreath, 2018). However, similar to the treatment of R^2 for models estimated under the Bayesian framework, the entire posterior distribution is used to evaluate the log-probability score. Thus, the log-probability score for a Bayesian model called the log-pointwise-predictive density ($\widehat{lppd}$) is defined as follows:

$$\widehat{lppd} = \sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\theta^s)\right)$$

where S is the total number of samples from the posterior distribution and θ^s is the s^{th} set of parameters sampled from the posterior distribution. When comparing models estimated from the same dataset, a higher $\widehat{lppd}$ indicates a better predictive accuracy.

A common methodology for assessing the predictive accuracy of statistical models is cross-validation. Cross-validation is when we leave-out a small part of the data and test the performance of the statistical model on that small part of the data using the model estimated from the rest of the data. A specific type of cross-validation is the leave-one-out (LOO) cross-validation, which will be the basis of comparing the national FG models presented in this chapter. However, a known difficulty of implementing LOO cross-validation is that it can be time-consuming and heavy on resources, especially in the Bayesian framework, as for each observation, a posterior distribution must be estimated. A technique to estimate the LOO cross-validation without the need to go through each sample is to use the relative importance provided by each sample in the posterior distribution (Vehtari et al., 2017). Vehtari et al. (2017) introduced a method and proved that the LOO cross-validation could be approximated with high accuracy without the need to conduct the actual LOO cross-validation. They showed that by using only the posterior distribution and the relative importance of each sample in the data, the LOO cross-validation is approximated by incorporating weights in the calculation of the $\widehat{lppd}$. The weights for the importance sampling undergo a smoothing process using the pareto distribution so that the weights become more reliable (McElreath, 2018) hence it is called the Pareto-Smoothed Importance Sampling (PSIS-LOO) cross-validation and is referred here as $\widehat{lppd_{psis-loo}}$. The $\widehat{lppd_{psis-loo}}$ is defined as follows:

$$\widehat{lppd_{psis-loo}} = \sum_{i=1}^{n} \log\left(\frac{\sum_{s=1}^{S} w_i^s p(y_i|\theta^s)}{\sum_{s=1}^{S} w_i^s}\right)$$

where the terms are the same as the definition of $\widehat{lppd}$ above with the addition of the importance-sampling weights, w_i^s . The $\widehat{lppd_{psis-loo}}$ will be the basis for assessing and comparing the predictive accuracy of the national FG models. Similar to the non-weighted version, a model with a higher $\widehat{lppd_{psis-loo}}$ indicates a better out-of-sample predictive performance.

In summary, the Bayesian R-squared will be the basis for measuring the fit of the models to the data, and the PSIS-LOO, an accurate approximation of the LOO cross-validation method, will be the basis for assessing the out-of-sample accuracy of the national FG models.

3.7 Summary Results

In this section, the results of the estimation of the national FG models are presented, followed by the assessment of model fit and out-of-sample prediction accuracy. As a reminder, the dependent variable is the log of the total volume of freight generated, $\log(y_i)$, and the independent variables are the population and GRP of the prefectures, the dummy for the prefecture, the dummy for the year, and the dummy for the goods type. The independent variables population and GRP are standardized to have a mean equal to zero and a standard deviation of one. Fit1 to fit3 is the conventional linear regression model where categorical variables such as prefecture, year, and goods type are incorporated in the model as dummy variables. However, the parameter estimates for the dummy variables of the prefecture, year, and goods are not shown in this section. Please refer to the appendix for the complete estimation results. Fit4 to Fit24 are all varying-intercepts model will have additional parameters estimated. These parameters are the standard deviation (sd) of the varying-intercepts. The sd provides information on how much variation there is among the particular varying intercept variables. Higher sd of varying-intercepts indicates that there is more heterogeneity among the categories in a specific varying-intercept variable.

Table 3.2 Estimated parameters of the national FG models

(mean)	intercept	population	GRP	sd (prefecture)	sd (year)	sd (goods type)
fit1*	14.7	0.77	0.07			
fit2*	14.71	0.82				
fit3*	14.78		0.84			
fit4	14.9	0.96	-0.33	0.45		
fit5	14.91	0.65		0.48		
fit6	14.91		0.54	0.62		
fit7	14.9	1.08	-0.45		0.2	
fit8	14.89	0.67			0.24	
fit9	14.9		0.55		0.18	
fit10	14.77	1.08	-0.45			1.28
fit11	14.77	0.67				1.29
fit12	14.78		0.55			1.29
fit13	14.9	0.99	-0.36	0.45	0.21	
fit14	14.89	0.66		0.48	0.23	
fit15	14.89		0.54	0.62	0.19	
fit16	14.75	0.75	-0.12	0.5		1.29
fit17	14.78	0.64		0.52		1.27
fit18	14.79		0.55	0.65		1.28
fit19	14.76	1.08	-0.45		0.24	1.28
fit20	14.75	0.67			0.22	1.3
fit21	14.77		0.55		0.21	1.28
fit22	14.76	0.86	-0.24	0.49	0.23	1.3
fit23	14.74	0.64		0.51	0.23	1.28
fit24	14.76		0.52	0.65	0.22	1.27

Table 3.2 gives a summary of the estimated parameters of the 24 national FG models. Based on the results, a recurring observation is that, except for the Fit1, all model specification that only has population or GRP as the independent variable will have a positive coefficient. To be specific, all models that only have population as the independent variable, namely, fit2, fit5, fit8, fit11, fit14, fit17, fit20, and fit23, have a positive coefficient ranging from 0.64 to 0.82. Similarly, all models that only have GRP as the independent variable, namely, fit3, fit6, fit9, fit12, fit15, fit18, fit21, and fit24, have a positive coefficient ranging from 0.52 to 0.84. The results of these two types of models give reasonable results, especially regarding the signs of the coefficients. As the population increases, the volume of freight generated also increases on the order of 0.64 to 0.82 log tons for each 1 standard deviation increase in population. On a similar note, as the GRP of a prefecture increase, the volume of freight generated also increases on the order of 0.52 to 0.84 log tons for each 1 standard deviation increase in GRP. However, with the exception of fit1, looking at the FG models with both population and GRP as independent variables in their linear equations, while the sign of the coefficients for population remains positive, the sign of the coefficient for GRP becomes negative. A negative coefficient for GRP is contrary to the expected influence of GRP on the total freight volume generated. A negative coefficient for GRP implies that holding constant the population, as the GRP of a prefecture increase, the total volume of freight generated will decrease. To be specific, the models with both population and GRP as independent variables are fit4, fit7, fit10, fit13, fit16, fit19, and fit22.

Table 3.3 Bayesian R-squared

	Estimate	Est. Error	Q2.5	Q97.5
fit1	0.8258	0.0036	0.8185	0.8324
fit2	0.8258	0.0035	0.8184	0.8323
fit3	0.8254	0.0036	0.8180	0.8321
fit24	0.8251	0.0037	0.8174	0.8320
fit23	0.8250	0.0037	0.8174	0.8318
fit22	0.8246	0.0037	0.8171	0.8315
fit18	0.8206	0.0037	0.8129	0.8275
fit17	0.8200	0.0038	0.8120	0.8270
fit16	0.8198	0.0038	0.8118	0.8268
fit19	0.7259	0.0064	0.7126	0.7376
fit10	0.7208	0.0065	0.7075	0.7330
fit20	0.7113	0.0068	0.6974	0.7241
fit11	0.7066	0.0070	0.6922	0.7193
fit21	0.6424	0.0089	0.6240	0.6587
fit12	0.6385	0.0090	0.6204	0.6557

	Estimate	Est. Error	Q2.5	Q97.5
fit15	0.3108	0.0170	0.2766	0.3428
fit14	0.3105	0.0170	0.2770	0.3431
fit13	0.3104	0.0170	0.2762	0.3433
fit6	0.3069	0.0171	0.2726	0.3394
fit4	0.3060	0.0170	0.2724	0.3384
fit5	0.3059	0.0175	0.2711	0.3393
fit7	0.2266	0.0168	0.1930	0.2591
fit8	0.2127	0.0161	0.1813	0.2444
fit9	0.1443	0.0155	0.1145	0.1751

Table 3.3 shows the tabulated Bayesian R-squared for all the fitted models. It can be observed that Fit1 to fit3, corresponding to the classical regression model formulations with the highest number of variables in the model, had the top 3 highest R-squared among all the 24 fitted models. Fit1, fit2, and fit3 all consist of intercepts or dummy variables for each factor in the variables prefecture, year, and goods. Fit1 considers both population and GRP as independent variables in the model, while fit2 only considers population, and fit3 only considers GRP. The implication of having the highest Bayesian R-squared among the fitted models is that the classical linear regression models describe the national freight volume generation data best in terms of goodness-of-fit. This is expected as among the 24 fitted models, fit1, fit2, and fit3 have the highest number of independent variables in their respective linear models. Following the top three classical linear regression models are the fit24, fit23, and fit22, which are model fits with varying-intercepts for the prefecture, year, and goods type. While fit22 includes all categorical variables as varying-intercepts, namely, the prefecture, year, and goods type, and both the population and GRP as independent variables, fit24 and fit23, which both also considers all categorical variables as varying-intercepts, but only consider GRP and population, respectively, had a higher Bayesian R-squared than fit22. However, as we will see in the following table, which summarizes the LOO cross-validation as approximated by the $\widehat{lppd}_{psis-loo}$ by taking the differences in the estimated $\widehat{lppd}_{psis-loo}$ of each of the model fits from the highest $\widehat{lppd}_{psis-loo}$, the model with the highest goodness-of-fit to the data, i.e., the model with the highest Bayesian R-squared, is not necessarily the best model for future predictions.

Table 3.4 lppd PSIS-LOO differences

Model comparisons:		
	elpd_diff	se_diff
fit24	0	0
fit23	-0.6	2.5

Model comparisons:		
	elpd_diff	se_diff
fit22	-1.5	3.4
fit3	-1.8	1.7
fit2	-2.2	3
fit1	-2.4	2.7
fit18	-17.1	6.1
fit17	-19.5	7
fit16	-20.8	7.2
fit19	-316.5	22.3
fit10	-328.7	22.6
fit20	-355	23.6
fit11	-365.4	23.9
fit21	-516.4	28
fit12	-522.8	28.1
fit14	-1025.4	31.9
fit13	-1025.6	31.9
fit15	-1026.4	32
fit5	-1027.7	32.2
fit4	-1028.5	32.2
fit6	-1028.5	32.2
fit7	-1092.4	32.6
fit8	-1105.6	32.4
fit9	-1168.7	32.8

Table 3.4 shows the differences in LOO cross-validation as approximated by the $\widehat{lppd}_{psis-loo}$ that was defined in the previous section. Recall that a model with a higher $\widehat{lppd}_{psis-loo}$ relative to other models indicates that the model will perform better in predicting out-of-sample data. In our context, the linear model with the highest $\widehat{lppd}_{psis-loo}$ will be the best predictor of national freight volume generation. Based on Table 3.4, fit24 has the highest $\widehat{lppd}_{psis-loo}$ as shown by the zero elpd_diff; the difference of the highest $\widehat{lppd}_{psis-loo}$ with the highest $\widehat{lppd}_{psis-loo}$ (itself) is zero. Fit24 corresponds to the varying-intercepts model, with the prefecture, year, goods type as varying-intercepts, an only GRP as the independent variable as follows:

Fit24: $\log(y_i) = \alpha + \alpha_{pref} + \alpha_{year} + \alpha_{goods} + \beta_{GRP}x_{grp}$

Recall that the result of the Bayesian R-squared for the estimated models indicates that fit1 is the best model for national freight with a linear model formulation as follows:

Fit1: $\log(y_i) = \alpha + \alpha_{pref1} + \cdots + \alpha_{pref46} + \alpha_{year1} + \cdots + \alpha_{year3} + \alpha_{goods1} \dots \alpha_{goods7} + \beta_{pop}x_{pop} + \beta_{GRP}x_{grp}$

Fit1 is a classical linear regression formulation with dummy variables and continuous variables as independent variables with a total of 59 variables. This is in contrast to the best model for the out-of-sample prediction that is the varying-intercept model fit24, which only has 5 variables in its linear model. As seen in the different best model between the Bayesian R-squared and the LOO cross-validation as approximated by $\widehat{lppd}_{psis-loo}$, the overfitting that has resulted from including 59 independent variables did not produce a model that is best for predicting future national freight volume generation. Furthermore, with only 5 predictor variables, in contrast to 59 independent variables, a model that performs better for out-of-sample predictions of national freight volume generation was estimated under a varying-intercepts model construction.

3.8 Summary and Conclusions

In this chapter, models for the national freight volume generation in Japan were estimated. The national freight volume data used has a total of 1,504 samples with 6 variables, namely, year, prefecture, the population (in 1,000s), gross regional product (GRP), type of goods, and weight in tons of goods generated. Specifically, the classical linear regression model and the varying-intercept model were compared. There was a total of 24 model specification used to estimate the national freight volume generation of Japan. The estimation of the classical linear regression model and the varying-intercept model were conducted under a Bayesian Inference framework because it can easily allow the estimation of complex models, such as those with varying-intercepts. The objective of this chapter is to demonstrate that a national freight generation model with a higher goodness-of-fit is not necessarily the best model for forecasting future freight volume generation.

The most common measure of goodness-of-fit for linear regression models is the R-squared, which explains the amount of variation in the data explained by the model, hence a high R-squared is desirable. However, in the Bayesian Inference setting of estimation, the traditional R-squared formula may be problematic because it can be greater than 1, whereas the estimated R-squared should be between 0 and 1 for it to be meaningful. Thus, a Bayesian version of R-squared was used to evaluate the goodness-of-fit of the estimated models. However, while the fit of a model to the data is measured by the R-squared, it does not provide any information on a model's performance to predict future values from future data. In the context of the national freight volume generation in Japan, the R-squared of a model does not say anything the model's performance on predicting future freight volume generation on future population and GRP.

A common and established method of estimating the out-of-sample performance of a model is to conduct cross-validation. There are multiple ways to conduct cross-validation. For the national freight volume generation models, the Leave-One-Out (LOO) cross-validation is used to evaluate the out-of-sample performance of the models. The LOO cross-validation method leaves-out one sample and estimates the model using the rest of the samples in the data. For example, in the national freight volume generation data of Japan with 1,504 samples, 1,503 samples are used to estimate the model and evaluated on the left-out sample regarding its predictive accuracy. The process is repeated a number of times equal to the total number of samples, and the results are averaged to get the average out-of-sample performance. However, there is a total of 24 national freight volume generation model specifications which means that the LOO cross-validation process would need to be done repeatedly for 24 times. This process will require the heavy use of computer resources, especially storage and memory, as well as long computation times. Thus, to evaluate the out-of-sample performance of the national freight volume models estimated in this chapter, a method that could approximate the LOO cross-validation with high accuracy and without the need to conduct the actual LOO cross-validation was used. The method is called the Pareto-Smoothed Importance Sampling (PSIS-LOO) cross-validation, which uses the relative importance of each sample in the data to the posterior distribution. The advantage of the PSIS-LOO cross-validation is that there is no need to repeatedly estimate the 24 models 1,504 times, once for each sample, using a training-test split, and averaging the errors to get the mean-squared-error of each model. The PSIS-LOO cross-validation only needs the posterior distribution of the models to approximate the LOO cross-validation; thus, it requires fewer computer resources and computation time.

The estimation results show that the national freight volume generation models with the most number of independent variables, i.e., the classical regression models, had the highest Bayesian R-squares. This indicates that the classical regression models fit the national freight volume generation of Japan best. This is expected because the classical regression formulation of the national freight volume generation includes the most number of predictor variables. The three national freight volume generation models with the highest Bayesian R-squared are classical regression models with 57 to 59 predictor variables; the variables are the population and GRP as continuous variables, the 47 prefectures as dummy variables, the 4 years as dummy variables, and the goods type as dummy variables. This implies that the classical regression model specification is the best model at describing the national freight volume generation of Japan. However, it was also shown that the models with the best goodness-of-fit to data were not the best at predicting out-of-sample data.

As discussed previously, the PSIS-LOO cross-validation method was conducted to determine the out-of-sample performance of the estimated national freight volume generation models. The results were summarized by tabulating the differences of the PSIS-LOO cross-validation measures from the model with the best PSIS-LOO cross-validation. The results showed that the models that performed best at out-of-sample prediction were the varying-intercepts model while the best models with Bayesian R-squared only following behind. This implies that models that have the best fit to data do not necessarily mean that they are the best for out-of-sample and future prediction. The best models based on the Bayesian R-squared not being the best for predicting future data demonstrates that the classical regression models with 57 to 59 predictor variables overfit the national freight volume generation data of Japan. In contrast, the result of the PSIS-LOO cross-validation shows that out-of-sample prediction is improved and overfitting avoided by specifying a varying-intercepts model for the national freight volume generation of Japan. The national freight volume generation model with the best PSIS-LOO cross-validation is the model with an intercept, which represents the overall mean freight volume generation, the prefecture, year, and goods type as varying intercepts to represent their varying effects and the GRP of the prefecture as a continuous independent variable. There are a total of 5 predictors in the best performing model for out-of-sample prediction, which is significantly lower than the model with the best Bayesian R-squared, which had 59 total predictor variables. Thus, it was shown that for the national freight volume generation of Japan, a varying-intercepts model with only 5 predictor variables outperforms the classical linear regression model with 59 predictor variables in out-of-sample prediction.

The results of the estimations of the different models the national freight volume generation of Japan and the evaluations of their respective Bayesian R-squared and LOO cross-validation as approximated by the PSIS-LOO method implies that a better national freight model that avoids overfitting and, correspondingly, performs better at prediction can be estimated by using a varying-intercepts model specification. It is often the case that applied freight models in general result to overfitting by using a simple model formulation such as the classical linear regression and incorporating as much independent variable. As seen in this chapter, predictive performance can be improved, and overfitting avoided by taking a step further in the model specification by incorporating varying effects in the form of varying intercepts into the linear model.

Chapter 4 LOGISTICS FACILITY ALLOCATION, SIZE, AND FREIGHT TRIP GENERATION OF TOKYO METROPOLITAN AREA

4.1 Introduction

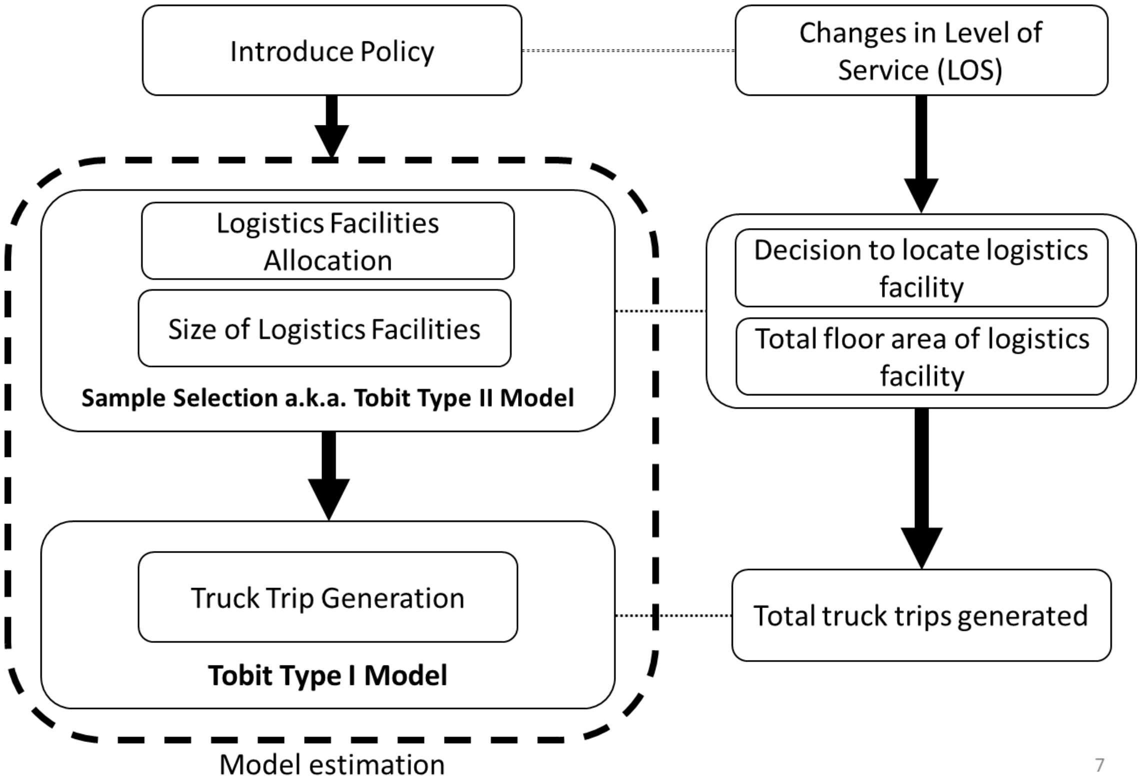
Logistics and freight are seldom topics of research in transportation. This motivates more studies that focus on logistics and freight systems in the context of transportation research and its impacts on the built environment and quality of life of people. Impacts due to generation and attraction of traffic from logistics facilities need to be considered in future transportation plans or city plans. While the relationship between land-use and transportation has been a well-known subject of research for person trip behavior, research that investigates together the relationship of allocation patterns of land-use and elements of logistics and freight networks such as the production of truck trips and location choice for logistics facilities are lacking. This chapter aims to analyze the relationship between logistics facilities and truck trip generation by utilizing the 4th and 5th Tokyo Metropolitan Area Urban Freight Survey (TMAUFS), which were conducted in 2003 and 2013, respectively.

Specifically, we aim to relate land-use allocation and truck trip generation by formulating a Logistics Floor Area model and a Truck Trip Generation model using utility theory with land-use variables and other area characteristics as inputs to both models. Ultimately, the estimated models will be used to conduct sensitivity analysis on the effects of infrastructure and policy changes on the total logistics floor area and truck trip generation.

The structure of this chapter is as follows: in section 4.2, we briefly present the 4th (2003) and 5th (2013) TMAUFS data and describe temporal changes from 2003 to 2013 with respect to the number of logistics facilities; section 4.3 presents the Truck Probe data portion of the 5th TMAUFS to give an overview of truck generation in Tokyo Metropolitan Area (TMA). In section 4.4 and 4.5, we formally develop and estimate the Logistics Floor Area model and the Truck Trip Generation model using TMAUFS data and relate both models together to demonstrate their practical application. Finally, section 4.6 concludes and summarizes this chapter.

4.2 Framework of Analysis

The framework of analysis for the logistics facility allocation, size, and freight trip generation of the Tokyo Metropolitan Area is shown in *Figure 4.1* below.



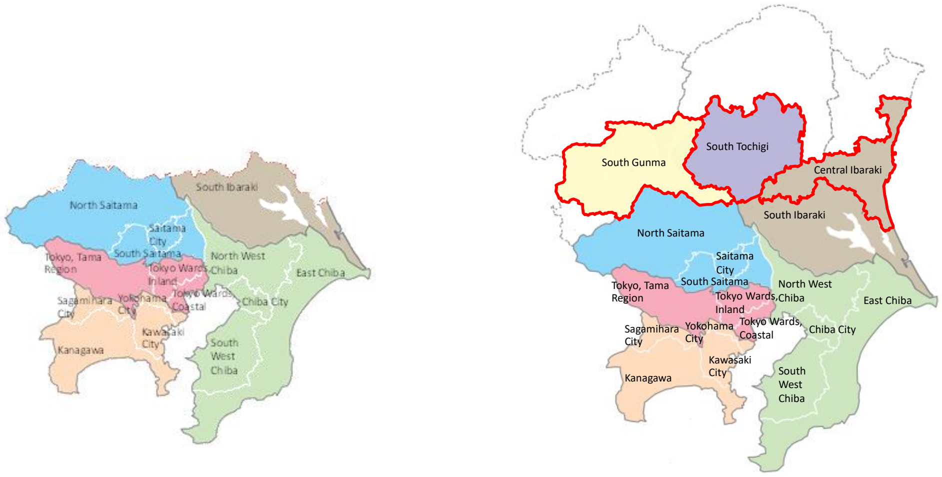
Introduce Policy
Changes in Level of Service (LOS)
Logistics Facilities Allocation
Size of Logistics Facilities
Sample Selection a.k.a. Tobit Type II Model
Decision to locate logistics facility
Total floor area of logistics facility
Truck Trip Generation
Tobit Type I Model
Total truck trips generated
Model estimation
7

Figure 4.1 Framework

4.3 Data Abstract

4.3.1 Basic Analysis using the 4th and 5th survey

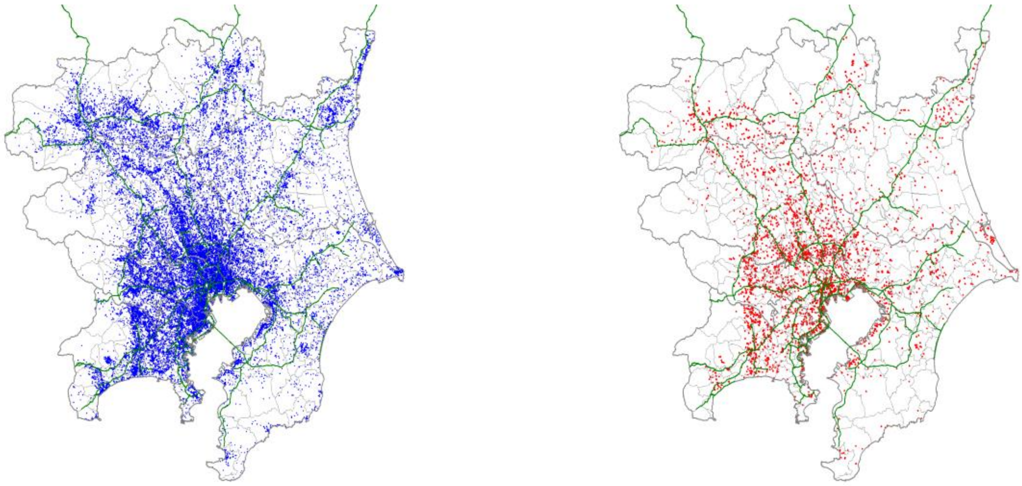
In this section, we briefly discuss the transition of the distribution of freight facilities in TMA from the 4th TMAUFS (2003) to the 5th TMAUFS (2013). The specific areas of analysis considered in the 5th survey are the areas of Tokyo, Kanagawa, Chiba, Saitama, and South Ibaraki so that a comparative analysis between the 4th and 5th survey can be conducted. This is because the survey areas covered in the 5th survey have been extended to South Gunma, South Tochigi, and Central Ibaraki, as shown in Figure 4.2 below.



□: New areas added for 5th survey

Figure 4.2. TMA freight survey areas: (a) 4th TMA freight survey (left), and (b) 5th TMA freight survey (right)

The units of analyses are logistics centers, which is one of the facility types defined in the survey to establishments in TMA. The distribution of all facility types surveyed in TMA and the distribution of all logistics facility respondents in TMA are shown in Figure 4.3 below.

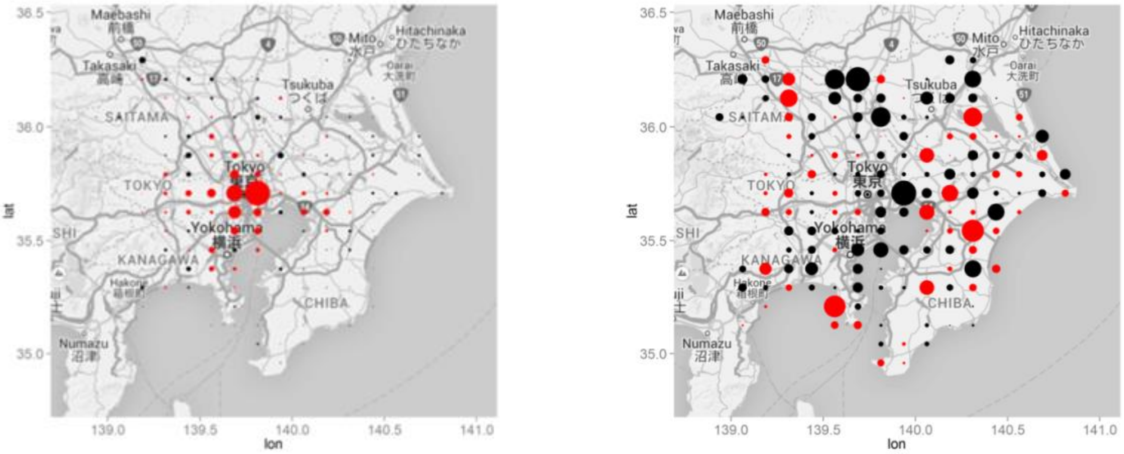


Total respondents: 44,000 Logistics facility respondents: 4,600

Figure 4.3. Distribution of respondents from all facility types (left) and Logistics facilities (right)

The specific variables of analyses are Total Floor Area (m^2), Number of Places with Outbound Trips, and Outbound Tonnage of Freight tallied from each type of establishment to represent the scale of facilities. Further, we focus on analyzing each secondary mesh unit, which is about 10-km^2 due to the 4th survey (Japan Geodetic System) and 5th survey (World Geodetic System) using different geodetic coordinate systems. Due to the difference in coordinate systems used, the longitude and the latitude of the two systems are different by about 460-m. However, we suppose that the 10-km^2 secondary mesh could ease the differences. Furthermore, we define the Number of Logistics Facilities as the sum of each establishment per mesh transformed by an expansion factor. We also define the Total Floor Area of Logistics Facilities, Number of Places with Outbound Trips, and the Outbound Tonnage of Freight as their respective averages considering an expansion factor to reflect relative magnitudes.

Figure 4.4 below shows the increase and decrease in the Number of Logistics Facilities and the Total Floor Area of Logistics Facilities. The relative sizes of the circles indicate the maximum absolute value of their respective percentage changes; black circles indicate an increase, and red circles indicate a decrease from the 4th survey to the 5th survey.



(a) The Number of Logistics Facilities

(b) Total Floor Area of Logistics Facilities

Figure 4.4. Increase and Decrease of (a) Number of Logistics Facilities (left), and (b) Total Floor Area of Logistics Facilities

As seen in Figure 4.4, the Number of Logistics Facilities in Tokyo is observed to have decreased and, by contrast, has increased in the suburbs (e.g., North Saitama, South Ibaraki). We suppose that this is due to the improved network of highways during the 5th TMAUFS relative to when the 4th TMAUFS was conducted (e.g., the completion of the Metropolitan Inter-City Expressway). Furthermore, the Total Floor Area of Logistics Facilities is observed to have increased around Tokyo Bay and the Tohoku Expressway. We suppose that this is due to logistics facilities in Japan moving toward increasing in size due to the consolidation of functions and services, which, as a result, lead to the decrease in small-scale logistics facilities and an increase in large-scale logistics facilities.

4.3.2 A Quantitative Understanding of the Increase and Decrease in the Specific Variables of Analysis using Multiple Regression

In order to analyze factors that might have influenced the increase and decrease of the specific units of analysis, we illustrate the increase and decrease quantitatively using conventional multiple regression analysis. We let the dependent variable (y_n^*) be the difference between the 4th and 5th surveys of each specific variables of analysis, namely the Number of Logistics Facilities, Total Floor Area of Logistics Facilities, Number of Places with Outbound Trips, and Outbound Tonnage of Freight as defined in equation (1) below:

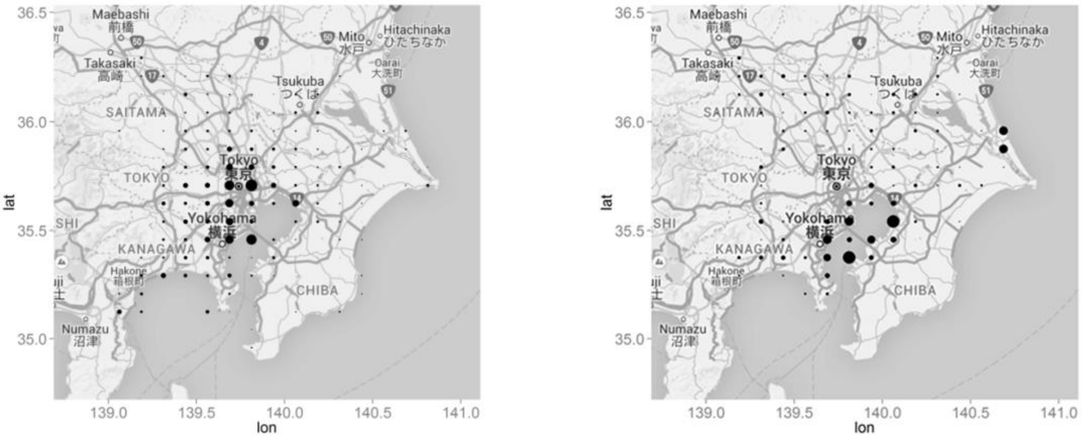
$$y_n^* = \Delta y_n = y_n^{5th} - y_n^{4th} \quad (1)$$

Table 4.1 shows the result of the multiple regression analysis.

Table 4.1. Estimation results of the multiple regression analysis

	Units	Estimate	Std. error	t-value	Pr(> t)
(Intercept)		-80.983	190.292	-0.4	6.71E-01
Residential area	km^2	1.943	1.128	1.7	8.70E-02
Commercial area	km^2	-46.055	3.296	-14.0	4.02E-29
Quasi-Industrial area	km^2	-11.236	3.123	-3.6	4.34E-04
Industrial area	km^2	-4.092	6.943	-0.6	5.57E-01
Restricted-Industrial area	km^2	7.359	1.958	3.8	2.42E-04
Urbanization control areas	km^2	0.059	0.604	0.1	9.23E-01
Extramural city planning areas	km^2	0.444	1.079	0.4	6.81E-01
Uninhabitable areas	km^2	-1.375	1.186	-1.2	2.48E-01
Seaside area	Dummy	-62.266	39.352	-1.6	1.16E-01
Metropolitan inter-city expressway	Dummy	12.177	22.073	0.6	5.82E-01
Inland	Dummy	-102.580	56.702	-1.8	7.24E-02
Suburb	Dummy	-131.451	51.960	-2.5	1.24E-02
Inhabitable land	km^2	-2.228	1.137	-2.0	5.17E-02
ln(Population)	1,000's	20.509	18.191	1.1	2.61E-01
ln(Working population in commuting distance)	1000's	18.753	16.979	1.1	2.71E-01
Distance to expressway IC	km	0.196	1.248	0.2	8.75E-01
Distance to port	km	0.243	0.401	0.6	5.46E-01
ln(Land prices)	1000 yen/km^2	10.841	18.442	0.6	5.58E-01
Distance from Tokyo Station	km	1.332	0.861	1.5	1.24E-01
Adjusted R-squared			0.745		
Number of samples (secondary meshes)			173		

The estimation results for the Number of Logistics Facilities as the dependent variable are convincing as reflected by its adjusted R-squared, which is higher than 0.7. By contrast, the other units of analysis, i.e., Total Floor Area of Logistics Facilities, Number of Places with Outbound Trips, and the Outbound Tonnage of Freight, as the dependent variables were not convincing due to their low adjusted R-squares (~0.2); the results of estimation for the aforementioned units of analysis were not shown here due space limitations. Therefore, we only consider the results of the estimation for the Number of Logistics Facilities. Figure 4.5 below shows the share of the explanatory variables for (a) Quasi-Industrial areas and (b) Restricted-Industrial areas where larger circles indicate larger shares in the data. Noting the result of the multiple regression analysis, cross-examining Figure 4.4 above and Figure 4.5 below reveals that the Number of Logistics Facilities tends to decrease in TMA, especially in areas where the share of Quasi-Industrial land-use is large. By contrast, in Restricted-Industrial areas, where residential buildings are restricted, the Number of Logistics Facilities is almost unchanged. This might be due to logistics facilities being built in more suitable places other than those in Quasi-Industrial areas.



(a) Quasi-Industrial areas

(b) Restricted-Industrial areas

Figure 4.5. Share of (a) Quasi-Industrial areas, and (b) Restricted-Industrial Areas

4.3.3 Truck Probe Data

In this section, we describe and discuss the truck probe data from the 5th TMAUFS in terms of truck trip generation in TMA. Among three (3) sources of truck probe data from the 5th TMAUFS, we used the data collected for one (1) week, from October 6th (Monday) to October 12th (Sunday), by an OBU (On-Board Unit) manufacturer because it has the most significant number of samples (22,995 samples), and use these data in estimating Truck Trip Generation models in Section 4. The truck trip generation data is categorized into four (4) levels: small, medium, large, and tractor; each level is categorized based on the maximum gross weight in tons. The truck trip generation data is an aggregation of truck trips generated at the tertiary level mesh; one data point of truck trip generation represents truck trips generated per 1-km^2 area (tertiary level mesh).

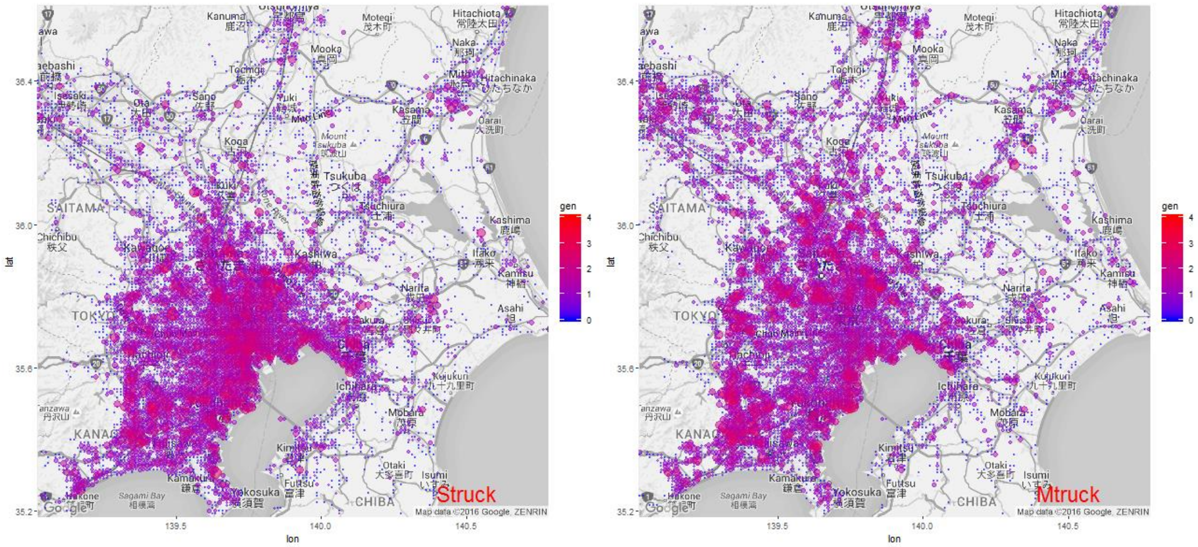


Figure 4.6. Common logarithm of one-week truck trip generation in Tokyo Metropolitan Area of (a) Small Trucks (left), and (b) Medium Trucks (right)

Figure 4.6 above shows the common logarithm[1] of the one-week truck trip generation of small trucks and medium trucks. The truck trip generation of small trucks (left) shows that, generally, small truck trips are concentrated in the center of TMA. This is due to small trucks primarily serving or operating for businesses in the city center as well as residents living in the city center and around Central Business Districts (CBDs). On the other hand, the truck trip generation of medium trucks (right) is more spread-out across TMA relative to small trucks, especially in the suburbs, north-west to north-east of TMA. The spreading-out of medium truck generation could primarily be explained by the inter-logistics facility movement of freight wherein freight is temporarily stored in an intermediary location before being hauled to its destination outside or inside TMA.

[1] $\log_{10}(x) = y$

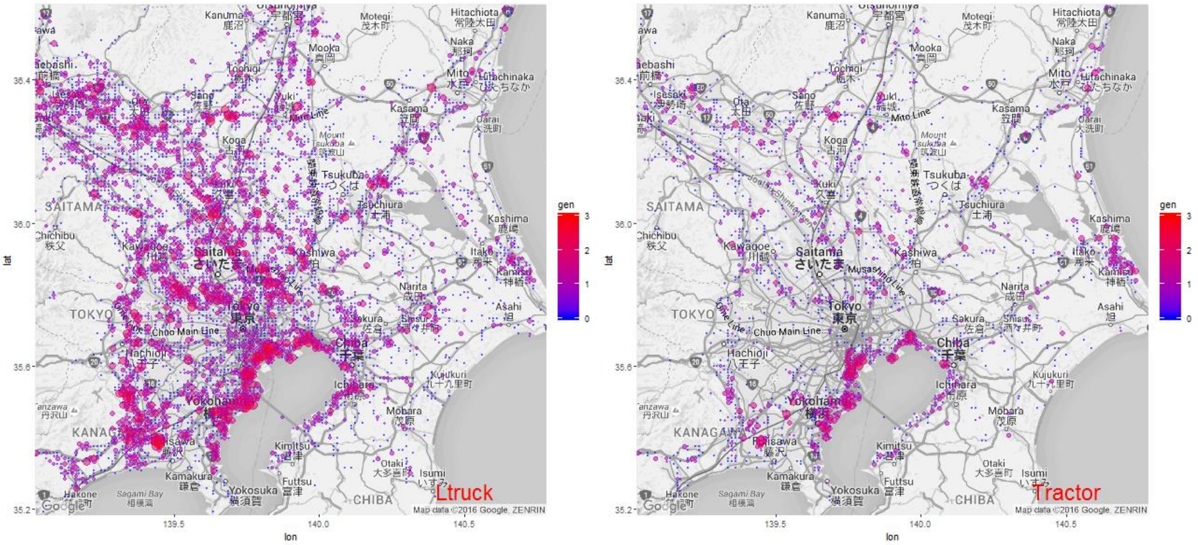


Figure 4.7. Common logarithm of one-week truck trip generation in Tokyo Metropolitan Area of (a) Large Trucks (left), and (b) Tractor Trucks (right)

Figure 4.7 shows the common logarithm of the one-week truck trip generation of large trucks and tractor trucks. It can be observed that for large trucks (left) and tractor trucks (right), there are large concentrations of trip generation along Tokyo Bay. This is expected because the Port of Tokyo and the Port of Yokohama are located along Tokyo Bay as well as logistics facilities servicing these ports. These ports cater to international shipping and container ships, and as such, their operations generate weak volumes of truck traffic inbound and outbound of Tokyo Bay. Furthermore, a large concentration of tractor trucks trips generation (right) can be observed in the eastern region of TMA, specifically in the areas of Kashima City and Kamisu City. This is due to Kashima City and Kamisu City being part of the Kashima Rinkai Industrial Zone, where about 1,500 factories of chemical, petrochemical, specialty chemical plants, steel, and oil refineries are located. Accompanying the Kashima Rinkai Industrial Zone is the Port of Kashima, which further contributes to tractor trucks trip generation from the eastern region of TMA due to inbound and outbound international shipping containers.

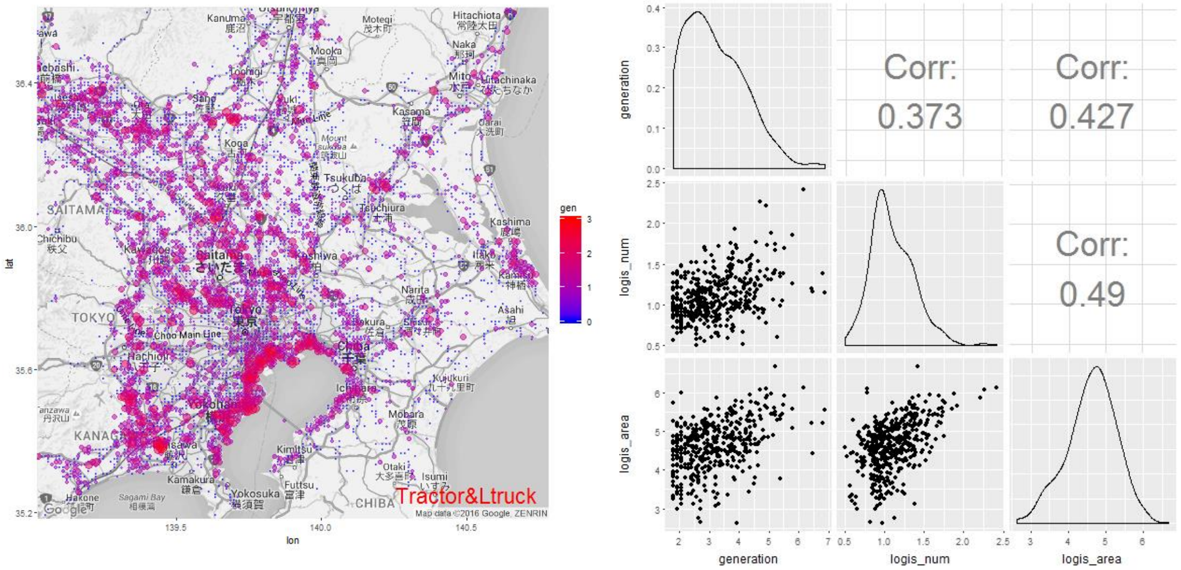


Figure 4.8. (a) Aggregation of Large Trucks and Tractor Trucks’ one-week truck trip generation in Tokyo Metropolitan Area; (b) Scatterplot of truck trip generation against the number of logistics facilities and total logistics facility floor area

Considering the size and nature of freight being transported by large trucks and tractor trucks, we further aggregate their one-week truck trip generation by combining their respective one-week trip generation. Figure 4.8 shows the combined one-week truck trip generation of large trucks and tractor trucks. We observe that large trucks and tractor trucks trip generation are concentrated around Tokyo Bay, where the Ports of Tokyo and Yokohama are located as well as within proximity of ring roads (circumferential highways) and radial roads in the suburbs where numerous logistics facilities, warehouses, and factories are located. It is important to note that there are residential areas located in high-concentration trip generation areas of large trucks and tractor trucks, especially along ring roads and radial roads; this exposes residents to safety risks. Furthermore, the high gross maximum weight of large trucks and tractor trucks exacerbates the deterioration and accelerate the wear-and-tear of roads, which increase road maintenance costs. Based on the scatterplot of truck trip generation against logistics facility count and total floor area in Figure 4.8 (b) above, we also observed a positive correlation between the trip generation of the tractor and large trucks with the number of logistics facilities and the total floor area of logistics facilities in an area. Given the observed positive correlation and the safety risks to residents and accelerating deterioration of roads caused by large trucks and tractor trucks, this chapter will specifically focus on the trip generation of large trucks and tractor trucks in analyzing the dynamics of logistics facility allocation and size, land-use, and truck trip generation in Section 4.

4.4 Logistics Facility Floor Area and Trip Generation Analysis

4.4.1 Logistics facility floor area model

As observed in the scatterplot in Figure 4.8 (b), we focus on the relationship of truck trip generation and logistics facility count and total floor area. Given that the total logistics facility floor area has a slightly higher correlation to truck trip generation than logistics facility counts, we deal with the former in developing the model, namely the Logistics Facility Floor Area (LFFA) model. In this section, we first develop an LFFA model to analyze factors that affect the total floor area of logistics facilities in an area, specifically in a 1-km^2 mesh, before proceeding to Truck Trip Generation model formulation, which will be discussed in Section 4.2.

A conventional multiple regression model the for the LFFA model, where we set the dependent variable (y^*) as the natural logarithm of the total logistics facility floor area scaled to size (Log.areaE) plus one, [(y^* = ln (Log.areaE + 1)], with the results shown in Table 4.2. However, due to the peculiarity of the data where numerous zero values were observed in the dependent variable, the estimated multiple regression model proved to be not a good fit for the data based on its low Adjusted R-squared (0.2744). Furthermore, using the estimation results in Table 4.2 for the prediction of Total LFFA results in underestimated values, which are due to many zero values in the data.

Table 4.2. Estimation result of conventional multiple regression model

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	-14.091	3.478	-4.05	5.3E-05
Population	-0.261	0.032	-8.10	9.2E-16
Working Population	0.632	0.237	2.67	7.6E-03
ACC.manuf	0.052	0.013	3.94	8.6E-05
ACC.cbd	1.019	0.109	9.36	< 2e-16
ICdistance	-0.030	0.020	-1.50	1.3E-01
TokyoPortDist	-0.018	0.004	-4.17	3.1E-05
landprice	-0.103	0.160	-0.64	5.2E-01
residence share	-0.178	0.493	-0.36	7.2E-01
commerical share	-2.253	1.146	-1.97	4.9E-02
quasiIndustrial share	5.096	0.760	6.70	2.6E-11
industrial share	1.905	1.003	1.90	5.8E-02
restricted industrial share	2.162	0.588	3.68	2.4E-04
road area	3.467	2.555	1.36	1.8E-01

vacant area	-1.224	0.828	-1.48	1.4E-01
Adjusted R-squared	0.2744			
Number of samples	2150			

Because of the underestimation due to many zero values in the data, there is a need to effectively select data points that will be included in the estimation. We do this by applying the Sample Selection model (Heckman, 1979), also known as the Tobit Type II model (Amemiya, 1984), which will be discussed and estimated in this section. The Tobit Type II model construction is as follows:

$$y_i^{S^*} = \boldsymbol{\beta}^S \boldsymbol{x}_i^S + \varepsilon_i^S \qquad \text{(Selection equation)} \tag{2}$$

$$y_i^S = \begin{cases} 0 \; if \; y_i^{S^*} < 0 \\ 1 \; if \; y_i^{S^*} \geq 0 \end{cases}, \tag{3}$$

where the dependent variable $y_i^{S^*}$is the untransformed value of the total number of logistics facilities (Log.num) in the 1-km^2 mesh in the observed data;

$$y_i^{O^*} = \boldsymbol{\beta}^O \boldsymbol{x}_i^O + \varepsilon_i^O \qquad \text{(Outcome equation)} \tag{4}$$

$$y_i^O = \begin{cases} 0 \quad if \; y_i^S = 0 \\ y_i^{O^*} \; if \; y_i^S = 1 \end{cases}, \tag{5}$$

where $y_i^{O^*}$is the natural logarithm of the total logistics floor area scaled to size plus one, [($y_i^{O^*}$ = ln (Log.areaE + 1)]; and

$$\begin{pmatrix} \varepsilon_i^S \\ \varepsilon_i^O \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & \sigma^2 \end{pmatrix} \right), \tag{6}$$

where equations (2) and (4) are the Selection and Outcome equations, respectively. We only observe the value of the latent outcome $y_i^{O^*}$ in equation (4) only if the latent selection variable $y_i^{S^*}$ is positive as described in the conditions in equations (3) and (5). Furthermore, it is assumed that the error terms follow a bivariate normal distribution, as shown in equation (6) above. The Maximum-Likelihood (ML) method is used to estimate the Sample Selection model. Equation (7) shows the likelihood function to be maximized as follows:

$$L^* = \prod_i \boldsymbol{\Phi}\left(-\boldsymbol{\beta}^S \boldsymbol{x}_i^S\right)^{1-y_i^S} \left\{ \boldsymbol{\Phi}\left(\frac{\boldsymbol{\beta}^S \boldsymbol{x}_i^S + \frac{\rho}{\sigma}\left(y_i^{O^*} - \boldsymbol{\beta}^O \boldsymbol{x}_i^O\right)}{\sqrt{1-\rho^2}} \right) \cdot \phi\left(y_i^{O^*} - \boldsymbol{\beta}^O \boldsymbol{x}_i^O\right) \right\}^{y_i^S}, \tag{7}$$

and the expected value of the outcome is shown in equation (8) as follows:

$$E\left[y_i^O | y_i^{S^*} \geq 0\right] = \boldsymbol{\beta}^O \boldsymbol{x}_i^O + \rho\sigma \frac{\phi\left(\boldsymbol{\beta}^S \boldsymbol{x}_i^S\right)}{\boldsymbol{\Phi}\left(\boldsymbol{\beta}^S \boldsymbol{x}_i^S\right)} \tag{8}$$

The input variables for the Sample Selection model are described in

. We emphasize here the inclusion of land-use variables as well as accessibility to manufacturing, CBDs, distance to the closest expressway interchange, and distance to the Port of Tokyo.

Table 4.3. Description of variables used in the Sample Selection model

Variable	**Description**
Population	Population covered by the mesh
Working Population	Working population (daytime) covered by the mesh
ACC.manuf[2]	Accessibility index to manufacturing sites
ACC.cbd[3]	Accessibility index to CBDs
ICdistance.km	Distance of the mesh to the closest interchange
TokyoPortDis.km	Distance to the Port of Tokyo
landprice.yen	Average land price in the mesh
roadArea.m2	Total road area in the mesh
vacantArea.m2	Total vacant area in the mesh
residence.rate	Share of residence land-use
commercial.rate	Share of commercial land-use
quasiIndustrial.rate	Share of quasi-industrial land-use
industrial.rate	Share of industrial land-use
r.industrial.rate	Share of restricted industrial land-use

[2] ACC.manuf$_i$ = $\sum_{j=1}^{J} M_j e^{-\ln(d_{ij})}$, where M_j is the total value of industrial products in area j and d_{ij} is the shortest distance between area i and area j.

[3] ACC.cbd$_i$ = $\sum_{j=1}^{J} B_j e^{-0.5\ln(d_{ij})}$, where B_j is the total working population in area j and d_{ij} is the shortest distance between area i and area j.

We utilize the Sample Selection package in the R programming language (Toomet and Henningsen, 2008) to estimate the model. Table 4.4 below shows the results of the Sample Selection model estimation for the LFFA model. Two other variations of the Sample Selection models were estimated, first of which include all variables in both the selection and outcome equations and the second, which is like Model 1 only that it includes all variables in the outcome equation. However, the estimation results showed parameter estimates that do not satisfy the conditions for utility maximization; thus, only Model 1 and Model 2 shown in Table 4.4 are further discussed in this chapter.

We highlight from the estimation results in Table 4.4 that in the selection equation results, accessibility to manufacturing areas (ACC.manuf), accessibility to CBDs (ACC.cbd), and distance to the Port of Tokyo are statistically significant and are consistent with utility maximization based on their parameter signs. This makes sense because logistics operations managers will strategically want to locate in areas where they will be able to minimize truck operating costs. Population is also statistically significant and satisfies parameter conditions based on its negative sign. This makes sense because when considering the large-scale operations of logistics facilities, developing logistics centers in highly populated areas is not ideal. Taking a look at land-use variables in the outcome equation results, the share of residential land-use (residence.rate) is consistent in sign (negative) and is statistically significant; this is line with population in the selection equation results and is interpreted the same way wherein high residential shares in an area is not ideal for locating logistics facilities more so for developing logistics facilities with large floor areas. Although the other land-use variables seem not significant except for quasi-industrial land-use (quasiIndustrial.rate), we left it as it is in the model to consider land-use in the model. The higher statistical significance of quasi-industrial land-use compared to industrial and restricted land-use could probably be due to logistics companies locating and building in more strategically located areas that have more accessibility to CBDs and manufacturing areas as well as areas closer to the Port of Tokyo which happen to be located and mixed with other less desirable structures for logistics facility operations such as residential and commercial structures.

Table 4.4. Result of Sample Selection (Tobit Type II) models

Probit selection equation:	Tobit Type II Model 1					Tobit Type II Model 2				
	Estimate	Std. error	t value	Pr(> t)		Estimate	Std. error	t value	Pr(> t)	
(Intercept)	-5.0562	1.1394	-4.438	9.09E-06	***	-5.626	1.184	-4.750	0.000	***
Population	-0.0787	0.0087	-9.067	< 2e-16	***	-0.078	0.010	-7.877	0.000	***
Working Population	0.1191	0.0731	1.629	0.1034		0.163	0.076	2.147	0.032	*
ACC.manuf	0.0172	0.0038	4.514	6.35E-06	***	0.013	0.004	3.147	0.002	**
ACC.cbd	0.3770	0.0341	11.052	< 2e-16	***	0.326	0.041	7.981	0.000	***
ICdistance.km	-0.0149	0.0068	-2.209	0.0272	*	-0.014	0.007	-2.088	0.037	*
TokyoPortDis.km	-0.0062	0.0014	-4.395	1.11E-05	***	-0.006	0.001	-3.867	0.000	***
Landprice.yen	-0.0391	0.0508	-0.771	0.4408		-0.021	0.052	-0.409	0.683	
residence.rate						0.225	0.153	1.468	0.142	
commercial.rate						-0.445	0.388	-1.147	0.251	
quasiIndustrial.rate						1.237	0.258	4.801	0.000	***
industrial.rate						0.478	0.328	1.455	0.146	
r.industrial.rate						0.529	0.197	2.681	0.007	**
roadArea.m2	0.6471	0.7594	0.852	0.3941		0.668	0.804	0.831	0.406	
vacantArea.m2	-0.5159	0.2539	-2.032	0.0422	*	-0.290	0.268	-1.080	0.280	
Outcome equation:										
	Estimate	Std. error	t value	Pr(> t)		Estimate	Std. error	t value	Pr(> t)	
(Intercept)	10.795	0.271	39.880	< 2e-16	***	10.913	0.334	32.703	< 2e-16	***
residence.rate	-1.498	0.233	-6.441	1.19E-10	***	-1.603	0.250	-6.419	0.000	***
commercial.rate	-0.149	0.495	-0.301	0.76342		0.160	0.510	0.314	0.754	
quasiIndustrial.rate	1.251	0.404	3.099	0.00194	**	0.601	0.530	1.134	0.257	
industrial.rate	-0.045	0.561	-0.081	0.9356		-0.194	0.633	-0.307	0.759	
r.industrial.rate	0.846	0.353	2.400	0.01641	*	0.594	0.406	1.462	0.144	
Error terms:										
	Estimate	Std. error	t value	Pr(> t)		Estimate	Std. error	t value	Pr(> t)	
sigma	2.009	0.087	23.041	< 2e-16	***	2.016	0.099	20.442	< 2e-16	***
rho	-0.635	0.084	-7.575	3.58E-14	***	-0.642	0.097	-6.582	0.000	***
Log-Likelihood:	-2830.29					-2811.753				
AIC:	5696.58					5669.506				
	2150 observations (1310 censored and 840 observed)									
	18 free parameters (df = 2132)					23 free parameters (df = 2127)				

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

We apply the formula for the expected value of the outcome as in equation (8) above using the estimated parameters in Table 4.4 and the averages of the dependent variables to compute the calibrated average LFFA. Table 4.5 below shows the average calibrated values of LFFA, including the expected values from the conventional multiple regression previously mentioned. Furthermore, we included sensitivity analysis in the form of policy changes and infrastructure improvements such as decreasing the distance to the closest expressway interchange and increasing the share of quasi-industrial, industrial, and restricted industrial land-use.

As shown in Table 4.5, as we decrease the accessibility distance of a 1-km^2 mesh to the closest expressway interchange by 5-km, Model 1 and Model 2 results in an 8% and 7.8% increase in the Logistics Facility Floor Area, respectively. Furthermore, as the share of quasi-industrial, industrial, restricted industrial land-use are increased by 5%, Model 1 and Model 2 result in increases of 19.7% and 13.3%, respectively. This is invaluable information, especially to city planners and road infrastructure managers, because the impacts of such improvements or policy changes are quantified.

We can also observe from the rightmost column of Table 4.5 the expected values from the results of the conventional multiple regression, which clearly underestimate values for the average LFFA (~3-m^2). The difference between the Sample Selection model and the conventional multiple regression model is evident. This is because of the feature of the Sample Selection Model that can better deal with zero values statistically.

Table 4.5. Sensitivity analysis of the Expected Values of the Outcome

Sensitivity Analysis	Tobit Type II				Multiple Regression	
	Model 1		Model 2			
	m^2	% increase	m^2	% increase	m^2	% increase
Average Area of Logistics Facilities	4,922		5,035		2.89	
Distance to closest IC decreased by 5 km	5,317	8.0%	5,426	7.8%	3.29	13.9%
Share of quasi.Ind, Ind, res.Ind increased by 5%	5,891	19.7%	5,704	13.3%	3.61	25.1%

Instead of the average increase in total floor area, we were also able to evaluate the increase in total floor area for each 1-km^2 mesh, this time, considering the completion of all the planned and under-construction ring (loop) roads of the "Three Loop Roads of the National Capital Region" illustrated in Figure 4.9 and visualize the percentage increase in Total Floor Area as shown in Figure 4.10 below. These are reflected as Future Level of Service (LOS), that is, the changes in the variables of accessibility index for manufacturing areas and CBDs (ACC.manuf and ACC.cbd) as well as the distance to the closest expressway interchange (ICdistance.km) when all the ring roads are completed and operational.

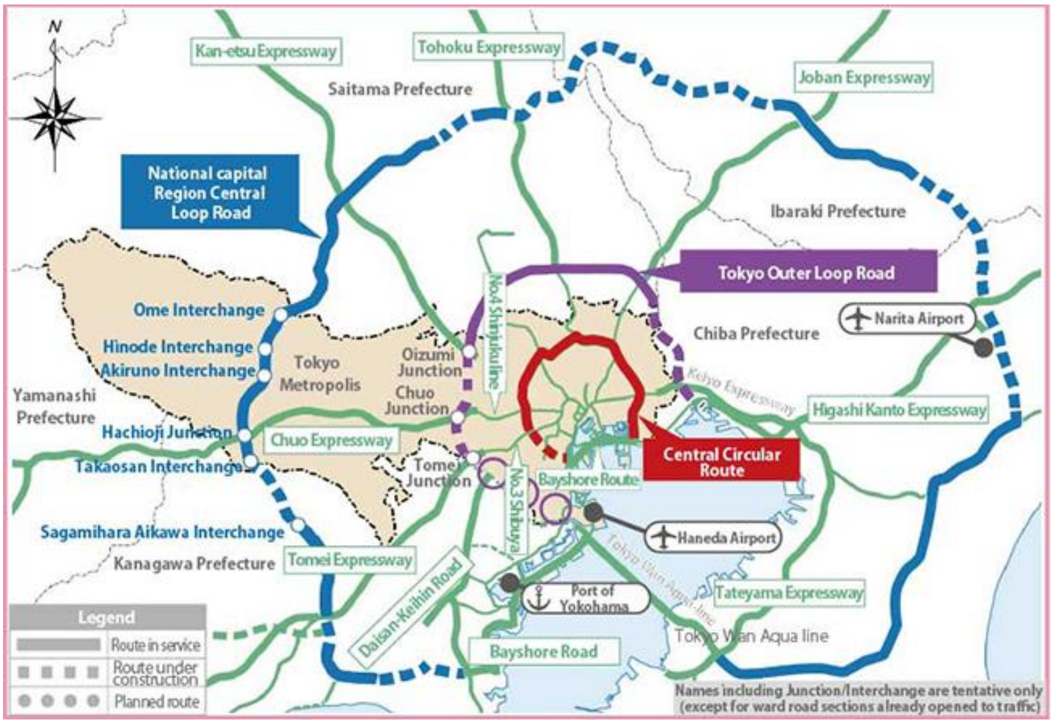
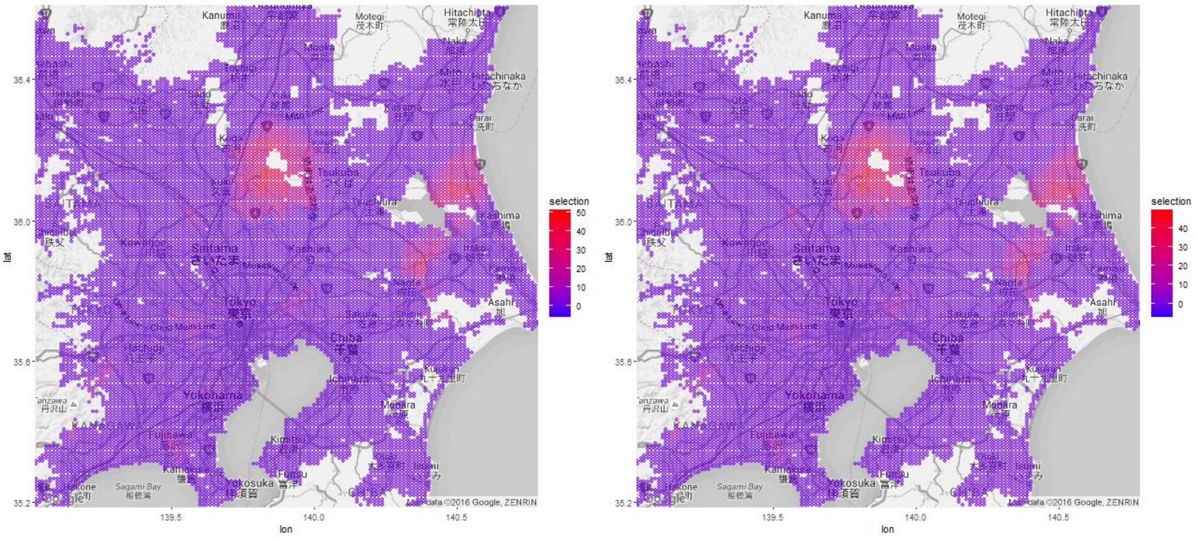


Figure 4.9. Three Loop Roads of the National Capital Region

Source: Tokyo Bureau of Construction



(a) Model 1

(b) Model 2

Figure 4.10. Percentage increase in Total Floor Area

At a glance, we can see that there will be significant increases in Total LFFA in the north and north-eastern region of TMA. The completion of the remaining sections of the Ken-O road[4]

[4] Ken-O road is also known as the Metropolitan Inter-City Expressways/National Capital Region Central Loop Road

(Figure 4.9) in the north to north-eastern regions of the TMA results in the total LFFA to increase in surrounding areas.

Although the percentage increase in total floor area of Model 1 and Model 2 in Figure 4.10 are almost similar, Model 2 can be considered relatively better given its higher Log-Likelihood of -2811.753 compared to the Log-likelihood of Model 1 of -2830.29. Furthermore, the Akaike Information Criterion (AIC), which determines the level of predictive error of the model, is lower for Model 2, hence the better model overall especially for out-of-sample predictions; this can be attributed to the inclusion of land-use variables in the Selection portion Model 2 which contributes to a better-estimated model.

4.5 Truck trip generation model

In this section, we formulate truck trip generation models, specifically for tractor trucks and large trucks. We emphasize in the model formulation the relationship of land-use composition and allocation in an area to its truck trip generation by including the corresponding shares of different land-use classifications in an area as inputs to the truck trip generation model.

Furthermore, we also take into consideration the relationship of the establishment of logistics facilities in an area to its truck trip generation by including as inputs to the model the total number of logistics facilities and the total floor area occupied by logistics facilities in the area.

A conventional multiple linear regression model was initially estimated for truck trip generation with the dependent variable (y^*) as the natural logarithm of the total truck trips generated (y) plus one in a 1-km^2 mesh, [($y^* = \ln(y + 1)$)]. However, the estimation results proved to be a poor fit to the data due to their low adjusted R-squares shown in Table 4.6 below.

Table 4.6. Estimation results of conventional multiple regression analysis

	Exluding Zero Generation				Including Zero Generation			
	Estimate	Std. Error	t-value	Pr(>|t|)	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	2.489	0.315	7.89	3.5E-15	1.319	0.158	8.35	< 2e-16
Log.numE	0.058	0.041	1.42	1.6E-01	0.017	0.030	0.56	5.7E-01
Log.areaE	0.040	0.009	4.36	1.3E-05	0.070	0.007	10.27	< 2e-16
Population	-0.053	0.004	-11.88	< 2e-16	-0.046	0.003	-15.49	< 2e-16
ICdistance.km	-0.060	0.014	-4.37	1.3E-05	-0.118	0.008	-15.56	< 2e-16
portDistance.km	-0.320	0.030	-10.53	< 2e-16	-0.244	0.017	-14.36	< 2e-16
Landprice.yen	0.087	0.019	4.53	6.1E-06	0.118	0.010	12.06	< 2e-16
residence.rate	-0.052	0.059	-0.87	3.8E-01	0.093	0.035	2.62	8.8E-03
commercial.rate	-0.247	0.145	-1.71	8.7E-02	0.172	0.100	1.73	8.4E-02
quasiIndustrial.rate	0.757	0.115	6.59	4.9E-11	1.460	0.084	17.44	< 2e-16
industrial.rate	1.053	0.164	6.43	1.3E-10	2.146	0.128	16.75	< 2e-16
r.industrial.rate	1.456	0.090	16.15	< 2e-16	2.336	0.070	33.62	< 2e-16
Adjusted R-squared	0.2257				0.2938			
Number of samples	5840				18077			

Moreover, the conventional multiple linear regression underestimates predictions of truck trip generation (almost 0 generation). This is due to the numerous zero truck trip generation values in the 1-km^2 mesh data.

To consider the peculiarity in the data, we estimate a more appropriate model, namely the Tobit regression model. Tobit regressions can handle numerous zero generation values in the data better. The Tobit model construction is as follows:

$$y_i^* = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2), \qquad (9)$$

where the latent variable y_i^* is the natural logarithm of the total truck trips generated plus one [(y_i^*= ln(y$_i$ + 1)]; and

$$y_i = \begin{cases} y_i^*, & if\ y_i^* > 0 \\ 0, & if\ y_i^* \leq 0, \end{cases} \qquad (10)$$

where the (k = 9) input variables are the same as in

of the Sample Selection model above. Also, included is an independent and normally distributed error term, ϵ_i, with mean 0 and standard deviation, σ, in the latent formulation. Instead of observing y_i^* directly, which in this case, is truck trip generation (zero or non-zero), we observe y_i as in equation (10) above; we observe the truck trip generation y_i^* if it is positive, and 0, otherwise. It is clearly seen here that the1-km^2 mesh data that have zero generation are appropriately considered in the Tobit model with the "potential" truck trip generated as a function of the input variables. To estimate coefficients β_k, we maximize the likelihood function of equations (9) and (10), as shown in equation (11) below.

$$L = \prod_{i}^{N} \left[\frac{1}{\sigma}\phi\left(\frac{y_i - X_i\beta}{\sigma}\right)\right]^{d_i} \left[1 - \mathbf{\Phi}\left(\frac{X_i\beta}{\sigma}\right)\right]^{1-d_i} \tag{11}$$

Table 4.7 below shows the estimation results of the Tobit regression model for truck trip generation.

Table 4.7. Estimation results of the Tobit Model (Tractor & Large Trucks)

	Estimate	Std. Error	z-value	
(Intercept):1	4.59	0.22	20.64	***
(Intercept):2	0.63	0.01	61.12	***
Log.areaE	0.13	0.01	22.53	***
Population	-0.06	0.01	-8.33	***
ICdistance.km	-0.38	0.02	-18.36	***
portDistance.km	-0.55	0.04	-13.07	***
residence.rate	0.82	0.09	9.25	***
commercial.rate	0.87	0.25	3.53	***
quasiIndustrial.rate	2.79	0.20	13.74	***
industrial.rate	4.21	0.31	13.71	***
r.industrial.rate	4.23	0.17	25.49	***
Log-likelihood:	-17,539.39			
AIC:	35,098.78			
No. of samples	18,077			

Signif. Codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Based on the results of Table 4.7, we can see that all estimates are statistically significant and satisfy the conditions for utility maximization; in other words, the signs (positive or negative) of the estimates/parameters are consistent with the conditions for utility maximization. We specifically note that the estimate for the total land area of logistics facilities (Log.areaE) is highly statistically significant. This is consistent with the findings that historically, in TMA, the number of logistics facilities is generally decreasing, and the average land area of logistics facilities is increasing due to the consolidation of functions and diversification of business operations of logistics facilities. This means that as the total land area of logistics facilities in an area increases, the trip generation of the tractor and large trucks also increases. The increase in truck trip generation, especially for tractor and large trucks, as the total land area of logistics facilities increase can possibly be attributed to the increase in allowable parking space for trucks of all types in the logistics centers. Hence allowing for a larger fleet of trucks and generally a larger scale of logistics operations. Furthermore, as the distance to the closest expressway interchange (ICdistance.km) and the distance to the Port of Tokyo (portDistance.km) decreases, the generation of tractor and large trucks increases; this is consistent with the findings in the LFFA model that the closer an area is to an expressway interchange and to the Port of Tokyo, the higher the likelihood that a logistics facility will locate in that area.

All land-use classifications are statistically significant and positively affect the trip generation of the tractor and large trucks. It is not surprising that the three types of industrial land-use (i.e., quasi, industrial, and restricted) increase the truck trip generation of the tractor and large trucks because these are areas where logistics facilities are built and are being operated. On the other hand, although the estimate for commercial land-use is positive and statistically significant for tractor and large trucks, the magnitude is relatively low compared to the three industrial land-uses. This is probably due to narrow roads and the presence of many pedestrians and shoppers as well as private cars in commercial areas, making it difficult for tractor and large trucks to maneuver, especially during time-constrained operations. However, we note that the share of residence land-use is statistically significant and positively affects the trip generation of the tractor and large trucks. This is most likely due to the prevalence of logistics facilities, factories, and warehouses in the suburbs where a lot of residential areas are also located. Furthermore, the existence of mixed land-use patterns (e.g., quasi-industrial land-use) wherein residential and industrial structures are mixed in an area might contribute to the significance of the residential share to the trip generation of the tractor and larger trucks.

Finally, in order to link the results of the LFFA model to the Truck Trip Generation model, we evaluate the average truck trip generation per day from the calibrated average LFFA in Section 4.1 and the estimated parameter for total logistics facility floor area (0.13) from the Truck Trip Generation model in Table 4.7. The equation for the average trip generation per day is as follows:

$$trip\ gen\, per\ day = floor\ area\ \times \frac{rate}{no.\ of\ days\ in\ a\ week} \times expansion\ factor, \qquad (12)$$

where:

trip gen per day: is the average Tractor and Large Truck trips generated per day per Logistics Facility with *"floor area"*;

floor area: is the average floor area (in 1,000's m^2) of Logistics Facilities evaluated from the Logistics Facilities Floor Area model in Section 4.1;

rate: is the parameter estimate for the Logistics Floor Area (Log.areaE) variable in the Truck Trip Generation Model;

expansion factor: 1/(share of Tractor and Large Trucks) in the data = 1/0.017;

no. of days in a week: the total number of days the data was collected in a week (7 days)

Table 4.8. The sensitivity of Average Truck Trip Generation per Day

	Tobit Type II						Multiple Regression		
	Model 1			Model 2					
	m^2	Average Truck Generation per day	% increase	m^2	Average Truck Generation per day	% increase	m^2	Average Truck Generation per day	% increase
Average Floor Area of Logistics Facilities	4,922	5.34		5,035	5.46		2.89	0.0031	
Distance to closest IC decreased by 5 km	5,317	5.76	8.0%	5,426	5.88	7.8%	3.29	0.0036	13.9%
Share of quasi.Ind, Ind, res.Ind increased by 5%	5,891	6.39	19.7%	5,704	6.19	13.3%	3.61	0.0039	25.1%

Table 4.8 shows the sensitivity analysis of average truck generation per day for a logistics facility with average floor area based on the estimation results of both the LFFA model and the Truck Trip Generation model. Again, we can see that the conventional multiple linear regression analysis gave underestimated results. In contrast, we observe realistic output from the combined LFFA model and Truck Trip Generation model.

4.6 Summary

We were able to show the transition and dynamics of logistics facility development from the 4th TMAUFS (2003) to the 5th TMAUFS (2013), focusing on the total number of logistics facilities and total floor area of logistics facilities in an area. The results of the analysis showed that the total number of logistics facilities is decreasing, especially in central TMA. Furthermore, the total number of logistics facilities decreased in quasi-industrial land-use areas along with little to no changes in restricted-industrial land-use areas. These observations indicate that there might be other suitable areas and factors being considered by logistics managers where logistics facilities would be developed. The decrease in the number of logistics facilities might also be due to the consolidation of functions and services of different logistics facilities, which lead to logistics facilities with larger floor areas. We also presented the Truck Probe data portion of the TMAUFS and observed that there are areas where truck trip generation is concentrated depending on the category of trucks. For instance, small and medium trucks generation are mostly clustered around the center of TMA, where most CBDs are located. On the other hand, large and tractor trucks are mostly concentrated around Tokyo Bay, around ports, as well as along the periphery of expressways in the suburbs. The general difference observed as to where truck trips are generated indicates that the location of logistics facilities might have an influence on where trucks are originating primarily for large and tractor trucks. Hence, we formulated the LFFA model and the Truck Trip Generation model focusing on the tractor and large trucks due to safety risks that they impose on the environment, especially to people and road infrastructure.

We were able to show in the LFFA model that accessibility to CBDs and manufacturing areas increases the probability of logistics facilities locating in an area. Also, decreasing the distance to expressways interchanges and to Port of Tokyo through infrastructure improvements such as the completion of the Ken-O expressway at the western area of TMA, increases the probability that logistics facilities will be located and developed in an area. In terms of land-use allocation in an area, the results showed that the share of residential land-use in a 1-km^2 area is significant in decreasing the total floor area of logistics facilities in that area. The negative effect of the share of residential land-use makes sense, especially for tractor and large trucks considering their relative size with medium and small trucks. This is supported by the negative effect of the population on the probability of logistics facilities being developed in an area. This is because the large volume and surface area of the tractor and large trucks make it more challenging for truck drivers to maneuver in highly populated areas such as in residential areas. Moreover, we were able to estimate the average Total Logistics Floor area from the estimated model as well as to conduct policy sensitivity analysis for infrastructure improvements such as shortening the distance to expressway interchanges and increasing the share of the three types of industrial land-uses.

Regarding the trip generation of tractors and large trucks, the results of the Truck Trip Generation model showed that the total floor area of logistics facilities positively affects the trip generation of tractors and large trucks in an area. Furthermore, the share for all land-use classifications is highly statistically significant and positively affects the trip generation of tractor and large trucks, especially on the share quasi-industrial, industrial, and restricted-industrial land-use. The results of the Truck Trip Generation model also showed that the distance of an area to the closest expressway interchange and to the Port of Tokyo negatively affect tractor and large trucks trip generation; meaning, as areas become further away from an expressway interchange and Port of Tokyo, the fewer the trips for tractor and large trucks generated. These are essential findings, especially for city planners and road managers because the development of logistics facilities, land-use allocation as well as the development of expressways and expressways interchanges will have an impact on the trip generation of tractors and large Trucks.

Finally, we were able to link the LFFA model and Truck Trip Generation model by evaluating truck trips generated using the estimation results of the Truck Trip Generation model and average logistics facility floor area from the LFFA model, as shown in Table 4.8 in Section 4. This chapter contributes to the research of logistics and freight movements by developing a modeling framework that could be used to analyze the effects of land-use policy changes and infrastructure improvements to logistics facility size and truck trip generation by estimating separate models for total floor area of facilities and for truck trip generation and linking them together to forecast travel demand of trucks. Lastly, the modeling framework that we have demonstrated can be replicated in other cities. This is because we have simply applied statistical methods using data from an urban freight survey in modeling logistics land-use location choice and floor area in conjunction with truck trip generation. However, we stress the importance of a well-conducted urban freight survey such as the TMAUFS that include surveys to logistics firms and truck probe data as well as a cohesive database of land-use allocation. These are the keys to implementing statistical models that are not only for theoretical applications but also for practical purposes such as sensitivity analysis of specific policy changes.

Chapter 5 SPARSE REGRESSION AS A METHOD FOR TRIP GENERATION MODELING IN KANTO, JAPAN

5.1 Introduction

The rapid development of information technology has revolutionized methods in collecting and storing data. There are now numerous methods for data collection and storage, which result in datasets that are not only large in sample size but also have many variables. Such datasets are typically referred to as Big Data (Fosso Wamba et al., 2015), and the challenge is to utilize such large datasets and determine which variables are necessary for the objective of the researcher. In this chapter, such data will be dealt with within the context of Truck Trip Generation (TTG) in the Tokyo Metropolitan Area (TMA).

Logistics facilities in the central areas of TMA are decreasing and relocating to the outskirts of TMA (Lidasan et al., 2017). Furthermore, with the completion of the Ken-Ō Expressway, which is the outermost ring-road of TMA, logistics facilities have been relocating in areas near expressway interchanges, increasing the number of logistics areas in the suburbs and outskirts of TMA (Lidasan et al., 2017). This consequently increases TTG in the outskirts of TMA, and it is of interest to investigate areas with noticeable increases in TTG.

5.2 Framework of Analysis

The framework of analysis is as follows: first the two methods for selecting variables in a regression context were compared, namely, the stepwise multiple regression, and a sparse regression method. The stepwise multiple regression method utilizes the R-squared and Akaike Information Criterion (AIC) while the sparse regression utilizes the Mean-Squared-Error (MSE) and k-fold cross-validation to evaluate models. These two measures are essentially different because they measure different things: the R-squared measures model fit to data while the MSE and cross-validation measures out-of-sample performance. Because of the differences in model evaluation of the two models, there needs to be a common measure for them to be comparable. But in the sparse regression context, the R-squared cannot be calculated because it utilizes cross-validation which is mainly for measuring out-of-sample performance. Thus, to compare the two methods, the MSE of the stepwise multiple regression method is also computed and compared to the results of the sparse regression results. Furthermore, a generalization of the sparse regression method is used to calibrate truck trip generation values across years to make them directly comparable. This method is useful for scaling the truck trip generation when there is uncertainly on whether the increase in truck trip generation is due to the actual increase in the number of trucks or from the increase in the number of tacographs, equipment that measures various data regarding movement of truck.

5.3 Data Abstract and Methodology

Truck probe data was acquired through a digital tachometer manufacturer in Japan. Digital tachometers installed in trucks collected data about its movement, i.e., they recorded the year, month, and time of departure from origin and arrival to the destination as well as the total travel time (in seconds) and total travel distance. It also recorded whether the truck is moving or at a stop during the data gathering period. The data was collected every day for three months (April to June) each year from 2015 to 2017.

Economic census data, land-use distribution data, and accessibility data of TMA were also acquired. The economic census data survey items can be classified into four major categories, namely, number of establishments, the number of establishments by employee size, total number of workers, and the number of companies by capital class. These four survey items are disaggregated further to different industries and business classifications. The land-use distribution data contains shares of different land-use classification while the accessibility data contains accessibility measures indices such as accessibility to CBDs, accessibility to commercial areas, accessibility to manufacturing areas, and similar indices to other points-of-interests.

The acquired truck probe data was then combined with the economic census data, land-use distribution data, and accessibility data of TMA, all of which are stored in the tertiary mesh/grid[5] units based on Japanese standards. The final dataset has 521 independent variables, 495 of which are from the economic census data with the other 26 from the land-use distribution and accessibility data. The total number of samples is 13,822.

[5] Tertiary mesh/grid unit refers to a 1-km^2 spatial unit; the truck probe data is organized based on this spatial reference

The truck trip generation modeling in this chapter focuses on extra-large and large trucks based on Japanese standards for truck sizes because previous studies showed that the distribution of TTG of these two truck sizes is similar in TMA. Figure 1 below shows the 2017 distribution of TTG in TMA for extra-large trucks and large trucks.

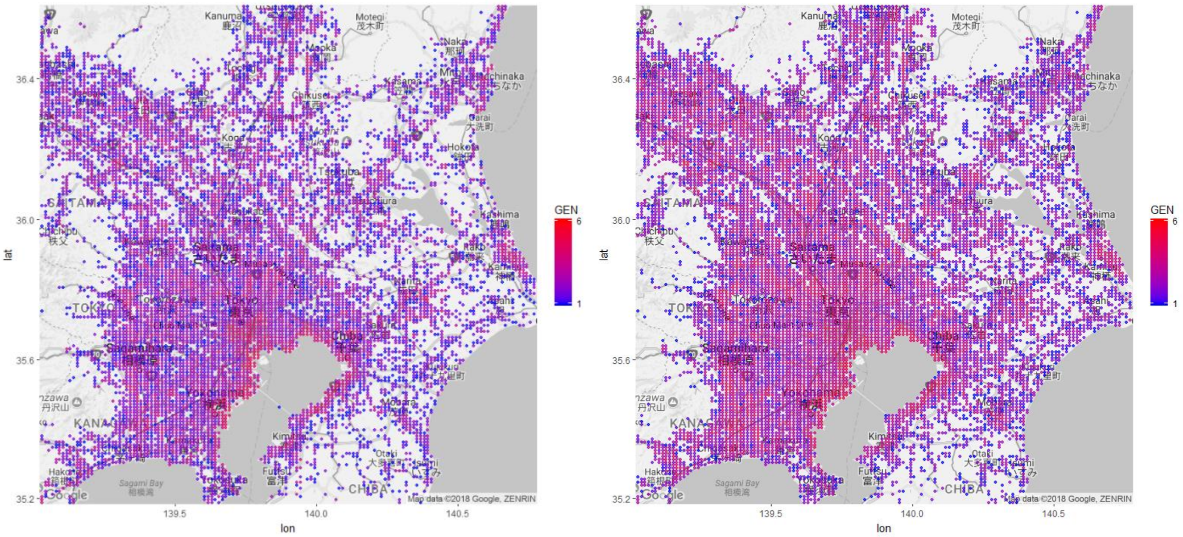


Figure 5.1 2017 Distribution of Truck Trip Generation in TMA for Extra-large trucks (left) and Large trucks (right)

5.4 Model Estimation of Truck Trip Generation in the Tokyo Metropolitan Area

The final dataset has 521 independent variables, with a total of 13,822 samples. Given this large dataset, an efficient statistical model is required to analyze the TTG and select the independent variables that have predictive value. Two statistical modeling frameworks were used to estimate TTG, namely Stepwise Multiple Regression and Sparse Regression. Stepwise Multiple Regression removes (backward elimination/selection) independent variables along the estimation process starting from a model that includes all independent variables and ends up with a final model that maximizes the log-likelihood with the minimum AIC. On the other hand, Sparse Regression utilizes the Lasso penalty in the objective function and estimates a regression model where some of the coefficients of the independent variables are zero in the final model, indicating that they do not have any influence on the dependent variable (Hastie et al., 2015). Furthermore, as an extension of the Sparse Regression framework used in this chapter, the application of Fused Lasso is explored, which is a generalization of the Lasso penalty. The advantage of the two modeling frameworks is that the final estimated model will only contain the relevant independent variables.

5.4.1 Stepwise Multiple Regression

The stepwise regression model estimated in this chapter utilized the backward elimination/selection procedure. The backward selection model starts with all candidate variables in the model, and at each step, the variable that is the least significant is removed. This process continues until no nonsignificant variables remain (James et al., 2013). The significance level at which variables can be removed from the model can be specified by the researcher. However, the default setting in the R programming environment (R Development Core Team 2018) was used. Variable selection was conducted based on minimizing the Akaike-Information-Criterion (AIC) while maximizing the log-likelihood function. The AIC is defined as AIC $= 2k - 2\ln(\hat{L})$ where k is the total number of independent variables in the linear model and $\hat{L}$ is the final log-likelihood. The model definition for the stepwise multiple regression is follows a log-linear construction as follows:

$$\ln(TTG) = \beta_0 + \sum_{\mathbf{1}}^{\boldsymbol{k}} \beta_k x_k \tag{13}$$

$\ln(TTG)$: natural logarithm of TTG

β_0: intercept

β_k: parameter of the independent variable to be estimated

x_k: independent variable *k*

Due to the limitation in space, only the Multiple R-squared, Adjusted R-squared, and the total number of coefficients is presented.

Table 5.1. Stepwise Multiple Regression

Year	Multiple R-squared	Adjusted R-squared	Total No. of Coefficients
2015	0.3814	0.3655	177
2016	0.3691	0.3556	153
2017	0.3852	0.3734	147

From the results shown in Table 1 above, it is noted that the R-squared values are relatively low and may indicate that the model is not a good fit to the data. Nevertheless, the Stepwise Regression model has successfully reduced the total number of coefficients to significantly half of the total initial number of independent variables.

5.4.2 Sparse Regression (Lasso)

Sparse Regression is a linear regression that utilizes the lasso penalty to regularize the regression coefficients by adding a constraint to the least-squares regression's objective function. Regression coefficients are shrunk to zero, and coefficients shrunk to zero can be interpreted as not having any influence on the dependent variable, i.e., on the TTG. The final model will then effectively have a smaller number of non-zero coefficients in the final model than the total number of independent variables in the dataset. The objective function to be minimized for the Sparse Regression model is as follows (Friedman et al., 2010; Hastie et al., 2015):

$$\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2N}\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right\} \tag{14}$$

N: total number of samples

y_i: Total Truck Trip Generation

x_{ij}: independent variable j of sample i

β_j: coefficients to the estimated

λ: tuning parameter to be estimated by cross-validation

The second term in the objective function is the lasso constraint, which regularizes the coefficients during the estimation process. This constraint is also known as the L_1 norm. The λ to be estimated by cross-validation determines the strength of the lasso constraint. The higher the λ is, the stronger the coefficients are constrained and shrunk to zero, and the lower the λ is, the weaker the constraint is and allows for non-zero coefficients. Estimation of the coefficients of the Sparse Regression model by minimizing equation 2 above was done using the "glmnet' package (Friedman et al., 2010) in R. Figure 2 to 4 below shows the estimation path for the sparse regression model and its corresponding cross-validation plot for years 2015 to 2017, respectively.

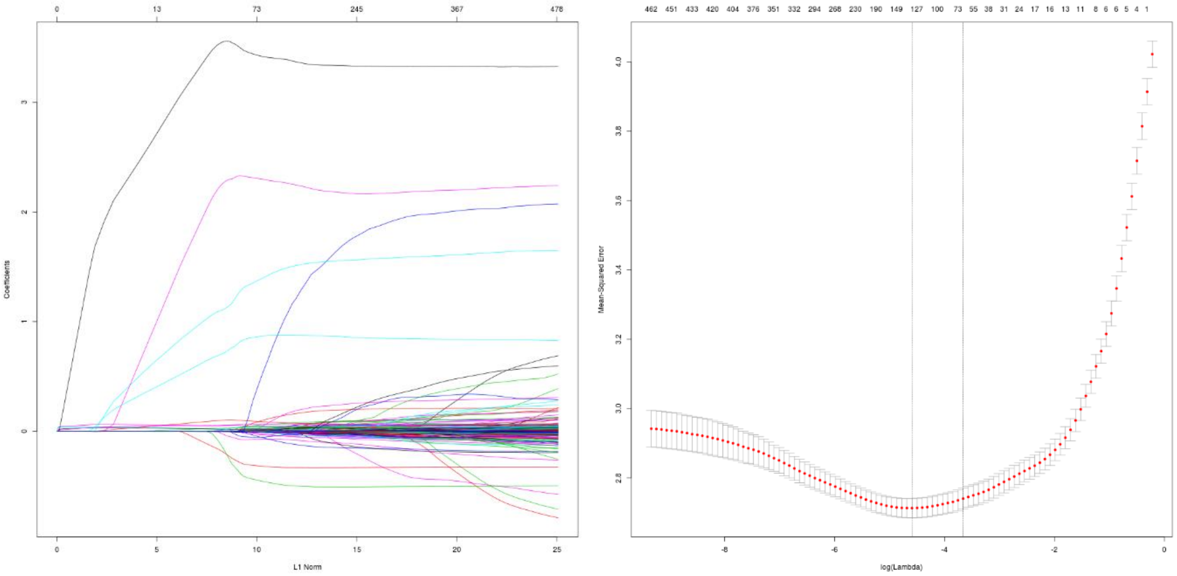


Figure 5.2 2015 coefficient estimation path (left) and cross-validation plot (right)

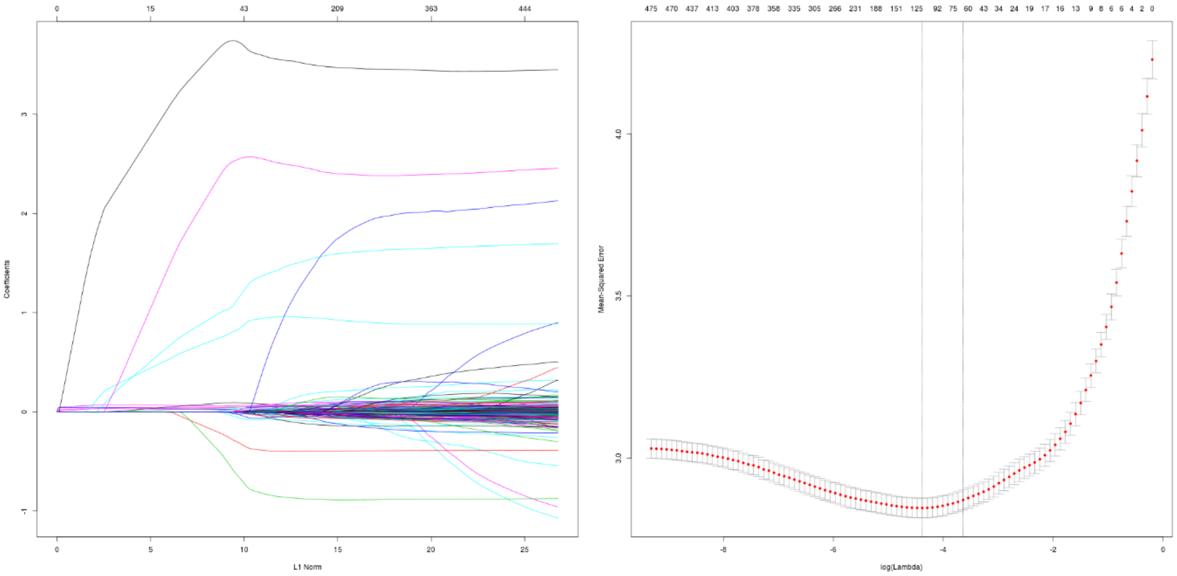


Figure 5.3 2016 coefficient estimation path (left) and cross-validation plot (right)

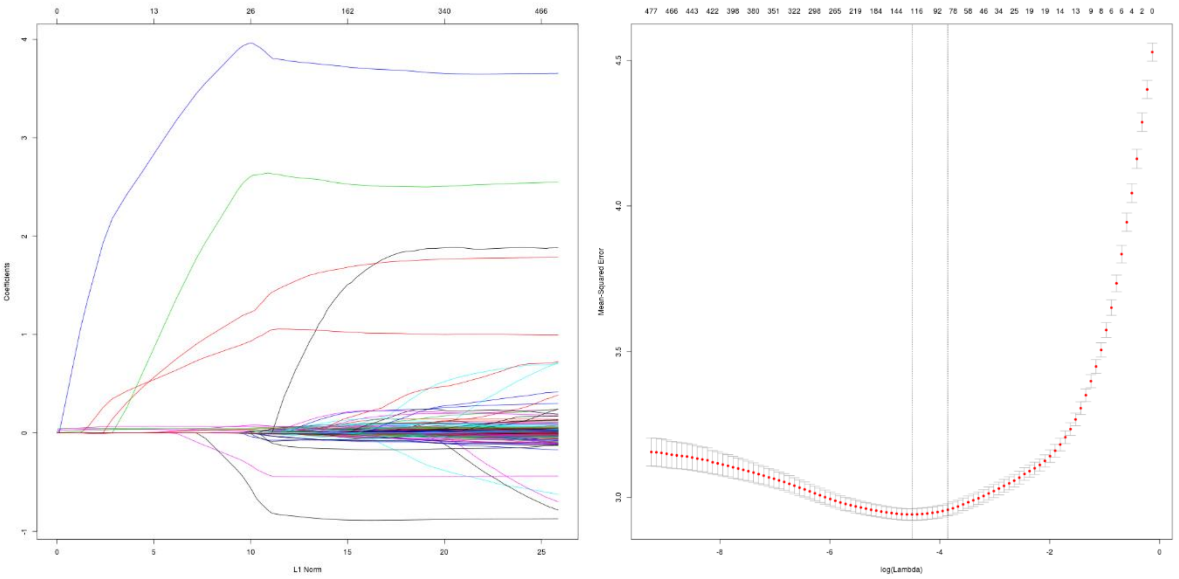


Figure 5.4 2017 coefficient estimation path (left) and cross-validation plot (right)

As shown in the figures above, different values of the lasso constraint yield different values of the coefficients, with some coefficients regularized to zero. The path of each coefficient can be seen, and at which value of the lasso constraint has a specific coefficient started to be non-zero. On the other hand, the cross-validation plots (right plot of each figure) shows the different λ tuning parameter and their corresponding Mean-Squared-Error (MSE). Two lines are marked in the plot: λ with the minimum MSE (dotted), and the largest λ with MSE within 1 standard error (SE) from the minimum MSE. Table 2 below shows the summary of sparse regression estimation results for the years 2015 to 2017.

Table 5.2. Sparse Regression Results Summary

Year	Total number of Non-zero Coefficients	Minimum MSE λ	MSE	SE
2015	136	0.01015336	2.7124	0.0279
2016	117	0.01251322	2.8462	0.0302
2017	122	0.01102266	2.9412	0.0200

It can be observed from Table 5.2. Sparse Regression Results Summary that Sparse Regression has reduced the number of coefficients from the original 521 independent variables significantly. The corresponding tuning parameter λ with the lowest MSE is also presented. The estimated non-zero coefficients from the minimum MSE λ could then be used for forecasting purposes. Due to space constraints, only the 2017 TTG Sparse Regression estimation is shown in Appendix 1.

5.5 Comparative Analysis of the Examined Models

The Stepwise Multiple Regression model and the Sparse Regression model use different methods of model evaluation. In order to compare the two models, a similar measure must be used. Because the multiple R-squared and adjusted R-squared for Sparse Regression cannot be computed, the MSE and SE of the Stepwise Multiple Regression model are computed by conducting model validation using a training and test set split. Before estimating the Stepwise Multiple Regression for each year, as in the previous section, and calculating the MSE and SE for each year through model validation, an initial estimation using all years is conducted to determine the appropriate training-test split as shown in Table 5.3 below.

Table 5.3. Training-Test Split Estimate

MSE	Training-Test Split
4.078533	50%-50%
3.433448	60%-40%
3.552372	70%-30%
3.803803	80%-20%
3.944834	90%-10%

The appropriate training-test split was 60% training set and 40% test set of the dataset and from this split based on the split with the lowest MSE. The Stepwise Multiple Regression model was estimated using 60% of the data for each year and validated using the remaining 50% test set. Table 5.4 and Table 5.5 below show the summaries for Stepwise Multiple Regression and Sparse Regression, respectively.

Table 5.4. Stepwise Multiple Regression 60%-40% Split Validation

Year	Number of Coefficients	MSE	SE
2015	177	3.2111	0.0191
2016	153	3.1587	0.0199
2017	147	3.3632	0.0209

Total estimation time (sec)	114,809 sec
Total estimation time (min)	1,913 min
Total estimation time (hr)	31.8 hr

Table 5.5. Sparse Regression Summary

Year	Number of Coefficients	MSE	SE
2015	136	2.7124	0.0279
2016	117	2.8462	0.0302
2017	122	2.9412	0.02
Total estimation time (sec)		97 sec	
Total estimation time (min)		1.61 min	

Sparse regression had a fewer number of coefficients in the final model compared to stepwise multiple regression. Furthermore, sparse regression had a lower MSE than the stepwise model for all years. This shows that even with fewer coefficients in the model, sparse regression had better predictive power than the stepwise model. The relatively higher MSE of the stepwise model is a confirmation of the low adjusted R-squared from the estimation results of the stepwise model. Also, considering the estimation times of both models, the Stepwise Multiple Regression model took 31.8 hours, while it only took 1.61 minutes to estimate the Sparse Regression model. This shows that the Sparse Regression model is a better model overall.

5.6 Fused Lasso

Establishing from the previous section that the Sparse Regression model with the lasso penalty is a better overall model than the stepwise multiple regression model, the application of Fused Lasso (Arnold and Tibshirani, 2014; Hastie et al., 2015), a generalization of the lasso penalty, to analyze the changes in TTG in TMA is explored. The Fused Lasso solves the following optimization problem:

$$\min_{(\boldsymbol{\theta}\in\mathbb{R}^N)} \left\{ \frac{1}{2}\sum_{i=1}^{N}(y_i - \theta_i)^2 + \lambda_1 \sum_{i=1}^{N}\left|\theta_j\right| + \lambda_2 \sum_{i\sim i\prime}^{N}\left|\theta_i - \theta_{i\prime}\right| \right\} \qquad (15)$$

The second term in the objective function is the original lasso penalty which shrinks the parameters θ_j to zero. The Fused Lasso objective function adds a second penalty, i.e., the third term in the objective function, which encourages neighboring coefficients to be similar. In a 2-dimensional graph, neighboring coefficients are defined as coefficients adjacent to each other. The Fused Lasso allows the researcher to highlight the critical areas in a noisy graph by smoothing.

In the context of TTG of TMA, the changes in TTG 2015 through 2017 cannot be directly compared because 1) the number of trucks installed with a digital tachometer is increasing throughout the years and may lead to misleading results, and 2) it will be challenging to highlight changes as well as concentration areas of TTG because of the noisy nature of the TTG plot. However, a comparison can be made by estimating a Fused Lasso model of the TTG. Because the 2015 TTG data was known to be collected under different conditions from the 2016 and 2017 data, only the 2016 and 2017 TTG Fused Lasso models will be estimated. Figure 5.5 shows the original TTG for 2016 and 2017 in grayscale. Changes from 2016 to 2017 is almost indiscernible.

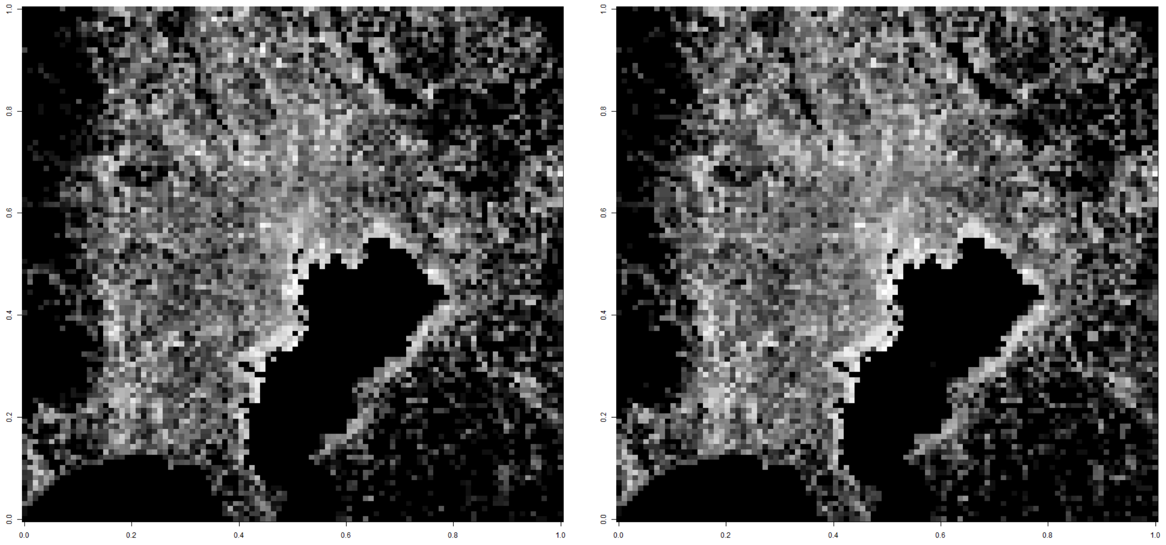


Figure 5.5 Original Truck Trip Generation for 2016 (left) and 2017 (right)

The original TTG graphs shown above were the inputs in estimating the Fused Lasso model. Unlike the Stepwise Multiple Regression model and Sparse Regression model in the previous section, only the intercept is estimated, that is, the TTG is estimated directly from the TTG graphs; hence, the resulting parameter vectors are itself TTG calibrated by the Fused Lasso model. Figure 5.6 below shows the sum of estimated TTG coefficients plotted against their respective λ. It is noted that the sum is constant through all the λ's for each year, meaning the coefficients only get redistributed depending on the tuning parameter λ but not change the sum. The Fused Lasso model for TTG was estimated through the "genlasso" package (Arnold and Tibshirani, 2014) in R.

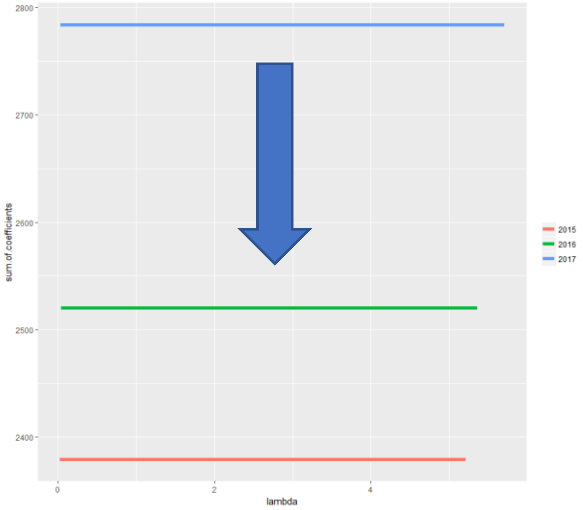


Figure 5.6 Plot (Wickham, 2009) of the sum of estimated

The sum of estimated coefficients allows the conversion of 2017 TTG estimates to be comparable to the 2016 TTG, as shown by the arrow in Figure 5.6, by multiplying a conversion factor which is the ratio of the sum of coefficients of 2016 to 2017 (2520.201/2784.156) as shown in equation (4) below:

$$TTG2017_{conv} = TTG2017_{orig} \times CF = TTG2017_{orig} \times \frac{2520.201}{2784.156} \qquad (16)$$

The conversion of 2017 TTG coefficients now allows the comparison of 2017 TTG and 2016 TTG. Three (3) tuning parameter λ's at different levels where chosen, and their corresponding coefficients were plotted, as shown in Figure 5.7 to Figure 5.9 below. It can be clearly seen that different levels of the tuning parameter λ, represent different strengths of the additional penalty in the Fused Lasso model: higher λ means stronger regularization and vice versa.

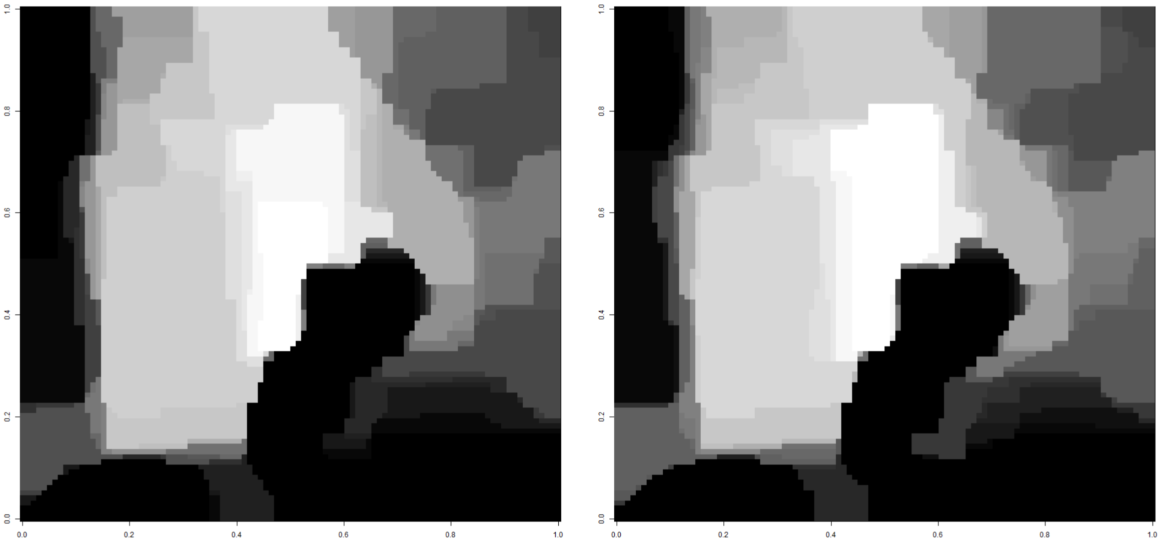


Figure 5.7 Calibrated Truck Trip Generation for 2016 (left) and 2017 (right) (λ=0.4806)

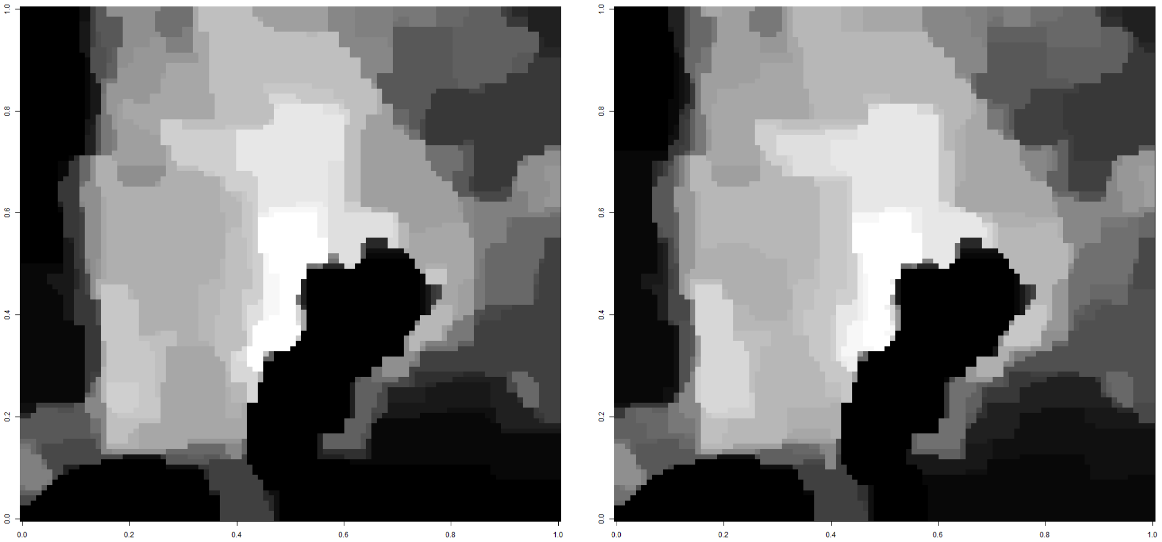


Figure 5.8 Calibrated Truck Trip Generation for 2016 (left) and 2017 (right) (λ=0.2777)

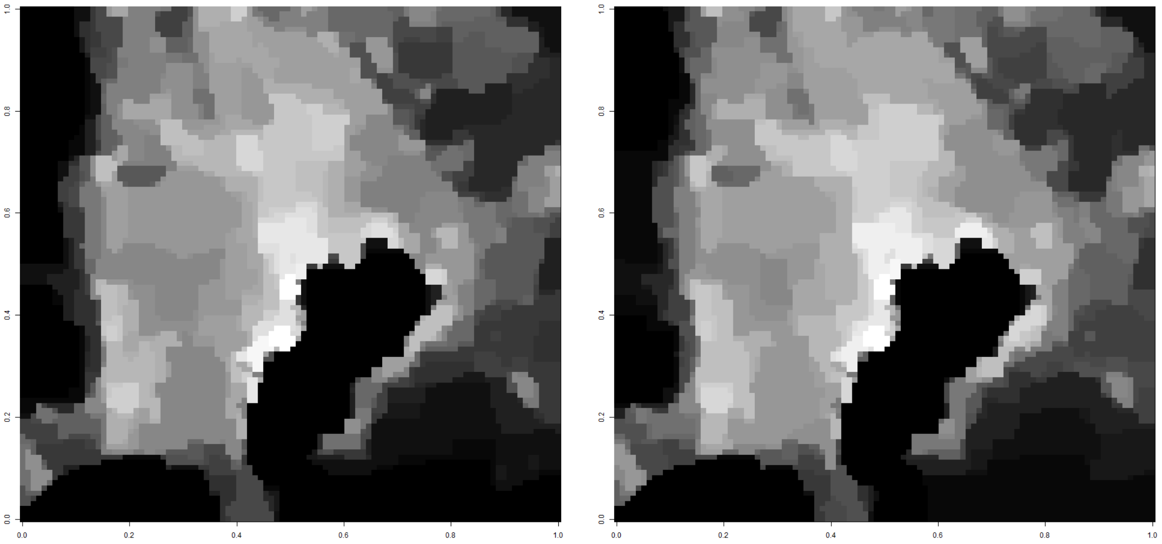


Figure 5.9 Calibrated Truck Trip Generation for 2016 (left) and 2017 (right) (λ=0.1696)

The lighter pixels in the calibrated TTG graphs indicate higher TTG than the darker pixels. Changes in TTG from 2016 to 2017 may not be obvious, but the calibrated and converted 2017 TTG can now be directly compared with 2016 TTG by taking the difference of their coefficients. Figure 5.10 to Figure 5.12 below show the difference between the 2017 TTG and the 2016 TTG. It can be seen from the figures that there are apparent increases in TTG from 2016 to 2017. Figure 5.12, with the lower λ tuning parameter, shows only a slight difference in TTG indicated by the relatively lighter pixel areas. However, examining Figure 5.10 with the higher λ tuning parameter, the difference becomes more apparent. Lighter pixels on the western regions of TMA are observed, which means that was an increase in 2017 TTG from 2016 TTG. This can be explained by 1) the relocation of logistics facilities from central TMA to the outskirts of TMA, and 2) the construction of logistics facilities near the Ken-O Expressway interchanges.

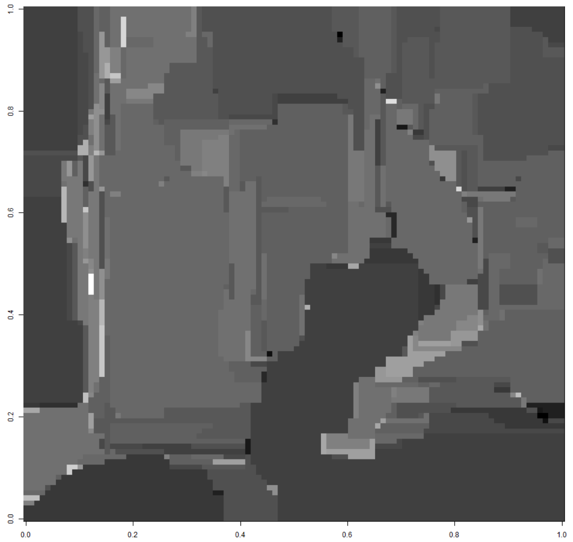


Figure 5.10 Difference between 2017 and 2016 TTG (λ=0.4806)

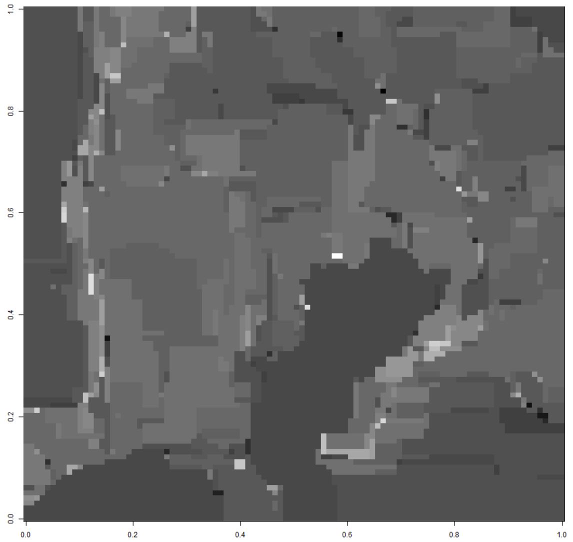


Figure 5.11 Difference between 2017 and 2016 TTG (λ=0.2777)

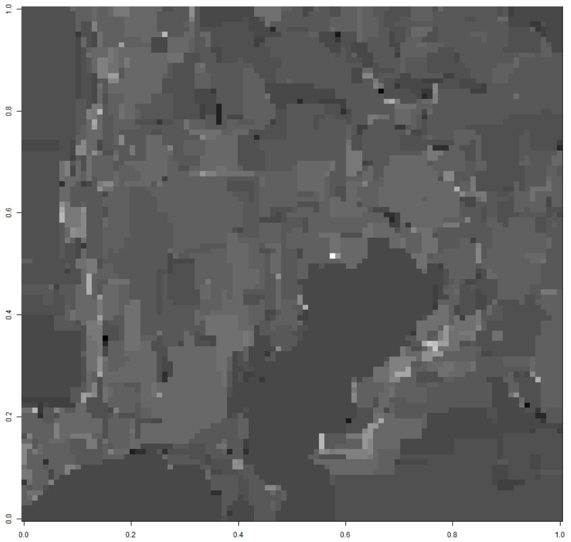


Figure 5.12 Difference between 2017 and 2016 TTG (λ=0.1696)

5.7 Summary

Truck probe data collected from digital tachometers were combined with economic census data, land-use distribution data, and accessibility data to create one dataset for modeling TTG. Two modeling frameworks for variable selection were compared, namely, Stepwise Multiple Regression, and Sparse Regression. Stepwise Multiple Regression estimates the coefficients of a model and does a model selection with varying combinations of independent variables by evaluating the AIC. On the other hand, Sparse Regression utilizes the lasso penalty to regularize estimated coefficients by encouraging shrinkage to zero. It was shown that the Sparse Regression framework is the better overall model for modeling TTG in TMA in terms of the number of parameters estimated, the MSE, and the computation time. The application of Fused Lasso as an extension of the Sparse Regression Modeling framework was also explored in comparing the TTG of 2016 and 2017. Because the number of trucks installed with digital tachometers is increasing through the years, a direct comparison cannot be made. However, it was shown that the Fused Lasso model could estimate the calibrated TTG coefficients for both years and convert the 2017 TTG to be comparable to the 2016 TTG. Thus, the difference in TTG from 2017 to 2016 was highlighted, and it was confirmed that there was an increase in TTG, especially in the western regions of TMA.

Chapter 6 TRUCK TRIP GENERATION MODELING CONSIDERING SPATIAL AUTO-CORRELATION IN KANTO AND KANSAI JAPAN USING SPATIAL REGRESSION

6.1 Introduction

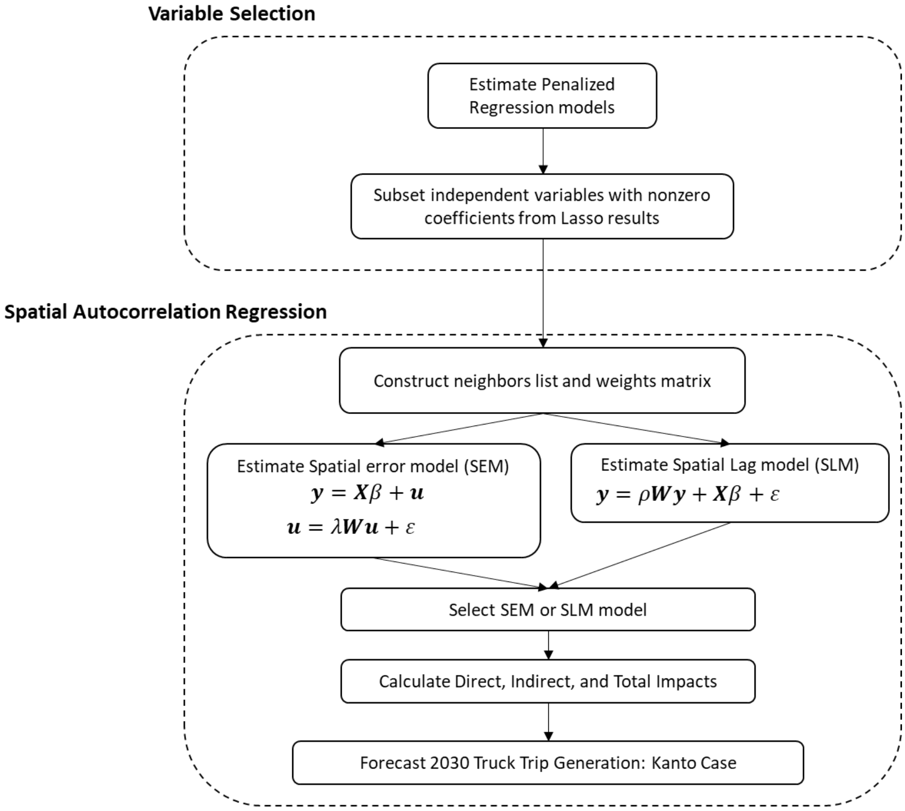
Freight transport and logistics is currently an active area in applied transportation research not only for its significance in economic development but also due to the externalities that it brings to society and the environment. For instance, trucks in Asia constitute only 9% of the transportation shares but emit 54% of the total CO_2 (Clean Air Asia, 2010). The transportation sector in the United States produces 29% of its total greenhouse gas (GHG) emissions, and 82% of which are from light-duty vehicles (which includes light-duty trucks) and medium/heavy-duty trucks (EPA, 2019). In Japan, the transportation sector constitutes 18% of the CO_2 emissions, of which about 35% are from trucks (Japan Automobile Manufacturers Association Inc., 2018). While efforts are being made to reduce GHG emissions from the transportation sector through hybridization and electrification, vehicles related to freight and logistics are posing a challenge, particularly medium and heavy-duty trucks, because of their high ton-kilometer demand for goods transport. For these reasons, there is a push to improve the understanding and description of freight and logistics systems, which, however, are hindered by issues and challenges regarding its complexity and data availability. In particular, the availability of data and tools for freight transport forecasting is a recurrent issue mentioned when it comes to modeling freight systems. Government departments responsible for transport policy have increasingly become concerned about the lack of availability of operational tools to forecast freight transport and understand possible effects of policy measures (Tavasszy and de Jong, 2014). Also, despite the abundance of data in logistics and transport operations, data is often proprietary and difficult to access (Tavasszy and de Jong, 2014). Given the complexity of freight systems and the difficulty of obtaining and lack of readily available data to develop practical models of freight systems, freight transport modeling is highly dependent on data that is within reach of researchers and analysts. Furthermore, the underlying spatial distribution and correlation in freight systems and networks, particularly in terms of activity and location of freight and logistics related facilities, when not considered, will result in inaccurate forecasts.

Japan has relatively an abundance of publicly available freight and logistics related data. These data are collected and maintained by the Japanese Government through its respective ministries and bureaus. The private sector also has its own data, if not richer, due to the need to assess performance, which affects their bottom line, albeit more difficult to access. The challenge is how to make use of such data for observing changes in freight systems and modeling freight transport and conducting rapid screening of variables as well as considering the spatial nature of freight systems. In this chapter, it is the aim to present a two-step framework for conducting rapid screening and selection of variables, then modeling the freight trip generation that accounts for the spatial relations in the system.

6.2 Framework of Analysis

The framework of analysis is shown in Table 6.1. A two-step approach is proposed: first is to conduct variable selection using penalized regression in order to lessen the number of independent variables that will be used for in the spatial regression. The reason for this is to aid in preventing overfitting and speed up the computation time of the spatial regression. This will be followed by two spatial regression methods: a spatial lag model and a spatial error model. Then the impacts are calculated because the regression coefficients of spatial regression cannot be directly interpreted due to the feedback effects of the spatial associations. Finally, a forecasting case is conducted.

Table 6.1 Framework of analysis for spatial autocorrelation regression


Variable Selection
Estimate Penalized Regression models
Subset independent variables with nonzero coefficients from Lasso results
Spatial Autocorrelation Regression
Construct neighbors list and weights matrix
Estimate Spatial error model (SEM)
$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{u}$
$\boldsymbol{u} = \lambda \boldsymbol{W}\boldsymbol{u} + \varepsilon$
Estimate Spatial Lag model (SLM)
$\boldsymbol{y} = \rho \boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\beta + \varepsilon$
Select SEM or SLM model
Calculate Direct, Indirect, and Total Impacts
Forecast 2030 Truck Trip Generation: Kanto Case

6.3 Data Abstract

6.3.1 Economic Census for Business Frame in Japan

Pursuant to the Japan Statistics Act (Act No. 53 of 2007), Japan conducts an economic census that aims to identify the current state of business activities of establishments and enterprises and to have a comprehensive overview of the industrial structure of Japan. It consists of two surveys, namely, the "Economic Census for Business Frame" and the "Economic Census for Business Activity." The primary difference between the two economic census is the former aims to identify the basic structure of establishments and enterprises of all industries in Japan such as total number of companies for each industry classification as well as the corresponding total number of employees in each industry classification while the latter aims to further identify the situation of economic activities of establishments and enterprises by comprehensively surveying and investigating accounting items such as sales (income) and costs in all the industrial classifications. For our purposes, we only utilize the "Economic Census for Business Frame," hereafter referred to as the "Economic Census," specifically data on the number of companies and the number of employees for each industry classification with 18 total industry classifications. We also limit our scope of analysis to East Japan (Kanto) and West Japan (Kansai) because these regions are the two contiguous areas with the largest economic activities that both have a major sea container port. The Economic Census of East and West Japan consists of 330 variables. The variables are classified into two (2) main categories: a) the number of companies and b) the number of workers for each industry classification (Table 6.2).

Table 6.2 Economic Census Industry Classification of Japan

A	Agriculture and Forestry
B	Fisheries
C	Mining and Quarrying of Stone and Gravel
D	Construction
E	Manufacturing
F	Electricity, Gas, Heat Supply and Water
G	Information and Communications
H	Transport and Postal Services
I	Wholesale and Retail Trade
J	Finance and Insurance
K	Real Estate and Goods Rental and Leasing

L	Scientific Research, Professional and Technical Services
M	Accommodations, Eating and Drinking Services
N	Living-Related and Personal Services and Amusement Services
O	Education, Learning Support
P	Medical, Health Care and Welfare
Q	Compound Services
R	Services, N.E.C.

6.3.2 Truck Probe Data

Truck probe data for trucks were acquired from a Japanese digital tachograph manufacturer that outfits digital tachographs to trucks. Data on the movement of trucks such as the date and time of departure from origin and arrival to the destination, as well as total travel time (in seconds) and total travel distance recorded by the digital tachograph, are remotely sent to the manufacturer’s server for maintenance and monitoring. Truck classification in the probe data follows the standards used by Japanese expressway companies managing electronic tollways and expressways, which classify trucks as small, regular, medium, large, and extra-large. We limit our analysis to large and extra-large trucks because these two are found to exhibit similar spatial distributions in terms of truck trip generation (Lidasan et al., 2017) and are similar in size and weight.

The truck probe data for large and extra-large trucks acquired were collected every day for three months (April to June) each year from 2016 to 2018. Inspection of the distribution of truck trip generation for East and West Japan would seem that trucks are increasing year-by-year. However, there is a caveat that the number of trucks that are outfitted with digital tachographs is also increasing, which means that previously unaccounted trucks through data collected from the digital tachograph could lead to an erroneous conclusion that truck trip generation is increasing through the years. In order to appropriately confirm the increase (or decrease) in the truck trips generated, here we introduce a method to determine and visually confirm the increase in generation by estimating a Fused Lasso model (Arnold and Tibshirani, 2014; Hastie et al., 2015) to calibrate the truck trip generation per cell so that they become comparable. The Fused Lasso model is a generalization of the lasso penalty in penalized regression methods which solves the optimization problem shown in equation (1):

$$\min_{(\boldsymbol{\theta}\in\mathbb{R}^N)}\left\{\frac{1}{2}\sum_{i=1}^{N}(y_i-\theta_i)^2+\lambda_1\sum_{i=1}^{N}|\theta_j|+\lambda_2\sum_{i\sim i\prime}^{N}|\theta_i-\theta_{i\prime}|\right\} \tag{17}$$

$$\min_{(\boldsymbol{\theta}\in\mathbb{R}^N)}\left\{\frac{1}{2}\sum_{i=1}^{N}(y_i-\theta_i)^2+\lambda\sum_{i\sim i\prime}^{N}|\theta_i-\theta_{i\prime}|\right\} \tag{18}$$

where y_i is the natural logarithm of total truck trip generation of cell *i*, θ_j the estimated parameter, in this case, a constant term, and λ_1 and λ_2 are tuning parameters estimated by cross-validation. The second term is a lasso penalty which regularizes the parameters θ_j, and as λ_1 increases, the stronger the constraint is, and parameters are shrunk to zero. The third term in the objective function encourages neighboring coefficients to be similar. We only need to estimate one tuning parameter for Fused Lasso based on *Lemma 4.1* (Hastie et al., 2015), as shown in equation (2). In a 2-dimensional graph, neighboring coefficients are those adjacent to one another. The truck trip generation values were rescaled so that a greyscale graph of the truck trip generation can be plotted and serve as input to the Fused Lasso model. We estimate the model and sum the estimated coefficients of each year. The sum of the coefficients for a series of tuning parameter values is constant, as shown in the graph for East Japan in Figure 6.1 (plot for West Japan is not shown due to space limitations).

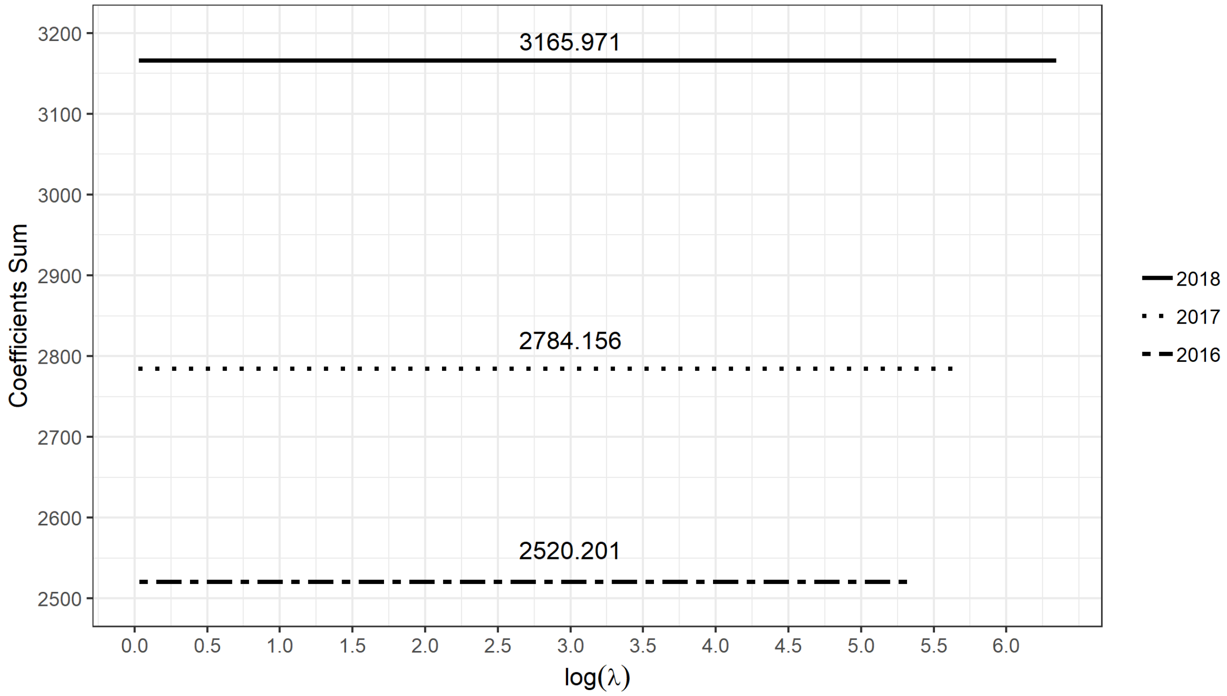


Figure 6.1 Kanto Fused Lasso coefficients sum plotted against tunning parameter $\boldsymbol{log(\lambda)}$

For the purpose of visualizing the change in truck trip generation, we consider the difference between 2016 and 2018. We do this by converting the Fused Lasso coefficients of 2018 to be comparable to 2016 by multiplying a conversion factor which is the ratio of the sum of coefficients of 2016 to 2018 (2784.156/3165.971) as shown in equation 2 below:

$$\theta'^{2018}_{i} = \theta^{2018}_{i} \times cf \qquad (19)$$

where θ^{2018}_{i} is the original calibrated truck trip generation for cell i, cf is the conversion factor, θ'^{2018}_{i} is the converted truck trip generation for comparison purposes. This scales down the 2018 sum of coefficients line (solid line) to the 2016 sum of coefficients line (short-long dashed line) in Figure 1. The same process was done for West Japan calibrated values. Conditional on the tuning parameter λ, the greyscale plot of the estimated parameters will have varying degrees of regularization, with higher λ's having stronger regularization on the parameters than lower λ's. A plot with relatively high regularization would highlight areas where truck trip generation is high, which are shown in lighter-colored cells. Now that the 2016 and the 2018 truck trip generation have been calibrated by the Fused Lasso model, their differences were computed conditional on a chosen tuning parameter λ. The differences were plotted to highlight areas with significant changes in truck trip generated visually. Figure 6.2 and Figure 6.3 show the difference between 2016 and 2018 truck trip generation.

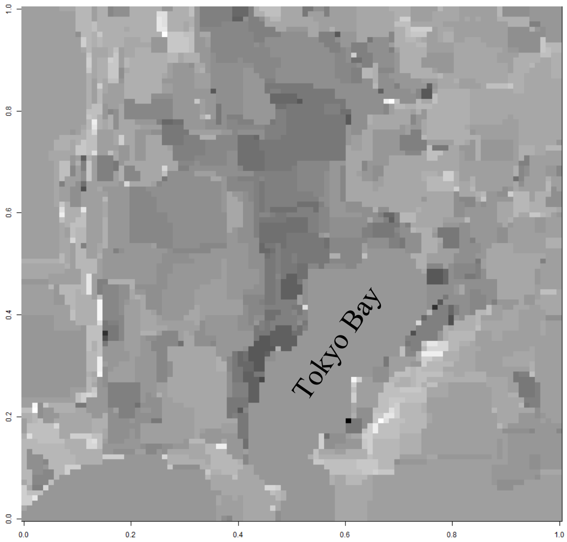


Figure 6.2 East Japan Fused Lasso calibrated generation difference ($\lambda = 0.4802$)

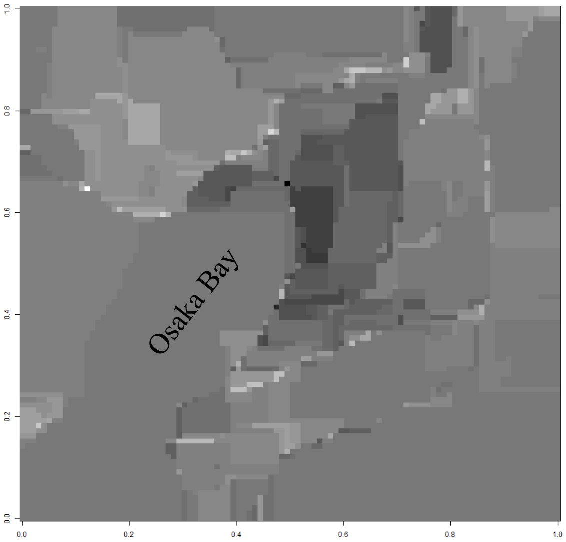


Figure 6.3 West Japan Fused Lasso calibrated generation difference ($\lambda = 0.3785$)

We confirm that there are regions where large and extra-large trucks trips increased in both East and West Japan since 2016. East Japan (Figure 6.2) had a noticeable increase in the western region, which is an indication of decentralization to the suburbs, and along ring roads and consolidation of functions of freight transport and logistics-related facilities (Sakai et al., 2016), There are also increases along the eastern side of Tokyo Bay. West Japan (Figure 6.3) also had a noticeable increase in truck trips generated since 2016, especially along the coastline of Osaka Bay. It can also be seen that there are increases going inland which, upon verification on a map, are expressways and, based on previous studies, areas surrounding expressways are found to be where freight and logistics related facilities tend to locate (Lidasan et al., 2017; Sakai et al., 2016).

6.3.3 Spatial Referencing of Data

A feature of surveys and data collected by the Japanese government is the use of a standardized spatial referencing system that makes it easy to georeference collected data from both a national and regional perspective. This makes it possible to identify the spatial distribution of survey items in Japan. There are three main levels of resolution (primary, secondary, and tertiary) in the said Japanese spatial referencing standard that is based on a grid of the Japanese national territory. Each cell in the grid has a corresponding code (referred locally as a “mesh code”), which represents a transformation of the longitude and latitude of the centroid of a cell for easier encoding. Although the two datasets mentioned above are taken from different sources, the standardized spatial referencing system of the Japanese government easily allows the combination of the two datasets to form a potentially more information-rich dataset. The resolution we used for our analysis is the tertiary level grid, which consists of 1-km by 1-km cells, i.e., one cell is a 1-km by 1-km data point in the data set. Based on this feature of the dataset, there is an inherent spatial relationship among the truck trip generation for each cell in the grid. In the following sections, we introduce a framework for the rapid screening of variables combined with truck trip generation modeling that accounts for the inherent spatial relations in the data.

6.4 Lasso Regression of Truck Trip Generation of East and West Japan

There is a total of 330 variables in each dataset of East and West Japan. However, using all 330 variables as inputs for spatial regression would be inefficient as not only would it a significant time estimation time too long, but not all variables may be relevant in modeling truck trip generation. Before estimating the spatial regression model of truck trip generation, we first estimate a penalized regression model for variable selection. To determine the penalty to be used for variable selection, we estimated penalized regression models with different penalties for each year and for number of companies, number of workers, and a combination of number of companies and number of workers and compared their Mean-Squared-Errors (MSE) as well as their coefficients of determination (R^2). Truck Trip Generation was log-transformed, and the variables were standardized by subtracting their mean and dividing by their standard deviation. The penalized regression models were estimated using the “oem” library in the R programming language (Huling and Qian, 2018). The summary of the results of penalized regression (Table 6.3) shows that for both East and West Japan, the Lasso penalty combined with variables of the number of companies and the number of workers has the lowest MSE. It can also be seen that from an initial number of 330 variables, all results show that the number of nonzero parameters estimated, i.e., the number of parameters that contribute to truck trip generation is below 330 variables. This will be the basis for selecting the variables that will be included in estimating the spatial regression model in the following section.

Table 6.3 Summary Results of Penalized Regression of Truck Trip Generation

		West Japan (Kansai)				East Japan (Kanto)			
	year	best model	MSE	R^2	No. of nonzero	best model	MSE	R^2	No. of nonzero
company	2016	SCAD	3.2217	0.232	35	Lasso	3.273	0.244	93
	2017	SCAD	3.368	0.251	51	Lasso	3.455	0.257	97
	2018	MCP	3.2597	0.274	51	Lasso	3.195	0.288	101
workers	2016	Lasso	3.2761	0.224	59	Lasso	3.375	0.228	67
	2017	Lasso	3.4432	0.239	69	Lasso	3.566	0.240	75
	2018	Lasso	3.3582	0.259	74	Lasso	3.347	0.261	70
company & workers	2016	Lasso	3.1349	0.262	129	Lasso	3.214	0.267	133
	2017	Lasso	3.2829	0.276	143	Lasso	3.389	0.280	147
	2018	Lasso	3.1876	0.298	144	Lasso	3.139	0.308	144

We have determined that the Lasso penalty will be used for variable selection and both the number of companies and the number of workers as variables for estimating the spatial regression model. The objective of this step is to conduct a rapid screening of variables for spatial regression estimation, and thus, the estimated coefficients won’t be analyzed. Nevertheless, the forest plot of the standardized coefficients of the Lasso regression results for both East (Kanto) and West (Kansai) Japan is shown in Figure 6.4. The number of companies in the road freight industry has the largest effect on truck trip generation, followed by machinery and equipment retail and warehousing. In terms of the number of workers, the manufacturing industry tends to have the biggest influence on the truck trips generated. While the number of workers in different manufacturing industries tends to increase truck trip generation, the same cannot be said of the number of companies because they tend to even out. However, these results do not consider the spatial relationship in the data.

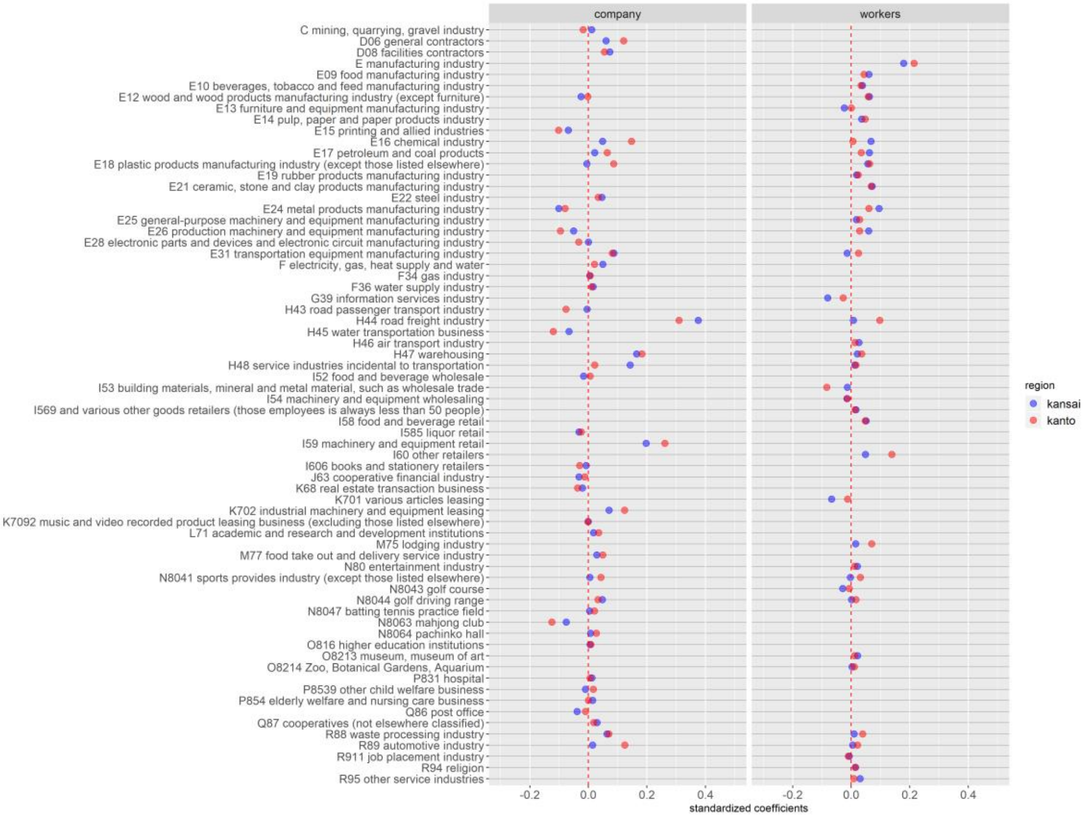


Figure 6.4 Estimated Coefficients of Lasso Regression (2018)

6.5 Spatial Regression

In this section, we finally model the spatial relationship after selecting the variables from the datasets of both East and West Japan in terms of truck trip generation of large and extra-large trucks. To estimate the spatial regression model, we need to define the neighborhood list and the corresponding weights matrix that will determine their spatial relationship. We defined neighbors using k-nearest neighbors and tested five (5) different k-nearest neighbors (KNNB) where $k \in \{4,8,12,16,20,24\}$. Based on the neighborhood lists, we also tested two definitions of weights matrix, a binary weights matrix, and an inverse-distance weights matrix defined as follows:

$$w_{ij} = 1 \quad \forall \quad j \in nb_i \qquad \text{(binary)} \qquad (20)$$

$$w_{ij} = \frac{1}{d_{ij}} \quad \forall \quad j \in nb_i \qquad \text{(inverse distance)} \qquad (21)$$

where w_{ij} is the weight of cell *i* to neighbor *j*, and d_{ij} is the distance of cell *i* to neighbor *j* for all *j* included in the neighbor list of *i*. The binary weights matrix assigns a weight of one (1) to all neighbors of cell *i* while the invers0e distance matrix assigns the inverse of the distance of cell *i* to cell *j*. Two spatial regression models have also been tested: a spatial lag model (SLM) and a spatial error model (SEM). The SLM model includes the spatial lag of the dependent variable, as shown in equations (6) and (7).

$$\boldsymbol{y} = \rho \boldsymbol{W} \boldsymbol{y} + \boldsymbol{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{22}$$

$$\boldsymbol{y} = (\boldsymbol{I}_n - \rho \boldsymbol{W})^{-1} \boldsymbol{X} \boldsymbol{\beta} + (\boldsymbol{I}_n - \rho \boldsymbol{W})^{-1} \boldsymbol{\varepsilon} \tag{23}$$

where $\boldsymbol{y}$ is the log-transformed truck trip generation, $\rho \boldsymbol{W} \boldsymbol{y}$ is the spatially lag term of the dependent variable where ρ is a parameter to be estimated determining the strength of influence of the spatial lag, and $\boldsymbol{W}$ is the weights matrix with elements w_{ij} as defined in equations (4) and (5) above, and ε is a normally distributed error term. On the other hand, the SEM model includes a spatially lagged error terms in addition to the normally distributed error term ε, as shown in equations (8) and (9).

$$\boldsymbol{y} = \boldsymbol{X} \boldsymbol{\beta} + \boldsymbol{u} \tag{24}$$

$$\boldsymbol{u} = \rho \boldsymbol{W} \boldsymbol{u} + \boldsymbol{\varepsilon} \tag{25}$$

where $\boldsymbol{y}$ is the log-transformed truck trip generation, $\boldsymbol{u}$ is the spatially lag error term which is composed of the spatial lag error term $\rho \boldsymbol{W} \boldsymbol{u}$ where ρ is a parameter to be estimated determining the strength of influence of the spatially lagged error term $\boldsymbol{u}$, and $\boldsymbol{W}$ is the weights matrix with elements w_{ij} as defined in equations (4) and (5) above, and $\boldsymbol{\varepsilon}$ is a normally distributed error term. SLM and SEM models were estimated by maximum likelihood using the “spdep” and “spatialreg” frameworks in the R programming language (Bivand et al., 2009) for the six (6) neighborhood definitions (KNNB where $k \in \{4,8,12,16,20,24\}$) and the two (2) weights matrix definitions (inverse distance and binary) for both East and West Japan.

Table 6.4 Summary of SLM and SEM models for West and East Japan

		West Japan (Kansai)				East Japan (Kanto)			
		Weights: Inverse Distance		Weights: Binary		Weights: Inverse Distance		Weights: Binary	
Type	knnb	R^2	Adj. R^2	R^2	Adj. R^2	R^2	Adj. R^2	R^2	Adj. R^2
Lag	4	0.390	0.385	0.293	0.287	0.385	0.379	0.297	0.290
Error	4	0.258	0.252	0.280	0.275	0.270	0.263	0.271	0.264
Lag	8	0.417	0.412	0.288	0.282	0.423	0.418	0.301	0.294
Error	8	0.115	0.108	0.273	0.267	0.223	0.215	0.260	0.253
Lag	12	0.425	0.420	0.316	0.310	0.447	0.441	0.373	0.367
Error	12	-0.086	-0.095	0.262	0.256	0.015	0.005	0.248	0.241
Lag	16	0.421	0.416	0.249	0.244	0.439	0.434	0.291	0.284
Error	16	-0.181	-0.190	0.268	0.262	-0.067	-0.078	0.256	0.249
Lag	20	0.419	0.415	0.180	0.174	0.442	0.436	0.272	0.265
Error	20	-0.225	-0.234	0.267	0.261	-0.250	-0.262	0.253	0.246
Lag	24	0.418	0.414	0.130	0.123	0.445	0.439	0.305	0.299
Error	24	-0.238	-0.248	0.265	0.259	-0.376	-0.389	0.252	0.245

The summary shows that for both East and West Japan and both weights matrix definition, the SLM model (type: lag) with neighborhood definition of 12 nearest neighbors (knnb: 12) had the highest adjusted R^2. It can also be seen that for SEM models with inverse distance weights matrix for KNNB ≥ 8, the coefficient of determination is negative, suggesting the models are worse than the mean of the data. Based on the estimation results, the SLM model using a neighborhood structure of 12 nearest neighbors with inverse distance weights matrix seems the appropriate model for both East and West Japan. A change in a single observation associated with any given explanatory variable in a spatial regression model will affect the cell itself (direct impacts) and potentially affect all other regions indirectly (indirect impacts) (LeSage and Pace, 2010). Average impacts are calculated from equations 10 to 12 (LeSage and Pace, 2010):

$$\overline{M}(r)_{direct} = n^{-1}tr\big(S_r(W)\big) \tag{26}$$

$$\overline{M}(r)_{total} = n^{-1}\iota'_n S_r(W)\iota_n \tag{27}$$

$$\overline{M}(r)_{indirect} = \overline{M}(r)_{total} - \overline{M}(r)_{direct} \tag{28}$$

where $\overline{M}(r)_{direct}$ are the average direct impacts of the independent variables, $\overline{M}(r)_{indirect}$ are the average indirect impacts of the independent variables, $\overline{M}(r)_{total}$ is the average total impacts of the independent variables, and $S_r(W) = (I_n - \rho W)^{-1} I_n \beta_r$. Table 6.5 shows the ten (10) highest positive and negative impacts to truck trip generation in East Japan, and Table 6.6 shows ten (10) highest positive and negative impacts to truck trip generation in West Japan for the SLM model using a neighborhood structure of 12 nearest neighbors with inverse distance weights matrix. Because independent variables are standardized by subtracting the mean and dividing by one standard deviation, the impact measures are on the same scale; thus, their relative magnitudes can be compared. The absolute value of the indirect impacts of the independent variables is larger than their direct impacts suggesting that not taking account of the spatial nature of truck trip generation in East and West Japan would lead to erroneous estimates of the parameters. Most of the independent variables with the highest (both positive and negative) impacts are from the retailing, road freight forwarding, manufacturing, and wholesale.

Table 6.5 Ten (10) Highest Impacts in East Japan

East Japan (Kanto)						
		Independent Variable	Type	Direct	Indirect	Total
10 Highest Positive Impacts	1	I59 machinery and equipment retailing	company	0.194	0.316	0.509
	2	I58 retail trade in food and beverage	workers	0.174	0.283	0.457
	3	I60 other retail business	workers	0.170	0.278	0.448
	4	H44 road freight forwarding industry	company	0.168	0.275	0.443
	5	E manufacturing industry	workers	0.159	0.260	0.419
	6	M75 accommodation industry	workers	0.109	0.179	0.288
	7	E16 chemical industry	company	0.095	0.156	0.251
	8	E18 plastic product manufacturing industry (excluding others)	company	0.086	0.140	0.226
	9	H47 warehouse industry	company	0.080	0.131	0.211
	10	K702 industrial machinery and equipment rental business	company	0.068	0.110	0.178
10 Highest	1	I54 machine tool wholesale business	company	-0.198	-0.323	-0.521
	2	N8063 mahjong club	company	-0.103	-0.169	-0.272

	3	E26 machine tool manufacturing industry for production	company	-0.100	-0.163	-0.263
	4	P832 general clinic	company	-0.097	-0.158	-0.254
	5	E24 metal product manufacturing industry	company	-0.090	-0.147	-0.237
	6	I2 retail trade	workers	-0.090	-0.146	-0.236
	7	K68 real estate business	company	-0.078	-0.128	-0.206
	8	E30 information and communication machinery and equipment manufacturing industry	workers	-0.075	-0.122	-0.197
	9	P833 dental clinic	company	-0.062	-0.101	-0.163
	10	H45 water transportation industry	company	-0.058	-0.095	-0.153

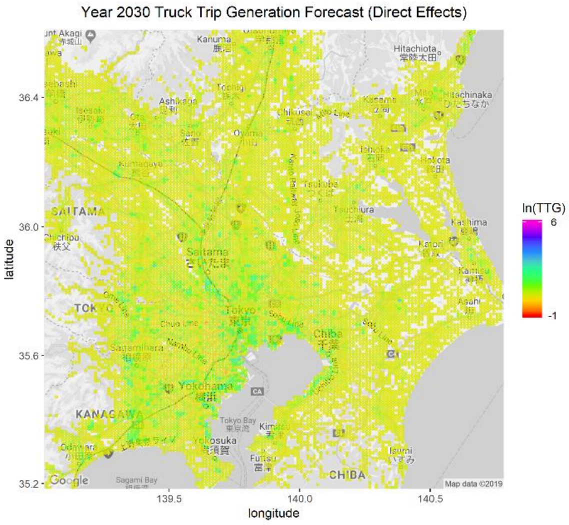
Table 6.6 Ten (10) Highest Impacts in West Japan

West Japan (Kansai)						
		Independent Variable	Type	Direct	Indirect	Total
10 Highest Positive Impacts	1	I50 ~ 55 wholesale	company	0.209	0.302	0.511
	2	I561 department stores, supermarkets	company	0.150	0.216	0.366
	3	I569 and various other goods retailers (employees less than 50 people)	company	0.144	0.208	0.353
	4	H44 road freight industry	company	0.141	0.203	0.343
	5	I59 machinery and equipment retail	company	0.119	0.172	0.292
	6	I60 other retailers	workers	0.113	0.163	0.276
	7	I50 ~ 55 Wholesale	workers	0.109	0.157	0.266
	8	E manufacturing industry	workers	0.108	0.156	0.264
	9	H48 service industries incidental to transportation	company	0.090	0.130	0.219
	10	I58 food and beverage retail	workers	0.089	0.129	0.219
10 Highest Negative Impacts	1	I56 various goods retail	company	-0.183	-0.265	-0.448
	2	I55 other wholesale	company	-0.118	-0.170	-0.289
	3	P835 medical treatment industry	company	-0.093	-0.134	-0.227
	4	I54 machinery and equipment wholesaling	workers	-0.093	-0.134	-0.227
	5	I52 food and beverage wholesale	company	-0.087	-0.125	-0.212
	6	E26 production machinery and equipment manufacturing industry	company	-0.084	-0.121	-0.204
	7	N804 sports provides industry	workers	-0.078	-0.113	-0.191

West Japan (Kansai)						
		Independent Variable	Type	Direct	Indirect	Total
	8	I53 building materials, mineral and metal material, such as wholesale trade	workers	-0.076	-0.110	-0.187
	9	E24 metal products manufacturing industry	company	-0.069	-0.100	-0.169
	10	G39 information services industry	workers	-0.051	-0.074	-0.125

6.6 Forecasting Truck Trip Generation: East Japan Case

To visualize the effect of not considering the spatial relations of truck trip generation in the data, we forecast truck trip generation in East Japan from the population projections released by the National Institute of Population and Social Security Research (IPSS) of Japan. This requires re-estimating the SLM mode with the 12-nearest-neighbor-structure and inverse distance matrix, which now includes the 2018 population for each cell in East Japan. The population projection of IPSS for 2030 was used to forecast the 2030 truck trip generation in East Japan. When considering only the direct impacts (without the effects of spatial lags), the truck trip generation is underestimated, as shown in Figure 6.5. The forecasts using direct impacts is equivalent to an ordinary least squares (OLS) regression. However, when forecasting the 2030 truck trip generation considering the spatial lags of truck trip generation, we get a more realistic forecast which appropriately models the areas with a concentration of truck trip generation especially near the coast of Tokyo Bay and to the western area of East Japan as shown in Figure 6.6.



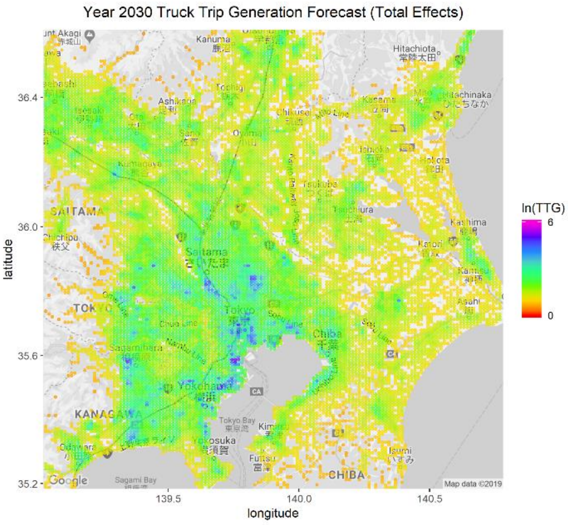


Figure 6.5 Truck Trip Generation Forecast for 2030 (Direct Impacts)

Figure 6.6 Truck Trip Generation Forecast for 2030 (Total Impacts)

6.7 Summary

A two-step approach to modeling truck trip generation was presented, particularly in the two regions with the highest economic activity in Japan. The most common issue with modeling freight transport is the lack of accessible data and the complexity of freight systems. Japan conducts an economic census that aims to identify the basic structure of establishments and enterprises of all industries in Japan, such as the total number of companies and the total number of employees in each industry classification. We utilize these publicly available data to model truck trip generation in East and West Japan. The truck trip generation data from a digital tachograph manufacturer that outfits digital tachographs in trucks was combined with the economic census data, and because both are encoded using the Japanese standard on spatial referencing data, the final dataset will have inherent spatial relations and allows for spatial regression modeling. Both the East and West Japan datasets consist of 330 independent variables, 165 of which are the total number of companies, and the remaining 165 representing the total number of employees for each industry classification. In order to efficiently estimate the spatial regression model, we first select the variables that will be used as input variables by estimating a penalized regression model. The penalized regression model results showed that the Lasso penalty was the best model and was used as the basis for variable selection. The results also showed that utilizing the total number of companies and the total number of employees provides a better model, and thus, both types of variables are used. From a total of 330 independent variables, the penalized regression (Lasso) has estimated 144 nonzero coefficients for both East and West Japan for the year 2018 which is less than half of the original 330 independent variables. This indicates that only less than half of the independent variables influence truck trip generation for both East and West Japan. However, the penalized regression model does not account for the spatial relations in the data. So, using the nonzero coefficients estimated as the basis for selecting the independent variables that influence the truck trips generated, we estimate a spatial regression model. Different neighborhood structures and weights matrices were tested in estimating the spatial regression models. Two specific model formulation was estimated; namely, the spatial lag model (SLM) and the spatial error model (SEM) were estimated. The results show that for both East and West Japan, an SLM with a neighborhood structure of 12 nearest neighbors and an inverse distance weights matrix had the highest adjusted coefficient of determination. However, estimated coefficients cannot be directly interpreted due to the feedback effects of the spatial lags of the dependent variable in the model, so total impacts must be calculated. Finally, we showed that using total impacts to forecast the 2030 truck trip generation for the East Japan case would result in more realistic values as compared to only considering direct impacts, which is akin to forecasting truck trip generation using ordinary least squares (OLS) model.

Chapter 7 SUMMARY, CONCLUSIONS, AND IMPLICATIONS

7.1 Summary and Conclusion

Freight volume generation and freight trip generation have known to be modeled using the classical method of regression with the desired output such as freight volume or freight vehicle volume as the dependent variable, and socio-economic variables as independent variables. Practical models that exist mainly differ in the functional structure of the linear model, such as whether there is an intercept or not and the type of socio-economic variables that are included in the linear equation. However, they mostly still follow the classical linear regression method. This poses problems of underfitting and mostly of overfitting. Underfitting is when the model doesn't get enough information from the data in order to properly represent the relations of the socio-economic variables to freight volume and freight trip generation. On the other hand, overfitting, is when the model learns too much from the data and becomes too sensitive that it poorly performs when it comes to forecasting. The overfitting problem is more common in freight models because practitioners tend to want to include as many variables as they can to improve the fit of the model to the data, but too much fit to the data comes at the price of the model poorly performing in prediction. A method of estimating the national freight volume generation in Japan as a varying-intercept model was presented in chapter 3, which deals with the overfitting issue. The varying intercept models had the best out-of-sample cross-validation performance, which shows that a varying intercept model is better for prediction than a model with a lot of independent variables.

Another issue with modeling freight trip generation is that freight trip generation is not independent of other factors such as the location of freight-related facilities. In the context of determining where and how much freight trips are produced; freight trips are a product of the decision of where logistics facilities are located and the socio-economic and locational variables. In addition, due to the limitations posed to where logistics facilities can locate, the natural spatial distribution of logistics facilities occurs. In the case of the Tokyo Metropolitan Area, logistics facilities are relocating to the suburbs, near expressway interchanges in the fringes of the Tokyo Metropolitan Area, or around Tokyo Bay. In the Kansai region, a similar observation can be seen of freight trips being concentrated around Osaka Bay. This has implications for freight trip generation as the spatial distribution of logistics facilities have now an effect on freight trips generated, especially when trying to determine where freight trips are generated. Modeling freight trip generation through simple linear regression will fail to consider the unobserved effects of the spatial distribution of logistics facilities and will lead to poor forecasts. For the issues of spatial factors in freight trip generation modeling, chapter 4 showed how to consider the location choice of logistics facilities in estimating freight trip generation. Chapter 4 presented a two-step approach of first modeling location choice and floor area, and using the estimated parameter for the logistics floor area, the truck trips generated can be estimated. Chapter 5 compared the lasso penalty for sparse regression and stepwise regression for variable selection. It was shown that the sparse regression framework is the better overall model for modeling freight trip generation in Tokyo Metropolitan Area in terms of the number of parameters estimated, the MSE, and the computation time. The application of Fused Lasso was also explored in comparing the freight trip generation of 2016 and 2017. Because the number of trucks installed with digital tachometers is increasing through the years, a direct comparison cannot be made. However, it was shown that the Fused Lasso model could estimate the calibrated coefficients for both years and convert 2017 to be comparable to 2016. Thus, the difference in freight trip generation from 2017 to 2016 was highlighted, and it was confirmed that there was an increase, especially in the western regions of the Tokyo Metropolitan Area.

In chapter 6, a two-step approach to modeling truck trip generation was presented. To efficiently estimate the spatial regression model, variables that will be used as input variables by estimating a penalized regression model. The penalized regression model results showed that the Lasso penalty provided that best model and was used as basis for variable selection. From a total of 330 independent variables, the penalized regression (Lasso) has estimated 144 nonzero coefficients for both East and West Japan for the year 2018, which is less than half of the original 330 independent variables. This indicates that only less than half of the independent variables influence the truck trips generated. The results show that for both East and West Japan, an SLM with a neighborhood structure of 12 nearest neighbors and an inverse distance weights matrix had the highest adjusted coefficient of determination. However, the estimated coefficients cannot be directly interpreted due to the feedback effects of the spatial lags of the dependent variable in the model, so the total impacts must be calculated. Finally, the total impacts were used to forecast the 2030 truck trip generation of East Japan, which resulted in more realistic values.

7.2 Recommendation

More studies should be done in testing the application of models for freight transportation modeling that do not overfit the data especially in the growing availability of sources of data other than surveys. Practitioners should start moving away from bad practices, like adding more variables just to improve the fit of the model. The reason is freight transportation models are used for forecasting, and improving fit will not necessarily improve forecasts but can be detrimental. Finally, the spatial aspects and spatial dependencies in the freight transport system should be considered by default because transportation and land-use are intertwined, and land-use has a spatial aspect to it inherently.

7.3 Implications

The implication for freight transport modeling is that we can do better by avoiding overfitting through including a lot of variables. Not only will the model take longer to estimate but it will perform poorer when it comes to forecasting because the model will learn too much from the current freight transport data that when future conditions are doesn't look like the data the model was trained with, the predictions will not be reliable. A degree of flexibility through a more skeptic set of model parameters, i.e., regularized parameters that are often smaller than the unbiased parameters from the classical regression model are needed. Also, better forecasts can be conducted by being able to consider the spatial aspects dependencies in the freight transport system.

7.4 Issues for Further Study

Other issues that need further study is the consideration of varying effects in the predictors themselves (as opposed to varying-intercepts) through varying coefficients. In the case of the national freight volume generation, these are the varying effects in population and GRP. The reason varying coefficients must be considered is that different areas have different levels of economic activity as well as different consumption patterns.

REFERENCES

Allen, J., Browne, M., Woodburn, A., 2010. Integrated transport policy in freight transport, in: Givoni, M., Banister, D. (Eds.), INTEGRATED TRANSPORT FROM POLICY TO PRACTICE. Routledge, New York, pp. 75–98.

Amemiya, T., 1984. Tobit models: A survey. J. Econom. 24, 3–61. https://doi.org/10.1016/0304-4076(84)90074-5

Arnold, T.B., Tibshirani, R.J., 2014. genlasso: Path algorithm for generalized lasso problems. Cran.

Bivand, R., Müller, W.G., Reder, M., 2009. Power calculations for global and local Moran's I, Computational Statistics and Data Analysis. https://doi.org/10.1016/j.csda.2008.07.021

Bürkner, P.C., 2018. Advanced Bayesian multilevel modeling with the R package brms. R J. https://doi.org/10.32614/rj-2018-017

Bürkner, P.C., 2017. brms: An R package for Bayesian multilevel models using Stan. J. Stat. Softw. https://doi.org/10.18637/jss.v080.i01

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M.A., Guo, J., Li, P., Riddell, A., 2017. Stan: A probabilistic programming language. J. Stat. Softw. https://doi.org/10.18637/jss.v076.i01

Chow, J.Y.J., Yang, C.H., Regan, A.C., 2010. State-of-the art of freight forecast modeling: Lessons learned and the road ahead. Transportation (Amst). https://doi.org/10.1007/s11116-010-9281-1

Clean Air Asia, 2010. GHG and Air Pollutant Indicators for Transport and Energy Sectors in Asia [WWW Document]. URL https://cleanairasia.org/ghg-and-air-pollutant-indicators-for-transport-and-energy-sectors-in-asia-2/ (accessed 6.24.19).

de Jong, G., Vierth, I., Tavasszy, L., Ben-Akiva, M., 2013. Recent developments in national and international freight transport models within Europe. Transportation (Amst). https://doi.org/10.1007/s11116-012-9422-9

El-maghraby, A., 2000. Truck Trip Generation Models for Trailer Operation. Transp. Res. Rec. J. Transp. Res. Board 1719, 1–9.

EPA, 2019. Fast Facts on Transportation Greenhouse Gas Emissions [WWW Document]. URL https://www.epa.gov/greenvehicles/fast-facts-transportation-greenhouse-gas-emissions (accessed 7.7.19).

Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G., Gnanzou, D., 2015. How “big data” can make big impact: Findings from a systematic review and a longitudinal case study. Int. J. Prod. Econ. 165, 234–246. https://doi.org/10.1016/j.ijpe.2014.12.031

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. 33. https://doi.org/10.18637/jss.v033.i01

Gelman, A., Goodrich, B., Gabry, J., Vehtari, A., 2019. R-squared for Bayesian Regression Models. Am. Stat. https://doi.org/10.1080/00031305.2018.1549100

Geurs, K.T., van Wee, B., 2004. Accessibility evaluation of land-use and transport strategies: Review and research directions. J. Transp. Geogr. 12, 127–140. https://doi.org/10.1016/j.jtrangeo.2003.10.005

Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical Learning with Sparsity: The Lasso and Generalizations, Crc. CRC Press, Boca Raton, FL. https://doi.org/10.1201/b18401-1

Heckman, J., 1979. Sample specification bias as a selection error. Econometrica 47, 153–162.

Hesse, M., 2004. Regulation of Regional Distribution Complexes 95, 162–173.

Hesse, M., 2002. Logistics real estate markets : indicators of structural change , linking land use and freight transport. ERSA 2002-Conference From Ind. to Adv. Serv. 1–16.

Holguín-Veras, J., Jaller, M., Destro, L., Ban, X., Lawson, C., Levinson, H., 2011. Freight Generation, Freight Trip Generation, and Perils of Using Constant Trip Rates. Transp. Res. Rec. J. Transp. Res. Board 2224, 68–81. https://doi.org/10.3141/2224-09

Holguin-Veras, J., Jaller, M., Sanchez-Diaz, I., Campbell, S., Lawson, C., 2014. Freight Generation and Freight Trip Generation Models, in: Freight Transport Modelling. Elsevie, London and Waltham, pp. 43–87.

Holguín-Veras, J., López-Genao, Y., Salam, A., 2002. Truck-Trip Generation at Container Terminals: Results from a Nationwide Survey. Transp. Res. Rec. J. Transp. Res. Board 1790, 89–96. https://doi.org/10.3141/1790-11

Hong, J., 2007. Transport and the location of foreign logistics firms: The Chinese experience. Transp. Res. Part A Policy Pract. 41, 597–609. https://doi.org/10.1016/j.tra.2006.11.004

Huling, J.D., Qian, P.Z.G., 2018. Fast Penalized Regression and Cross Validation for Tall Data with the oem Package VV.

Iwakata, M., Okamoto, N., Ishida, H., Hyodo, T., 2015. A Study of Freight Facility Location in Tokyo Metropolitan and its Future. J. East. Asia Soc. Transp. Stud. 11, 722–738.

James, E.H., Ginn, J.R., James, F.J., Kain, J.F., Straszheim, M.R., 1972. Land-Use — Transportation Planning Studies, in: Empirical Models of Urban Land Use: Suggestions on Research Objectives and Organization. National Bureau of Economic Research Volume, pp. 6–16.

James, G., Tibshirani, R., Hastie, T., 2013. Linear Model Selection and Regularization, in: An Introduction to Statistical Learning with Applications in R. Springer New York, pp. 203–264. https://doi.org/10.1007/978-1-4614-7138-7

Japan Automobile Manufacturers Association Inc., 2018. The Motor Industry of Japan 2018.

Kulpa, T., 2014. Freight Truck Trip Generation Modelling at Regional Level. Procedia - Soc. Behav. Sci. 111, 197–202. https://doi.org/10.1016/j.sbspro.2014.01.052

LeSage, J., Pace, R.K., 2010. Motivating and Interpreting Spatial Econometric Models, in: Introduction to Spatial Econometrics. CRC Press, pp. 25–43. https://doi.org/10.1201/9781420064254.ch2

Lidasan, A.H.S.B., Umeda, S., Hyodo, T., 2017. Characteristics of Logistics Facilities Allocation , Size and Truck Generation by Tokyo Metropolitan Area Urban Freight Survey. Int. J. Oper. Res. 14, 139–155.

Lindsey, C., Mahmassani, H.S., Mullarkey, M., Nash, T., Rothberg, S., 2014. Regional logistics hubs, freight activity and industrial space demand: Econometric analysis. Res. Transp. Bus. Manag. 11, 98–104. https://doi.org/10.1016/j.rtbm.2014.06.002

McElreath, R., 2018. Statistical rethinking: A bayesian course with examples in R and stan, Statistical Rethinking: A Bayesian Course with Examples in R and Stan. https://doi.org/10.1201/9781315372495

Neal, R.M., 2011. MCMC using hamiltonian dynamics, in: Handbook of Markov Chain

Monte Carlo. https://doi.org/10.1201/b10905-6

Newman, P.W.G., Kenworthy, J.R., 1996. The land use-transport connection: An overview. Land use policy 13, 1–22. https://doi.org/10.1016/0264-8377(95)00027-5

R Developement Core Team, 2018. R: A Language and Environment for Statistical Computing. R Found. Stat. Comput.

Rao, C., Goh, M., Zhao, Y., Zheng, J., 2015. Location selection of city logistics centers under sustainability. Transp. Res. Part D Transp. Environ. 36, 29–44. https://doi.org/10.1016/j.trd.2015.02.008

Rodrigue, J.P., Comtois, C., Slack, B., 2016. The geography of transport systems, The Geography of Transport Systems. https://doi.org/10.4324/9781315618159

Sakai, T., Kawamura, K., Hyodo, T., 2016. Logistics Facility Distribution in Tokyo Metropolitan Area: Experiences and Policy Lessons. Transp. Res. Procedia 12, 263–277. https://doi.org/10.1016/j.trpro.2016.02.064

Sorratini, J., Smith, R., 2000. Development of a Statewide Truck Trip Forecasting Model Based on Commodity Flows and Input-Output Coefficients. Transp. Res. Rec. J. Transp. Res. Board 1707, 49–55. https://doi.org/10.3141/1707-06

Tadi, R.R., Balbach, P., 1994. Truck Trip Generation Characteristics of Nonresidential Land Uses. ITE J. 43–47.

Taniguchi, E., Thompson, R., Yamada, T., van Duin, R., 2001. Introduction, in: City Logistics Network Modelling and Intelligent Transport Systems. Pergamon, Oxford, pp. 1–9.

Tavasszy, L., de Jong, G., 2014. Introduction, in: Modelling Freight Transport. Elsevier, London, pp. 1–11.

Tavasszy, L.A., Ruijgrok, K., Davydenko, I., 2012. Incorporating Logistics in Freight Transport Demand Models: State-of-the-Art and Research Opportunities. Transp. Rev. https://doi.org/10.1080/01441647.2011.644640

Tjur, T., 2009. Coefficients of determination in logistic regression models - A new proposal: The coefficient of discrimination. Am. Stat. https://doi.org/10.1198/tast.2009.08210

Toomet, O., Henningsen, A., 2008. Sample Selection Models in R: PackagesampleSelection.

J. Stat. Softw. 27, 1–23. https://doi.org/10.1007/s00221-005-0322-5

Vehtari, A., Gelman, A., Gabry, J., 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat. Comput. https://doi.org/10.1007/s11222-016-9696-4

Wagner, T., 2010. Regional traffic impacts of logistics-related land use. Transp. Policy 17, 224–229. https://doi.org/10.1016/j.tranpol.2010.01.012

Wegener, M., 2004. Overview of land-use transport models. Transp. Geogr. Spat. Syst. 127–146. https://doi.org/10.1007/s10654-011-9614-1

Wickham, H., 2009. Elegant Graphics for Data Analysis, Media. Springer-Verlag New York, New York. https://doi.org/10.1007/978-0-387-98141-3

Woudsma, C., Jakubicek, P., Dablanc, L., 2015. Logistics sprawl in North America: methodological issues and a case study in Toronto. Transp. Res. Procedia IN SUBMISS, 474–488. https://doi.org/10.1016/j.trpro.2016.02.081

Woudsma, C., Jensen, J.F., Kanaroglou, P., Maoh, H., 2008. Logistics land use and the city: A spatial-temporal modeling approach. Transp. Res. Part E Logist. Transp. Rev. 44, 277–297. https://doi.org/10.1016/j.tre.2007.07.006

APPENDIX

2017 TTG Sparse Regression Estimation Results

Variable	Variable Sub-label	Coef.
(Intercept)		1.9546
(Number of establishments) A to R All industries,	A~R	0
C to E secondary industry,	C~E	0
C05 Mining, quarrying, gravel sampling	C05	0
(Number of establishments) D Construction industry,	D	0
D 06 Comprehensive construction industry,	D06	0
D07 Construction work by job (excluding facility construction work)	D07	0
(Number of establishments) D 08 Equipment construction industry,	D08	0
E Manufacturing industry,	E	0
E09 Foodstuff manufacturing industry	E09	0.0058
(Number of establishments) E10 Beverages · Tobacco · Feed manufacturing industry,	E10	0.0569
E11 Textile industry,	E11	0
E12 Wood and wood product manufacturing industry (excluding furniture)	E12	0.0009
(Number of establishments) E13 Furniture / Fittings Manufacturing industry,	E13	-0.0461
E14 Manufacture of pulp, paper, and paper processed goods,	E14	0
E15 Printing and related business	E15	-0.0049

(Number of establishments) E16 Chemical industry,	E16	0
E17 Petroleum products / coal product manufacturing industry,	E17	0.0097
E 18 Plastic product manufacturing industry (excluding others)	E18	0.0184
(Number of establishments) E 19 Rubber product manufacturing industry,	E19	0
E20 Leather, same product / fur manufacturing industry,	E20	0.0002
E21 Ceramic · Soil product manufacturing industry	E21	0
(Number of establishments) E22 Steel industry,	E22	0
E23 Nonferrous metals manufacturing industry,	E23	-0.0199
E24 Metal product manufacturing industry	E24	-0.0004
(Number of establishments) E25 Machinery and equipment for manufacturing machinery,	E25	0
E 26 Production machinery and equipment manufacturing industry,	E26	-0.0066
E 27 Industrial machinery and equipment manufacturing industry	E27	0
(Number of establishments) E28 Electronic parts / devices ·	E28	0
Electronic circuit manufacturing industry, E29 Electrical machinery, and equipment manufacturing industry,	E29	-0.0026
E 30 Information and communication machinery and equipment manufacturing industry	E30	0
(Number of establishments) E31 Transportation machinery and equipment manufacturing industry,	E31	0.0064
E32 Other manufacturing industry,	E32	0
F to R tertiary industry	F~R	0
(Number of establishments) F Electricity, gas, heat supply, water supply industry,	F	0
F33 Electric industry,	F33	0
F34 gas industry	F34	0
(Number of establishments) F35 heat supply industry,	F35	0
F36 water industry,	F36	0.0694
G information communication industry	G	0
(Number of establishments) G 37 Communications industry,	G37	0
G38 Broadcasting business,	G38	0
G39 Information service industry	G39	0
(Number of establishments) G40 Internet accompanying service industry,	G40	0
G41 Video / voice / text information system work,	G41	0
H Transportation industry, postal service	H	0
(Number of establishments) H42 Railway industry,	H42	-0.0209
H43 road passenger transportation business,	H43	-0.0092
H44 Road Freight Forwarding Industry	H44	0.0595
(Number of establishments) H45 Water transport industry,	H45	-0.0758
H46 Air Transport Industry,	H46	0.0265
H47 warehouse industry	H47	0.0434
(Number of establishments) H48 Service industry incidental to transportation,	H48	0
H49 Postal business (including credit facilities business),	H49	0.0625
I Wholesale and retail trade	I	0
(Number of establishments) I1 Wholesale business,	I1	0
I50 Various goods Wholesale business,	I50	0
I 51 Pharmaceuticals, clothing, etc. wholesale business	I51	0
(Number of establishments) I 52 Food and beverage wholesale business,	I52	0
I53 Building materials, wholesale industry such as mineral and metallic materials,	I53	0
I54 Machine tool wholesale business	I54	-0.0011
(Number of establishments) I55 Other wholesale business,	I55	0
I2 retail industry,	I2	0
I 56 Various product retailers	I56	0.0075
(Number of establishments) I561 Department store, general supermarket,	I561	0
I569 Various other merchandise retailers (employees are always 50	I569	0
, I57 Textile, clothing, personal belongings retailing	I57	0
(Number of establishments) I 58 Retail trade in food and beverage,	I58	0
I 581 Various grocery retailers,	I581	-0.0005
I585 liquor retail trade	I585	-0.0118

(Number of establishments) I59 Machinery and equipment retailing,	I59	0.0144
I60 Other retail business,	I60	0
I 603 Pharmaceuticals and cosmetics retail trade	I603	0
(Number of establishments) I 606 Book · stationery retailing,	I606	0
J Financial industry, insurance industry,	J	0
J62 Banking business	J62	0
(Number of establishments) J 622 Bank (excluding Central Bank),	J622	0
J63 Cooperative organization Finance industry,	J63	0
J631 Small business etc. Finance industry	J631	0
(Number of establishments) K Real estate industry, rental goods business,	K	0
K68 Real Estate Business,	K68	0
K69 Real estate leasing industry · Management industry	K69	0
(Number of establishments) K 70 Goods rental business,	K70	0
K701 Various goods rental business,	K701	0.0665
K702 Industrial machinery and equipment rental business	K702	0.0817
(Number of establishments) K 703 Office equipment leasing business,	K703	0
K 704 Automobile rental business,	K704	0
K705 Sports / entertainment equipment rental business	K705	0
(Number of establishments) K 709 Other goods Leasing business,	K709	0
K7092 Music and video record leasing business (excluding others),	K7092	-0.0265
K 7099 Goods not classified as other leasing business	K7099	-0.0011
(Number of establishments) L academic research, specialized / technical service industry,	L	0
L71 Academic and development research institution,	L71	0
L 72 Professional service industry (not classified elsewhere)	L72	0
(Number of establishments) L 73 Advertising business,	L73	0
L 74 technical service industry (not classified elsewhere),	L74	0
M accommodation industry, food service business	M	0
(Number of establishments) M75 Accommodation,	M75	0
M751 Ryokan, Hotel,	M751	0.0233
M7591 Company / group accommodation	M7591	0
(Number of establishments) M76 restaurant,	M76	0
M77 Takeaway / food delivery service,	M77	0
N Living related service industry, the entertainment industry	N	0
(Number of establishments) N 78 Laundry · Barber · Beauty · Bathroom business,	N78	0
N79 Other life-related services industry,	N79	0
N80 entertainment industry	N80	0
(Number of establishments) N801 Movie Theater,	N801	0
N802 entertainment venue (excluding others), entertainer team,	N802	0
N804 sports facilities offering business	N804	0
(Number of establishments)	N8041	0.0797
N8041 Sports facility offering (excluding others),	N8042	0
N8042 gymnasium, N8043 golf course	N8043	-0.1601
(Number of establishments) N8044 Golf practice range,	N8044	0
N8045 Bowling alley,	N8045	0
N8046 Tennis court	N8046	0
(Number of establishments) N8047 Batting / tennis practice range,	N8047	0.1063
N8048 fitness club,	N8048	0
N806 game room	N806	0
(Number of establishments) N8063 Mahjong Club,	N8063	-0.0349
N8064 Pachinko hall,	N8064	0.0417
N8065 game center	N8065	0
(Number of establishments) N 8069 Other play area,	N8069	0
O education, learning support industry,	O	0
O81 School Education	O81	0
(Number of establishments) O 811 Kindergarten,	O811	-0.0704

O 812 Elementary school,	O812	0
O 813 Middle School	O813	0
(Number of establishments) O 814 High school, Secondary school,	O814	0
O 815 special support school,	O815	0
O 816 Higher education institution	O816	-0.0154
(Number of establishments) O 817 vocational schools, various schools,	O817	0
O 82 Other education, learning support industry,	O82	0
O 8213 museum, museum	O8213	0
(Number of establishments) O 82 14 Zoo, botanical garden, aquarium,	O8214	0.1577
O823 Learning cram school,	O823	0
O8241 Music teaching work	O8241	-0.0316
(Number of establishments) O 824 Foreign language conversation teaching business,	O8245	0
O8246 Sports / health teaching professor,	O8246	0
P Medical, welfare	P	0
(Number of establishments) P83 Medical industry,	P83	0
P831 Hospital,	P831	0.0001
P832 general clinic	P832	-0.0136
(Number of establishments) P833 Dental clinic,	P833	0
P835 Therapeutic industry,	P835	0
P84 Health sanitation	P84	0
(Number of establishments) P85 Social insurance / social welfare / nursing care business,	P85	0
P853 Child Welfare Project,	P853	0
P8531 nursery school	P8531	0
(Number of establishments) P8539 Other child welfare projects,	P8539	0
P854 Welfare and long-term care business for the elderly,	P854	-0.0038
P855 Welfare service for people with disabilities	P855	0
(Number of establishments)	P859	0.0040
P859 Other social insurance · social welfare · nursing care business,	Q	0
Q combined service business, Q86 post office	Q86	-0.0736
(Number of establishments) Q 87 Cooperative associations (not classified elsewhere),	Q87	0.0659
R service industry (not classified elsewhere),	R	0
R88 Waste disposal industry	R88	0.0385
(Number of establishments) R 89 Automobile maintenance industry,	R89	0.0433
R90 Machine etc. Repair work (excluding others),	R90	0
R91 Employment placement / worker dispatch business	R91	0
(Number of establishments) R 911 Employment introduction industry,	R911	0
R 92 Other business services industry,	R92	0
R93 Political, Economic and Cultural Organizations	R93	0
(Number of establishments) R 933 Academic / Cultural Organization,	R933	0
R 94 Religion,	R94	0
R 95 Other service industry	R95	0
(Number of establishments By employee size)	1 to 4 people	0
A to R all industries (1)	5-9 people	0
	10 to 19 people	0
(Number of establishments By employee size)	20 to 29 people	0
A to R all industries (2)	30 - 49 people	0
	50 to 99 people	0
(Number of establishments By employee size)	100 to 299 people	0
A to R all industries (3)	Over 300 people	0
	(Repeat) Over 100 people	0
(Number of establishments By employee size)	Less than 20 people	0
A to R all industries	20 or more	0
(Number of establishments By employee size)	Less than 20 people	0
C Mining, quarrying, gravel sampling	20 or more	0.0541
(Number of establishments By employee size)	Less than 20 people	0

D Construction industry	20 or more	0
(Number of establishments By employee size)	Less than 20 people	0
E Manufacturing industry	20 or more	0.0448
(Number of establishments By employee size)	Less than 20 people	0
F Electricity, gas, heat supply, water supply industry	20 or more	0
(Number of establishments By employee size)	Less than 20 people	0
G information communication industry	20 or more	0
(Number of establishments By employee size)	Less than 20 people	0
H Transportation industry, postal service	20 or more	0.0620
(Number of establishments By employee size)	Less than 20 people	0
I Wholesale and retail trade	20 or more	0
(Number of establishments By employee size)	Less than 20 people	0
J Financial industry, the insurance industry	20 or more	0
(Number of establishments By employee size)	Less than 20 people	0
K real estate industry, goods rental business	20 or more	0
(Number of establishments By employee size)	Less than 20 people	0
L academic research	20 or more	0
(Number of establishments By employee size)	Less than 20 people	0
M accommodation industry, food service business	20 or more	0
(Number of establishments By employee size)	Less than 20 people	0
N Living related service industry, the entertainment industry	20 or more	0
(Number of establishments By employee size)	Less than 20 people	0
O education, learning support industry	20 or more	0
(Number of establishments By employee size)	Less than 20 people	0
P Medical, welfare	20 or more	0
(Number of establishments By employee size)	Less than 20 people	0
Q combined service business	20 or more	0.0245
(Number of establishments By employee size)	Less than 20 people	0
R service industry (not classified elsewhere)	20 or more	0
(Number of establishments By employee size)	1 to 4 people	0
E Manufacturing industry (1)	5-9 people	-0.0029
	10 to 19 people	0
(Number of establishments By employee size)	20 to 29 people	0
E Manufacturing industry (2)	30 - 49 people	0
	50 to 99 people	0.0409
(Number of establishments By employee size)	100 to 299 people	0.0070
E Manufacturing industry (3)	Over 300 people	0
	(Repeat) Over 100 people	0.0264
(Number of establishments By employee size)	1 to 4 people	0
I Wholesale and retail trade (1)	5-9 people	0
	10 to 19 people	0
(Number of establishments By employee size)	20 to 29 people	0
I Wholesale and retail trade (2)	30 - 49 people	0
	50 to 99 people	0
(Number of establishments By employee size)	100 to 299 people	0
I Wholesale and retail trade (3)	Over 300 people	0
	(Repeat) Over 100 people	0
(Number of establishments By employee size)	1 to 4 people	0
R service industry (not classified elsewhere) (1)	5-9 people	0
	10 to 19 people	0
(Number of establishments By employee size)	20 to 29 people	0
R service industry (not classified elsewhere) (2)	30 - 49 people	0
	50 to 99 people	0
(Number of establishments By employee size)	100 to 299 people	0
R service industry (not classified elsewhere) (3)	Over 300 people	0

	(Repeat) Over 100 people	0
(Number of establishments by opening time) All industries	Established before 1985	-0.0009
(Number of establishments by opening time) All industries	1985 - 1994 established	0
(Number of establishments by opening time) All industries	Established in 1995 - 2004	0
(Number of establishments by opening time) All industries	Established in 2005-2009	0
(Number of establishments by opening time) All industries	Established after 2010	0
A to R all industries	Total number of workers	0
C to E secondary industry	Total number of workers	0
C05 Mining, quarrying, gravel sampling	Total number of workers	0
D Construction industry	Total number of workers	0
D 06 Comprehensive construction work	Total number of workers	0
D07 Construction work by job (excluding facility construction work)	Total number of workers	0
D08 Equipment construction industry	Total number of workers	0
E Manufacturing industry	Total number of workers	0
E09 Foodstuff manufacturing industry	Total number of workers	0.0001
E10 Beverages · Tobacco · Feed manufacturing industry	Total number of workers	0.0007
E 11 Textile Industry	Total number of workers	0
E12 Wood and wood product manufacturing industry (excluding furniture)	Total number of workers	0.0044
E13 Furniture and accessories manufacturing industry	Total number of workers	0
E14 Manufacturer of pulp, paper, and paper processed goods	Total number of workers	9.2E-04
E15 Printing and related business	Total number of workers	-9.6E-05
E16 Chemical industry	Total number of workers	0
E17 Manufacturing of petroleum products and coal products	Total number of workers	0
E 18 Plastic product manufacturing industry (excluding others)	Total number of workers	0.0003
E19 Rubber product manufacturing industry	Total number of workers	0
E20 Leather, same product, fur manufacturing industry	Total number of workers	0
E21 Ceramic · Soil product manufacturing industry	Total number of workers	0.0010
E22 Iron and Steel Industry	Total number of workers	0
E23 Nonferrous metal manufacturing industry	Total number of workers	0
E24 Metal product manufacturing industry	Total number of workers	0
E25 Manufacturer of machinery and equipment for general-purpose machinery	Total number of workers	0.0002
E26 Machine tool manufacturing industry for the production	Total number of workers	4.3E-05
E 27 Industrial machinery and equipment manufacturing industry	Total number of workers	0
E28 Electronic parts / devices / electronic circuit manufacturing industry	Total number of workers	0
E29 Electrical machinery and equipment manufacturing industry	Total number of workers	-0.0001
E 30 Information and communication machinery and equipment manufacturing industry	Total number of workers	0
E31 Transportation machinery and equipment manufacturing industry	Total number of workers	0.0002
E32 Other manufacturing industry	Total number of workers	0
F to R tertiary industry	Total number of workers	0
F Electricity, gas, heat supply, water supply industry	Total number of workers	0
F33 Electric Industry	Total number of workers	0
F34 gas industry	Total number of workers	0
F35 Heat Supply Industry	Total number of workers	0
F36 water service industry	Total number of workers	0
G information communication industry	Total number of workers	0
G37 Telecommunications industry	Total number of workers	0
G38 Broadcasting industry	Total number of workers	0
G39 Information service industry	Total number of workers	-7.2E-06
G40 Internet accompanying service industry	Total number of workers	0
G41 Video / voice / text information system work	Total number of workers	0
H Transportation industry, postal service	Total number of workers	0
H42 railway industry	Total number of workers	-3.0E-05
H43 road passenger transportation business	Total number of workers	-0.0007
H44 Road Freight Forwarding Industry	Total number of workers	0
H45 water transportation industry	Total number of workers	0

H46 Air Transport Industry	Total number of workers	0
H47 warehouse industry	Total number of workers	0.0005
H48 Service industry incidental to transportation	Total number of workers	0
H49 Postal business (including credit facilities business)	Total number of workers	0
I Wholesale and retail trade	Total number of workers	0
I1 Wholesale trade	Total number of workers	0
I 50 Various goods Wholesale business	Total number of workers	0
I 51 Pharmaceuticals, clothing, etc. wholesale business	Total number of workers	0
I 52 Food and beverage wholesale business	Total number of workers	0
I53 Wholesale of building materials, mineral and metal materials, etc.	Total number of workers	0
I54 Machine tool wholesale business	Total number of workers	-4.8E-05
I55 Other wholesale business	Total number of workers	0
I2 Retail trade	Total number of workers	7.1E-05
I 56 Various product retailers	Total number of workers	0
I561 department store, general supermarket	Total number of workers	0
I569 Various other commodities Retail trade	Total number of workers	0
I 57 cloth, clothes, personal belongings retail	Total number of workers	0
I 58 Retail trade in food and beverage	Total number of workers	0.0001
I 581 Various grocery retailers	Total number of workers	0
I585 liquor retail trade	Total number of workers	0
I59 Machinery and equipment retailing	Total number of workers	0
I60 Other retail business	Total number of workers	0.0003
I 603 Pharmaceuticals and cosmetics retail trade	Total number of workers	0
I606 Book · stationery retailing	Total number of workers	0
J Financial industry, the insurance industry	Total number of workers	0
J62 Banking business	Total number of workers	0
J 622 Bank (excluding Central Bank)	Total number of workers	0
J63 Cooperative organization finance industry	Total number of workers	0
J631 Small business etc. Finance industry	Total number of workers	0
K real estate industry, goods rental business	Total number of workers	0
K68 Real Estate Business	Total number of workers	0
K69 Real estate leasing industry · Management industry	Total number of workers	-0.0003
K 70 Goods rental business	Total number of workers	-0.0006
K701 various goods rental business	Total number of workers	-2.8E-05
K702 Industrial machinery and equipment rental business	Total number of workers	0
K 703 Office equipment rental business	Total number of workers	-0.0008
K 704 Automobile rental business	Total number of workers	-0.0010
K705 Sports / entertainment equipment rental business	Total number of workers	0
K709 Other goods rental business	Total number of workers	0
K7092 Music and video record leasing business (excluding the separate document)	Total number of workers	0
K 7099 Goods not classified as other leasing business	Total number of workers	-0.0040
L academic research, professional and technical service industry	Total number of workers	0
L71 academic and development research institution	Total number of workers	0
L 72 Professional service industry (not classified elsewhere)	Total number of workers	0
L 73 Advertisement	Total number of workers	0
L 74 Technical service industry (not classified elsewhere)	Total number of workers	0
M accommodation industry, food service business	Total number of workers	0
M75 accommodation industry	Total number of workers	0.0001
M751 Ryokan, Hotel	Total number of workers	0
M7591 Company / group accommodation	Total number of workers	0
M76 restaurant	Total number of workers	0
M77 Takeaway / food delivery service	Total number of workers	0
N Living related service industry, the entertainment industry	Total number of workers	0
N 78 Laundry · barber · beauty · bathroom industry	Total number of workers	0
N79 Other life-related services	Total number of workers	0

N80 entertainment industry	Total number of workers	0
N801 Movie Theater	Total number of workers	0
N802 entertainment venue (excluding others), box office	Total number of workers	0
N804 sports facilities offering business	Total number of workers	0
N8041 Sports facilities offering (excluding others)	Total number of workers	0.0006
N8042 gymnasium	Total number of workers	0.0016
N8043 golf course	Total number of workers	-0.0006
N8044 Golf Driving Range	Total number of workers	0
N8045 Bowling alley	Total number of workers	-0.0011
N8046 Tennis court	Total number of workers	0
N8047 Batting / Tennis Practice Area	Total number of workers	0
N8048 fitness club	Total number of workers	-0.0004
N806 game room	Total number of workers	0
N8063 Mahjong club	Total number of workers	0
N8064 Pachinko hall	Total number of workers	0
N8065 game center	Total number of workers	0
N8069 Other playgrounds	Total number of workers	0
O education, learning support industry	Total number of workers	0
O81 School Education	Total number of workers	0
O811 Kindergarten	Total number of workers	0
O 812 elementary school	Total number of workers	0
O 813 Middle School	Total number of workers	0
O 814 High school, Secondary school	Total number of workers	0
O815 special support school	Total number of workers	0
O 816 Higher education institution	Total number of workers	0
O 817 vocational school, various schools	Total number of workers	0
O 82 Other education, learning support industry	Total number of workers	0
O 8213 museum, museum	Total number of workers	0.0001
O8214 Zoo, Botanical Gardens, Aquarium	Total number of workers	0.0028
O823 Learning cram school	Total number of workers	0
O8241 Music teaching work	Total number of workers	0
O 8245 Foreign language conversation teaching work	Total number of workers	0
O 8246 Sports · Health teaching work	Total number of workers	0
P Medical, welfare	Total number of workers	0
P83 Medical service	Total number of workers	0
P831 Hospital	Total number of workers	3.0E-05
P832 general clinic	Total number of workers	0
P833 Dental clinic	Total number of workers	0
P835 Therapeutic business	Total number of workers	0
P84 Health sanitation	Total number of workers	0
P85 Social insurance · social welfare · nursing care business	Total number of workers	0
P853 Child Welfare Project	Total number of workers	0
P8531 nursery school	Total number of workers	0
P8539 Other child welfare business	Total number of workers	0
P854 Welfare and long-term care business for the elderly	Total number of workers	-6.2E-05
P855 Welfare service for people with disabilities	Total number of workers	0
P859 Other social insurance · social welfare · nursing care business	Total number of workers	0
Q combined service business	Total number of workers	0
Q86 Post office	Total number of workers	0
Q 87 Cooperative association (not classified elsewhere)	Total number of workers	0.0002
R service industry (not classified elsewhere)	Total number of workers	0
R88 Waste disposal industry	Total number of workers	0.0002
R89 Automobile maintenance service	Total number of workers	0
R90 Machine etc. Repair work (excluding others)	Total number of workers	0
R91 Employment placement / worker dispatch business	Total number of workers	0

R 911 Employment placement business	Total number of workers	-0.0001
R 92 Other business services	Total number of workers	0
R93 Political, Economic and Cultural Organizations	Total number of workers	0
R933 Academic / Cultural Organization	Total number of workers	0
R 94 religion	Total number of workers	0.0002
R 95 Other service industry	Total number of workers	0
(By employee sizes) A to R whole industry 1 to 4 people	Total number of workers	0
(By employee sizes) A to R all industries 5 to 9 people	Total number of workers	0
(By employee sizes) A to R all industries 10 to 19 people	Total number of workers	0
(By employee sizes) A to R all industries 20 to 29 people	Total number of workers	0
(By employee sizes) A to R all industries 30 to 49 people	Total number of workers	0
(By employee sizes) A to R all industries 50 to 99 people	Total number of workers	0
(By employee sizes) A to R all industries 100 to 299 people	Total number of workers	0
(By employee sizes) A to R All industries over 300 people	Total number of workers	0
(By employee size) A to R all industries	Total number of workers	0
(By employee sizes) A to R all less than 20 people	Total number of workers	0
(By employee size) A to R All industries 20 or more	Total number of workers	0
(By employee sizes) C Mining, quarrying, gravel sampling, -19	Total number of workers	0
(By employee sizes) C Mining, quarrying, gravel sampling 20+	Total number of workers	0
(By employee sizes) D Construction industry less than 20 people	Total number of workers	0
(By employee size) D Construction industry 20 or more	Total number of workers	0
(By employee size) E Manufacturer less than 20 people	Total number of workers	0
(By employee size) E Manufacturing industry 20 or more	Total number of workers	0
(By employee size) F Electricity, gas, heat supply, water supply industry, -19	Total number of workers	0
(By employee size) F Electricity, gas, heat supply, water supply industry, 20+	Total number of workers	0
(By employee size) G Information and telecommunications industry less than 20 people	Total number of workers	0
(By employee size) G Information and telecommunications industry 20 or more	Total number of workers	0
(By employee sizes) H Transportation industry, postal work less than 20 people	Total number of workers	0
(By employee sizes) H Transportation industry, postal business 20 or more	Total number of workers	0
(By employee size) I Wholesale and retailing less than 20 people	Total number of workers	0
(By employee size) I Wholesale and retail business 20 or more	Total number of workers	0
(By employee sizes) J Finance industry, insurance industry Less than 20 people	Total number of workers	0
(By employee sizes) J Finance industry, insurance industry 20 or more	Total number of workers	0
(By employee size) K Real estate industry, goods rental business, -19	Total number of workers	0
(By employee size) K Real estate industry, goods rental business, 20+	Total number of workers	0
(By employee size) L academic research, specialized / technical service industry, -19	Total number of workers	0
(By employee size) L academic research, specialized / technical service industry, 20+	Total number of workers	0
(By employee size) M accommodation industry, food service business, -19	Total number of workers	0
(By employee size) M accommodation industry, food service business, 20+	Total number of workers	0
(By employee size) N Living related service industry, entertainment industry, -19	Total number of workers	0
(By employee size) N Living related service industry, entertainment industry, 20+	Total number of workers	0
(By employee size) O Education, learning support industry, -19	Total number of workers	0
(By employee size) O Education, learning support industry, 20+	Total number of workers	0
(By employee size) P Medical, welfare Less than 20	Total number of workers	0
(By employee size) P Medical, welfare 20 or more	Total number of workers	0
(By employee size) Q combined service business, -19	Total number of workers	0
(By employee size) Q combined service business, 20+	Total number of workers	0
(By employee size) E Manufacturing industry 1 to 4 people	Total number of workers	0
(By employee size) E Manufacturing 5-9 people	Total number of workers	0
(By employee size) E Manufacturing industry 10 to 19 people	Total number of workers	0
(By employee size) E Manufacturing industry 20 to 29 people	Total number of workers	0
(By employee size) E Manufacturing industry 30 to 49 people	Total number of workers	5.6E-05
(By employee size) E Manufacturing industry 50 to 99 people	Total number of workers	0.0002
(By employee size) E Manufacturing industry 100 to 299 people	Total number of workers	0
(By employee size) E Manufacturing industry 300 or more	Total number of workers	0

(By opening time) A to R All industries · Established before 1984	Total number of workers	0
(By opening time) A to R All industries · Established in 1985 and 1994	Total number of workers	0
(By opening time) A to R All industries · Established in 1995 - 2004	Total number of workers	0
(By opening time) A to R all industries · Established 2005 - 2009 years	Total number of workers	0
(By opening time) A to R All industries · Opened after 2010	Total number of workers	0
(Number of companies by capital class)	Less than 3 million yen	0
A to R all industries (1)	Below 300 ~ 5 million yen	0
	Less than 500 ~ 10 million yen	0
(Number of companies by capital class)	Less than 1,000 ~ 30 million yen	0
A to R all industries (2)	Less than 3,000 ~ 50 million yen	0
	Less than 5,000 to 100 million yen	0
(Number of companies by capital class)	Less than 1 billion yen	0
A to R all industries (3)	Less than 10 to 5 billion yen	0
	5 billion yen or more	0
(Number of companies by capital class)	Less than 3 million yen	0
E Manufacturing industry (1)	Below 300 ~ 5 million yen	0
	Less than 500 ~ 10 million yen	0
(Number of companies by capital class)	Less than 1,000 ~ 30 million yen	-0.0107
E Manufacturing industry (2)	Less than 3,000 ~ 50 million yen	-0.0281
	Less than 5,000 to 100 million yen	0
(Number of companies by capital class)	Less than 1 billion yen	0
E Manufacturing industry (3)	Less than 10 to 5 billion yen	0
	5 billion yen or more	-0.0698
(Number of companies by capital class)	Less than 3 million yen	0
I Wholesale and retail trade (1)	Below 300 ~ 5 million yen	0
	Less than 500 ~ 10 million yen	0.0043
(Number of companies by capital class)	Less than 1,000 ~ 30 million yen	0
I Wholesale and retail trade (2)	Less than 3,000 ~ 50 million yen	0
	Less than 5,000 to 100 million yen	-0.0172
(Number of companies by capital class)	Less than 1 billion yen	0
I Wholesale and retail trade (3)	Less than 10 to 5 billion yen	0
	5 billion yen or more	0
(Number of companies by capital class)	Less than 3 million yen	0
R service industry (not classified elsewhere) (1)	Below 300 ~ 5 million yen	0
	Less than 500 ~ 10 million yen	0
(Number of companies by capital class)	Less than 1,000 ~ 30 million yen	0
R service industry (not classified elsewhere) (2)	Less than 3,000 ~ 50 million yen	-0.0024
	Less than 5,000 to 100 million yen	0
(Number of companies by capital class)	Less than 1 billion yen	0
R service industry (not classified elsewhere) (3)	Less than 10 to 5 billion yen	0
	5 billion yen or more	0
Residential.area.km2		1.0374
population		-4.4E-05
num.of.workers		3.2E-07
acc.material		0
acc.assembly		0
acc.HHDgoods		-5.8E-09
acc.manufact		-1.4E-14
acc.market		0
acc.CBD		0
acc.nightPOP		0
acc.goodsSales		0
distance.to.nearest.IC		-1.2E-05
length.of.roads.over.13m.width		1.3E-05
length.of.roads		2.2E-07

num.of.roads.over.13m.width		0.05756
num.of.roads		0.00251
land.price..omit.negative.value.mesh.		0
residencial.area.rate		0.1792
commercial.area.rate		1.2520
quisi.industrial.area.rate		1.6457
Industrial.area.rate		2.5599
Exclusive.industrial.area.rate		3.7407
Urbanization.adjustment.area.rate		0
others.rate		-0.4472
out.of.city.plan.rate		-0.8758
Indiscriminate.place.rate		-0.0122