A Firm Population Synthesis Model for Urban Freight Transport Demand Modeling

A Part of Microsimulation Freight Transport Model for the Province of South Holland

Mesay Shemsu Nasir (4751213)

MSc Thesis Report

Master in Transport Infrastructure and Logistics

Delft University of Technology

A Firm Population Synthesis Model for Urban Freight Transport Demand Modeling

A Part of Microsimulation Freight Transport Model for the Province of South Holland

Author: Mesay Shemsu Nasir
Student Number: 4751213
Delft University of Technology

Master program: Transport Infrastructure and Logistics
Graduation Committee: Lóránt Tavasszy, Chair (TU Delft)
Michiel de Bok (TU Delft/ Significance)
Rob van Nes, (TU Delft)
Bilge Atasoy, (TU Delft)

Preface

This master thesis has been written in fulfillment of the requirements of the program: Master of Science in Transport Infrastructure and Logistics.

Several individuals offered significant and unwavering support to see this project through to completion. First, I would like to acknowledge Prof. Dr. Ir. Loránt Tavasszy for everything ranging from the opportunity to work with him to his enormous help throughout my project. Second, I would also like to acknowledge Dr.ir. Michiel de Bok for his continuous feedback during our progress meetings which helped me synthesize and steer the research work. Finally, I would like to acknowledge my graduation committee members Dr.ir. Rob van Nes and Dr. Bilge Atasoy for their valuable feedback on my work and willingness to answer any question I had.

Amongst my friends and professional associates, I would like to thank Raeed Mohammed a great deal. I would also like to thank Dr. Bikila T. Wodajo from my alma mater who was instrumental in the completion of this project.

I could try to thank my parents Shemsu N. Mohammed and Meskerem A. Hailé, but it is an impossible feat. They are the standards of humanity and altruism. I would also like to thank my siblings Elias and Sarah for their support.

Finally, I thank Him. He makes a path where there is none.

Mesay Shemsu Nasir

Summary

Most of the freight transport models used in practice today are aggregate. The vast array of actors present in the freight transport make modeling it on disaggregate basis a challenging task (Ben-Akiva & de Jong, 2008). An approach to carry out this immense modeling task was outlined by(De Bok et. al., 2018). In this approach, the model will start from observed statistics and incrementally build up to choice models for logistics decisions. Central to this concept, however, are the agents among whom firms are a significant share. However, the synthesis of a firm population in the context of urban freight transport has gotten scant attention in literature.

This research focuses on urban business-to-business freight generation (FG) modeling. It explores the relevant attributes of a firm that can be used to model freight demand and how to model firm agents as part of a synthetic population. Indeed, this is because actual firm population data are protected by privacy laws. The overarching research question encapsulating both the theoretical and practical aspects of the problem is:

How can a synthetic firm population for the Province of South Holland be developed for a microsimulation model of urban freight demand?

Several modeling approaches investigate freight transport demand, each with its own need for inputs that can predict the amount freight generated by a firm (Comi et. al.. 2012). These inputs to FG models are specific data about attributes of individual firms that fit into certain causal and/or correlative mechanisms in the models whereby FG is predicted. Review of literature has been used to compile a set of attributes that are relevant. This list is composed of location, employee size, floor area, commodity type, economic sector and fleet size. The list is a composition of attributes that have demonstrated ability to predict FG in modeling approaches that have varying degrees of explanatory power and degree of transferability. Then, it follows that each attribute is not equally relevant in predicting FG for all firms. This research is based on an explanation of freight demand based on commodities and economic sectors as major predictors while size attributes such as number of employees, floor area and fleet size help explain demand differences within a given sector.

Following the above, economic sectors were used to assess whether a certain firm type is worthy for consideration in population synthesis. Commodity-based FG statistics and input-output tables were used to assess which firm types to consider in the population synthesis. This led to the conclusion that firms of all economic sectors must be synthesized.

The spatial distribution of firms was given special attention. Here also, the type of economic sector of the firm was found to be highly relevant in explaining the firm location. Proximity/accessibility indicators, land rent and agglomeration externalities associated with co-location with other firms are shown to be relevant in linking firm-level economics to the consequent spatial distributions.

To synthesize the firm population, several techniques were explored from literature and compared across different criteria in a qualitative (quasi)multi-criteria selection that led to the choice of the IPU algorithm whose benefits and pitfalls were also discussed.

The FG-oriented definition of firm agents and the chosen population synthesis technique are then used to specify a model and build an implementation architecture according to the model specification. In this respect, a multi-dimensional IPU model was specified with: the input data requirements, the iterative adjustment functions, expected output, the convergence criteria and the goodness-of-fit measure to be adopted.

Once specified, the model was reduced into definite dimensions based on the list of firm attributes for which the required data was available. ***The source of data used in this research was publicly available data offered free of charge by CBS.*** Because this data set is limited, the dimensional width of the model was compromised significantly. The attributes floor area, employment size, commodity type and fleet size had to be left out as a result of data unavailability. Furthermore, the finest location marker available is the neighborhood.

Aggregate and disaggregate data on the economic sector were found in bundles of economic sectors. This has reduced the sectoral dimensions in the specification of the model from 21 SBI2008 sectoral categories to 8 bundles. Therefore, a firm specification had to adapt as:

firm = firm (economic sector bundle, neighborhood)

Firm-level sample data on the attributes which when cross-classified, give proportions that initialize the IPU model, were also unavailable in the data set. Hence, indicators of firm distribution were required. Several possible indicators, namely: urban density, accessibility, land rent and sector-specific factors were explored along with suggestions of possible formulations of the corresponding indicator factors to initialize the IPU model with. After deliberation on availability, urban density was chosen as an indicator.

The above choice comes with two implications. First, it is an indicator variant across zones only and negative correlation with sectors like agriculture make it necessary to find a proper indicator and/or indicator formulation (in this research a reciprocal of urban density) to capture this relationship.

The results of the model revealed a measure of surprising results as urban density showed good correlation for most sectors (bar agriculture) at municipal level and satisfactorily well for service sectors at the neighborhood level. However, detailed interpretations of the distribution of the correlation coefficient revealed more nuances.

The implemented firm synthesis model for South Holland is of certain utility, as it improves on disaggregation level from the COROPS (Davydenko et. al., 2013) used in practice, to firm sector bundles at the neighborhood level; although the former uses commodity types. The model specification remains generic to accept the use of more attributes and more detailed data. This makes the model potent and attractive for further research.

Keywords: ***Firm Population Synthesis, Microsimulation, Urban Freight Demand, Iterative Proportional Updating, South Holland***

Contents

1. Introduction

Even though passenger transport modeling has experienced a shift towards disaggregate modeling to cope with some deficiencies of aggregate models, an analogous shift in the field of freight transport is yet to be realized (Samimi et. al,, 2014). There is added complexity in freight transport because there are multiple actors involved in freight transport (Tavasszy & de Jong, 2014).

Several researchers have tried to render a framework to describe the multi-actor freight transport system. (De Bok et al., 2018) show a framework for freight transport whereby they identify the following actors and the markets they interact in. They specify:

- *Freight generators*: these are the producers and consumers of products
- *Logistics service providers (LSP)*: the companies responsible for the movement of products and
- *Policy makers*: the actors responsible for setting the context in which freight transport occurs.

In the above classification, it can be the case that a freight generator may source its own transportation.

Noteworthy is how firms make up a significant part of the population in the logistics arena with presence in freight generation and freight movement. In this regard, the firm is the basic unit of freight generation (FG) and Freight Trip Generation (FTG), as well as important logistics decisions such as location choice, transport mode choice, shipment size, delivery time, route choice, etc.

This lends itself well to a microsimulation approach with firms as decision making agents. In such a model, each firm will be an agent of some type within the supply chain and its position within the supply chain will determine the type of decisions the firm takes.

Furthermore, the firm population is evidently heterogeneous regarding its intrinsic attributes such as type of economic activity, number of employees, annual turnover, etc. Because the foundation of a microsimulation freight transport model is the line of reasoning that connects the firm's attributes to its freight transport demand, it is of scientific value to ascertain which of these intrinsic attributes of a firm govern its freight transport demand and to what extent.

1.1. Problem Statement

If the intrinsic attributes of a firm related to freight transport demand are known, then they can serve as inputs to models that utilize them to predict freight transport decisions and phenomena at disaggregate level. Using procedures that are mindful of the mathematical means by which the disaggregate results were obtained, they can of course be aggregated to a desired level (Holguı´n-Veras et. al., 2014 in Tavasszy & de Jong, 2014).

Noteworthy is that while different firms may have different attributes affecting their freight transport demand, one universal attribute is location. Location creates the spatial gap between point of making and point of use of a product, that justifies transport. Therefore, in a microsimulation model and for aggregations that follow from it, spatial distribution is of paramount importance. In acknowledgement of this fact, this research pays special attention to the attributes of firm agents as they are spread across a geographical area.

There is scant literature on microsimulation of freight transport. The work of (Samimi et al., 2014) for a national disaggregate freight transport model in the US is worth mentioning. However, as it was a national model, the level of detail in defining the actors was kept to a minimum to ease computational load on the model. Furthermore, a comprehensive microsimulation model for urban freight transport incorporating logistic decisions at the firm level is lacking in the Netherlands.

In view of this gap (De Bok et al., 2018) have developed a multi-stage approach to building a microsimulation freight transport model: the MASS-GT. The approach involves using a descriptive statistics-based setup for the first stage and then building up to choice models for logistics decisions. At the core of this concept lie the actors that generate freight and make decisions on how to transport it, namely: firms. Therefore, to make a potent model, one needs to understand the attributes and the causal and/or correlative mechanisms that relate them to urban freight demand. Hence, the question: ***"What attributes do we need to know about firms in order to define them as agents of a freight transport model?"*** is a scientific gap that needs to be addressed.

Furthermore, there is practical side to the problem. Data protection laws make individual firm data inaccessible. This means that once the logistically important firm attributes are known, ways of synthesizing a fictitious firm population are necessary. ***Hence, another gap in the problem is one of data syntheses regarding firm population for the context of The Netherlands.***

The research on the above two gaps can be important in three ways:

- significance in the immediate context of The Netherlands, i.e. addition to knowledge base.
- significance in building a generic model that can be transferable and
- significance of generated data output for further research

1.2. Research Questions

The research aims to answer the following research questions.

Main question: How can a synthetic firm population for the Province of South Holland be developed for a microsimulation model of urban freight demand?

The above research question has been broken down to the following sub-questions. The sub-questions are given along with the deliverables that will answer them.

1. Conceptual definition:

   What are the required attributes in the definition of a firm agent needed for a microsimulation model? How can these attributes be used to classify firms in the population of producer and consumer firm agents?

2. ***How will the spatial distribution of the firm population be characterized?***
   a. What theory(ies) will be used to locate a firm of specific attribute set?
3. Method:
   a. ***What technique of population synthesis will be selected to synthesize the firm population in the research?***
   b. ***How will a firm synthesis model be specified and implemented?***

1.3. Research Objective and Scope

In order to answer the above research questions, the following set of objectives have been formulated. **General objective:** ***To define and synthesize firm agents for use in agent-based modeling of freight transport.***

To achieve this objective, the research sets the following specific objectives, namely:

1. ***To identify the set of attributes that predict freight transport demand at the firm level.*** The attribute set serves to define the firm as a logistics agent of its own freight transport demand. This is to be accomplished by means of literature review. Achieving this objective answers the first sub question of the research.
2. ***To identify the attributes of a firm related to a firm's location behavior.*** The attributes of a firm that are relevant in explaining a firm's location are described; these explanations will borrow from location theories and empirical studies. In addition, zonal characteristics that attract particular types of firms are also described. This objective is tied to answering sub question 2 of the research.
3. ***To assess population synthesis techniques and choose a technique for synthesizing a population of firms based on the identified attributes.*** Under this objective the research assesses the state of the art in population synthesis. The capabilities and mechanisms of each technique are discussed to meet this objective after which a selection of a synthesis technique will be made for the context of this research. Literature review is used to achieve this objective. Meeting this objective seeks to answer sub question 3(a) of the research.
4. ***To specify a population synthesis model based on the chosen technique.*** This objective will be met by using the findings of the literature review on urban freight transport demand and population synthesis. The model specification will consist of a set of functions, requirements for input data, steps of the implementation algorithm and the type of output expected. Meeting this objective must answer sub question 3(b) of the research.
5. ***To use openly available firm data to synthesize a firm population for the province of South Holland.*** It is evident that firm-level data are private. As such, this objective is met by exploring the openly available data from different sources and devising ways in which open data can be used to meet the requirements set in the model specifications. Meeting this objective aims to answer sub question 3(b) of the research.
6. ***To verify the model with respect to model specifications.*** This objective will be achieved by making a comparative evaluation of the model output specifications to the outputs that were synthesized. This objective is tied to sub question 3(b) of the research as it affirms the delivery of the model as per the intended design.
7. ***To validate the population synthesis model.*** This objective will be achieved by making a quantitative validation of the model outputs against observed data. This objective is tied to the overarching main research question as it affirms the level of success in synthesizing the South Holland firm population.

The research will be limited to the following scope in attempting to meet the above objectives.

➔ ***Geographical limitations***

Although the research aims to build a model founded on transferable methods, the model estimation is carried out on data limited to the geographic boundaries of the province of South Holland.

➔ ***Logistic actors***

The model is scoped to look at firms as logistic actors; it includes the freight transport demands of producers and consumers. Noteworthy is that logistics service providers also have their own freight transport demand and are, thus, producers and consumers themselves. However, the model is limited to production-consumption relations among businesses. Therefore, households which are significant end-consumers of goods are not included in the population synthesis. Second, all firms are synthesized from a freight demand point of view only. This is to mean that all firms have the role of freight generators in the model.

➔ ***Logistics Decisions***

The model looks at interactions between actors at the level of production and consumption networks, i.e., the origin and destination of freight only. The kind of decisions such as shipment size and vehicle type choice, that are related to how freight is transported are not included in the synthesis model. Hence, the list of attributes of each type of firm are also deemed important for this research from the perspective of predicting the firm-level freight transport demand.

➔ ***Firm Location***

Since the intended population synthesis model is to be a part of a microsimulation model for freight transport, the location of a firm is necessary attribute to consider. A very relevant issue here is that of resolution. The finer the resolution of a firm location model, the more accurate it is in capturing the factors that pull the firm to a given location. This involves a clear understanding of not only the set of these factors themselves, but also their explanatory power when it comes to location at a given resolution. In this research, several factors are pointed out, but empirical investigation is limited to urban density.

1.4. Contribution

The contributions of this study are the following:

(a) *Definition of firm agents for microsimulation of urban freight transport demand*

Here a method of literature review on firm-level freight demand was adopted leading to an identification of attributes important for predicting freight demand for firms. A matching of commodity classes (NST2007) and economic sectors (SBI2008) was used to justify the type of firms to be considered in the synthesis while other firm-level attributes were obtained from literature review on freight demand modeling attempts at firm level.

(b) *Proposal of a generic method to synthesize firm population data*

A generic multi-dimensional model based on iterative proportional updating (IPU) was developed to synthesize the firm population. The proposed method contains an undefined number of dimensions which contributes to its transferability.

(c) *Use of the proposed generic method in a case study on the Province of South Holland*

The model was used to synthesize the firm population of South Holland. The generic formulation of the method was reduced to the number of dimensions needed to form a specific model.

(d) *Modeling of the spatial distribution of the firm population in the Province of South Holland*

The proposed model was able to explain the spatial distribution of firms in South Holland at the municipal level.

1.5. Research Approach and Report Outline

The following schematic shows the approach followed in conducting this research. It shows the sequential order of achievement of the research objectives given in the previous section.

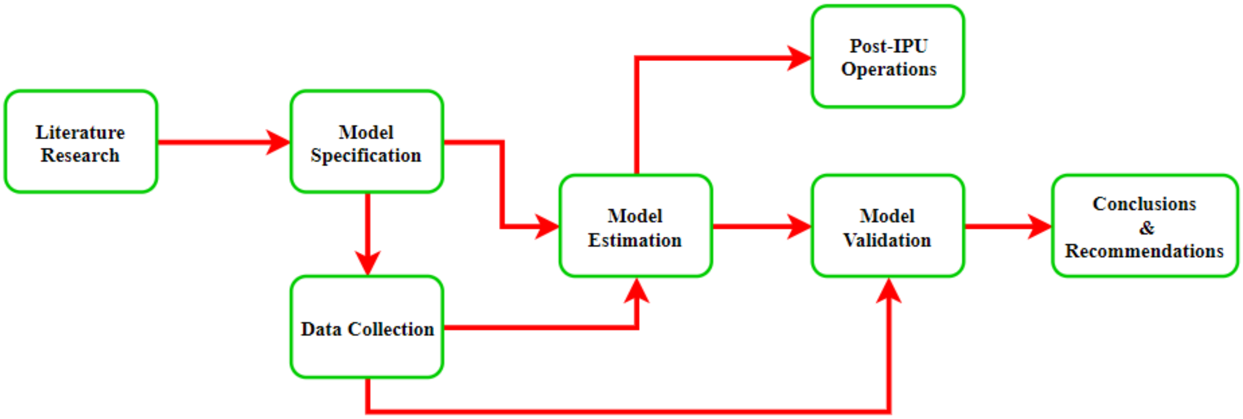


Figure 1: Schematic of the Research Approach

The report to follow will begin by presenting a literature review. This section has two subsections. The first subsection will study Urban freight transport demand and link it to firm population synthesis. This is done by qualifying firms regarding attributes: mainly their economic activities which motivate their freight transport demand. It culminates by building a set of firm-level attributes important for freight transport demand estimation.

The second subsection looks at the techniques of population synthesis currently implemented in research along with their associated merits and pitfalls. It also has a subsection dedicated to firm location in the context of firm synthesis. Ultimately, it makes the choice of a synthesis technique to be utilized in the forthcoming section of model specification.

The model specification section follows with the chosen synthesis technique to formulate the exact functions involved, requirements of inputs and expected outputs.
The case-study section then follows to describe and make an inventory of the available data and to make a comparison to the requirements stated in the model specification. Any changes to the model structure and/or input data as a result of available data are also elaborated on.

The model is then estimated on the available data. Since the data inventory was smaller than the required, an alternative post-IPU method for expanding the model was developed to add one more attribute (employee size) to the model.

The validation section presents the results of the model performance. This section reflects both on the case study and the generic model as they relate to the validation results.

Finally, the conclusions and recommendations summarize the findings and their connections to the research questions that motivated the research. Furthermore, they reflect on the limitations of both the research approach and the synthesis model to make recommendations for future research.

2. Literature Review

2.1. Urban Freight Transport Demand and its Relation to Firm Population Synthesis

The objective of this section is to define the firm as a logistics agent. It first shows that there are different types of actors, each with their own freight transport demands, amongst which firms can be classified. It further classifies these actor-type categories into the type of economic activities they are related to in order to refine the type of freight transport demand they have. Finally, it looks at previous work on what firm-level attributes have been found as justifiable predictors of said freight transport demand.

Urban freight transport is a complex phenomenon involving many actors. The number and type of actors involved is attached to the modeling framework being used to investigate the freight transport process. Below, Figure 2 shows a modeling framework for freight transport.

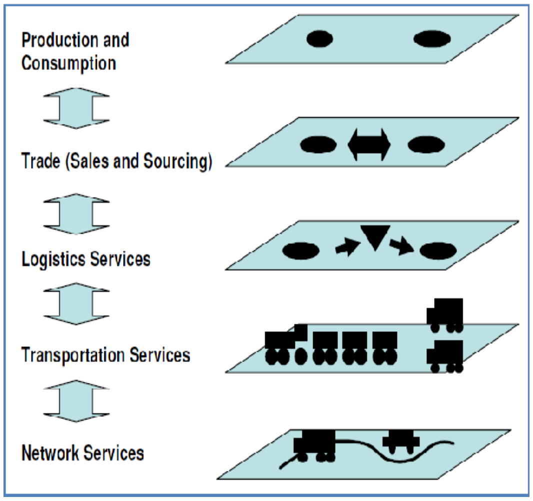


Figure 2: Freight transportation modeling frameworks (Tavasszy, 2006)

Given the above framework, actors with distinct roles in the freight transport process can be identified. For instance, (De Bok & Tavasszy, 2018) propose a multi-actor framework for freight transport whereby they identify the following actors and the markets they interact in.

1. Freight Generators: These are producers and consumers of products and interact with each other within the commodity market.
    a. Producers (Shippers): This is collective nomenclature to represent the supply end of the logistics chain; it can consist of manufacturers, retailers and wholesalers. Shippers are majorly interested in moving their goods via reliable transport services. These services are often obtained from hired third parties (integrators/carriers).
    b. Consumers are responsible for making an order of a certain size with an objective of getting the desired order at the desired time, i.e., fidelity and reliability. These entities can include

both the end users of products and producers that need intermediate products as raw materials to produce finished products. The decisions of these actors include sourcing, amount of product, the delivery deadline and decision on whether to outsource transportation services. Apart from paying for transport costs (included in the purchase price), the consumer is not responsible for freight transport.

Producers and consumers interact with each other within the commodity market. The prices in the commodity market reflect the value of a product and its transportation to a desired location. For the purposes of this research, we shall limit the freight generators to businesses excluding households.

2. Logistics Service Providers (LSP): These are companies responsible for the movement of products from production sites to consumption sites. They can be specialized third party services (3PLs) or own account logistics services (the producer itself moves the products). These actors interact with producers and consumers in the transport market. In this market the prices reflect the value of transporting goods to the consumer. On the other hand, LSPs interact with producers in the logistics market where the prices therein reflect the value of setting up and effecting a plan of delivery of goods from the producer to the consumer.
3. Policy makers: These are the actors responsible for setting the context in which freight transport occurs. They make regulations that constrain logistics choices and decisions. Their interaction is with LSPs within the infrastructure market. Prices in this market reflect the value of the levels of traffic services on the available infrastructure along with externalities.

The above actors are players in an environment that is fundamentally driven by motivations of economic prosperity (*Holguín*-Veras et. al in (L. Tavasszy & de Jong, 2014). This means that logistics needs and decisions are closely tied to the goods and services produced at firm level. It is, therefore, of interest to investigate how economic activity is related to freight transport demands. A way to do this is to further qualify their economic function by looking at their production activities. With relevance to this study, focus will be given to the classification of economic activities according to the system in The Netherlands known as SBI2008 (Standaard Bedrijfsindeling 2008).

In the above classification, there are 21 major economic activities which are further subdivided by sub-codes to specialized economic activities as well. This research, however, looks at the major uni-letter codes, namely:

- Agriculture, forestry and fishing (A),
- Mining and quarrying (B),
- Manufacturing (C),
- Electricity, gas, steam and air conditioning supply (D),
- Water supply; sewerage, waste management and remediation activities (E),
- Construction (F),
- Wholesale and retail trade; repair of motor vehicles and motorcycles (G),
- Transportation and storage (H),
- Accommodation and food service activities (I),
- Information and communication (J),
- Financial institutions (K),

- Renting, buying and selling of real estate (L),
- Consultancy, research and other specialized business services (M),
- Renting and leasing of tangible goods and other business support services (N),
- Public administration, public services and compulsory social security (O),
- Education (P),
- Human health and social work activities (Q),
- Culture, sports and recreation (R),
- Other service activities (S),
- Activities of households as employers; undifferentiated goods and service-producing activities of households for own use (T). Sectors S and T have been coupled as Miscellaneous Services.
- Extraterritorial organizations and bodies (U).

The above sectors produce specific commodities that have their own characteristics. While it is clear from the nomenclature what general category of activity each sector is involved in, there needs to be a systematic classification of the goods or services produced by each sector. This is particularly important to explain freight transport demand. There are many schemes to classify these commodities. A notable one is the NST 2007 shown in Appendix-A.

It is interesting to match the NST 2007 commodities (page 64) to the SBI 2008 sectors (page63). This matching has been done in Appendix-A (page 63). In doing so, one sees that the manufacturing sector, which is associated with most of the commodity groups, takes almost half of the share of the freight transport demand with construction, mining and agriculture being important sectors as well (see table-1). An equally important consideration in freight generation is the split of this very concept into freight production and freight attraction. Indeed, the origin-destination network of economic sectors is what motivates the transportation demand for freight. Mindful of this, the author has attached two extra columns to the data table in Appendix A to specify the possible origin and destination sectors of each commodity category.

Table-1 gives a summary of the origin-destination matching between commodities and sectors. Here, one observes that although from an aggregate perspective a few sectors can cover much of the demand, the full origin-destination network of economic sectors requires consideration. A good example of this can be the construction sector which attracts significant proportion (10.62%, see attraction end column) of freight whilst not being a producer of any commodity category. Here, one observes that although from an aggregate perspective a few sectors can cover much of the demand, the full origin-destination network of economic sectors requires consideration.

Table 1: Summary of Freight Origin-Destination Table (obtained by sector-commodity matching of (Eurostat, 2015) data)

SBI2008 Sector Category (Production End)	Freight Volume(1000tons)	Percentage Share	SBI2008 Sector Category (Attraction End)
A - Agriculture, forestry and fishing	57,714	10.10%	Several: C, G, I
B - Mining and quarrying	90,379	15.82%	Several: B, C, D
C - Manufacturing	283217	49.58%	All Sectors (Construction dominates with 10.62%; the transport and storage sector is worth mentioning at 6.19%)
E - Water supply; sewerage, waste management and remediation activities	43941	7.69%	C, E

This argument is more solidified by looking at data from OCED, (2015) that provides monetary values of the transactions between economic sectors via input-output tables. These data (see Appendix-B) are by no means direct measures of the tonnage of freight generated by firms. However, they can lend some insight regarding firm-to-firm interactions within the population. At first glance, the dominant intra-sector transactions are evident in top freight generators such as manufacturing, construction, transportation and storage and IT, suggesting possible chains of complementary sub-specializations within sectors. However, one also observes both dominant and significant shares of transactions distributed outside the diagonal elements of the table.

Having justified that the whole spectrum of sectors must be included in the synthesis model, we can now proceed to see what intrinsic attributes at firm level can serve as viable predictors of freight transport demand. At this juncture, we must further qualify what we mean by freight transport demand. Thus far, freight transport demand had been expressed as the amount of freight (tonnage) attracted to or produced by economic activity. This is properly termed as freight generation (FG). One can also express freight transport demand as the number of deliveries required to move the attracted or produced freight - appropriately termed as freight trip generation (FTG). FG and FTG are distinct concepts, as the latter is dependent on shipment size, vehicle type and delivery frequency decisions at the firm level (*Holguín*-Veras et. al in (L. Tavasszy & de Jong, 2014).

There have been several attempts to identify relevant attributes to model freight demand at firm level. (Novak et. al., 2011) determined that the size of firms as measured by the number of employees is a significant predictor for FG. (Bastida & Holguín-Veras, 2009) found commodity type and economic sector (industry segment) to be significant in estimating freight generation for shippers and receivers (producers and consumers in the wording of this research).(Park, et. al, 2015) suggested number of employees, revenue and floor area as potent predictors of urban freight generation in Korea, across all sectors of the SBI 2008 mentioned above.

As regards FTG, (Iding, & Tavasszy, 2002) found that firm size as measured by the floor area and the number of employees can be used in a regression model to predict FTG. (de Oliveira et. al., 2017) calibrated establishment-level regression models for pubs and bars in Belo Horizonte (Brazil) using the number of employees, the floor area and the number of operational days as explanatory variables. In both cases the explanatory variables were found to have positive correlations with FTG.

For their activity-based freight demand model, (Samimi et al., 2014) mention annual turnover of firms as important information to predict freight transport demand. (Eastman, 1980), reviewed early attempts to model freight demand at the firm level. These attempts required as data points the floor area, fleet size, turnover and employee size.

At this point, it is necessary to comment on the different attributes explored and their explanatory power of FG and FTG. Meaning, the above literature can be used as indications to compile an all-encompassing list of data points that can potentially predict FG/FTG at the firm level. However, the referred literature does not suggest all attributes are equally relevant across industry segments (and implicitly logistic actor roles). Furthermore, the attributes can be used in different modeling approaches with varying degrees of capturing the freight transport process which in turn make them suitable to inform policies with different planning horizons (Comi, 2012). Therefore, the match between the models and the type of policy decisions they are intended to inform is what truly constitutes the utility of having detailed firm-level data.

Hence, we can summarize this subchapter by defining the firm concisely with a pool of attributes that can predict freight transport demand as described in the equation below. An agent-based model of freight transport demand must then be based on individual data points of the attributes in the equation.

firm = firm (location, commodity type, economic sector, employee size, floor area, annual turnover, fleet size)

One notes above that location is added to the attribute list. This makes sense because, if this model is to be integrated into a larger freight transport model (MASS-GT), a spatial marker is necessary in attaching freight demand to the transport network.

2.2. Firm Population Synthesis

The objective of this subchapter is to look at the state of the art in population synthesis techniques and to help answer research question 3(a) regarding which population synthesis technique to use for the purpose of this research. It also gives some theoretical underpinning to how firm location is considered in firm population synthesis.

As mentioned in chapter one the research considers urban freight demand at firm level for business-to-business freight transport modeling scope. Population synthesis methods have seen widespread use for passenger transport (Ma, Mitchell, & Heppenstall, 2015), land-use and transport interaction (LUTI) models (Gerber et. al., 2018) and population geography(Lomax & Norman, 2016) that need synthetic population data. Previous attempts at firm population synthesis attempts include works by (Ryan, Maoh, & Kanaroglou, 2009), (Samimi et al., 2014) and (Abed et al., 2014). Each of these differ from the current research in terms of spatial coverage (national & regional), the size of the population synthesis problem and the firm agent definition. The synthesis of firms for the purpose of urban freight demand modeling is scant in literature.

2.2.1. Population synthesis techniques

This subchapter intends to explain the details of existing population synthesis techniques and eventually to select the method to be used in the model developed in this research.

Often, geographical areas are divided by administrations based on statistical data collection zones. These zones vary in hierarchy that mostly depends on the size of the geographical area covered. It is a fundamental premise of population synthesis that values for a lower-level geographical area, must aggregate to give values for a higher-level geographical area. The task of synthesizing a population is, thus, generating individuals whose classification according to designated attributes adds up to aggregate numbers observed for the population regarding said attributes.

Before going to describe available population synthesis techniques, we observe two commonalities that exist throughout literature. The first is that all the forthcoming techniques require observed values on attributes based on which a population is to be synthesized. These serve as constraints to which synthesis models stick to. Second is data on population heterogeneity across attribute classes. All the techniques forthcoming, utilize some way to "learn" the distributions that exist in the population. The most common way to know these distributions is to take a representative sample of the population itself. Assuming these data requirements can be fulfilled, we explore the following techniques.

Iterative Proportional Updating (IPU)

The first one is iterative proportional updating (IPU)(Axhausen & Müller, 2010). This is fundamentally a scaling algorithm that attempts to find the perfect set of factors with which to multiply a given sample from the population such that the aggregate values of the synthesized and the observed population are equal. In this method, the sample data will be used to make cross-classification tables for a set (usually a pair) of attributes: hence, a joint-distribution. The joint-distribution will then be scaled up linearly to fit aggregate constraints. This makes the synthesis heavily reliant on the accuracy of the sample data and can lack in terms of capturing true heterogeneity in the population (Farooq et. al., 2013).However, it is quite practical and easy, especially when dealing with discrete distributions.

Among several, (Fienberg, 1970), (Macgill, 1977) and (Ruschendorf, 1995) have done notable work in proving the convergence of the IPU process and the existence of a unique solution for it. Notable advantages that make IPU popular are its ease for implementation for a comparable performance with other methods (Ryan et. al., 2009).

The following are limitations of the IPU method.

1. *The zero-cell problem.* As a scaling method IPU cannot fit a cell entry of zero to any aggregate value. (Xin Ye et. al, 2009) note that inputting a small number to initialize the algorithm can lead to arbitrary bias. To ensure the number is small enough to avoid bias, they recommend 1 divided by the total population. While the spirit of this workaround is benign the population size to be synthesized then matters. Meanwhile Choupani & Mamdoohi, (2016), suggest aggregating adjacent attribute classes to deal with this problem. This can pose two problems: first is that the results that are obtained from bundled attribute classes need a method to unbundle after the completion of IPU. Second, the aggregation of attribute classes must be logical (e.g. there is no sense in aggregating male and female classes if the gender attribute is meant to be classified across this dichotomy).

2. *Non-matching constraints.* Often, there can be situations where IPU-based population syntheses must be carried out using data sources that offer aggregate constraints that do not match. The more attributes of classification exist, the more problematic this becomes especially with forecasting purposes in mind (Rich & Mulalic, 2012). A simple way to mitigate this problem can be to assign importance levels for attributes so that targets are adjusted to the important ranked values.
3. *Non-integer final outcomes.* This is another pitfall and can have significant effects if the non-integer outcome represents a cell of small value (hence a less-common attribute or instance) where a simple rounding up or down can have significant impact. Beckman et. al., (1996) used the post fitting process to mitigate this problem, whereby known individuals from a sample are allocated to the cross-classification tables. The allocation can treat the observed table values (integer or not) as probabilities to draw from; these probabilities can be used to perform simulations to render a population.
4. *Modifiable attribute cell problem.* (Otani et. al., 2012) point out an interesting aspect of the IPU algorithm that is a consequence of cells being organized in discrete classes. The way cross-tables are organized can lead to differently fitting synthetic populations – a problem called the Modifiable attribute cell problem. The best organization then is the one that minimizes the number of cells along the most important attribute. The computational complexity of such a problem warrants genetic algorithms to be employed (Otani et. al., 2012).

IPU has one aspect that is both fundamental to its working and demanding in practice. That is the level of detail in sample data requirements. To elaborate, a generic IPU algorithm (see formulation in chapter 3) works by setting up a cross table for N attributes with values categorized under a finite number of discrete classes. Then the sample data are used to fill in the cross table while known aggregate totals of the observed population for each of the discrete classes are set up in the marginals of the table. Example tables of both the input data and the IPU cross table are is shown below.

Table 2: Initial sample data on firms

Firm ID	Firm Size (No. Employees)	City
001	15	City-1
002	59	City-3
…	…	…
250	75	City-2

Table 3: Example of an IPU cross table

Firm Size (No. Employees)	City			Totals
	City-1	City-2	City-3	
0 – 20				15,500
20 – 60				10,500
60 - 150				4,000
Totals	10,000	15,000	5,000	

This means that to initialize such a table, we need data detailed enough to be classified across each attribute class. This leaves us with the following two scenarios.

i. Required level of detail is available. If data on the values of all attributes for a sample of agents are present, we can classify each agent in the cross table and summarize the aggregate numbers into initial joint distributions between the attributes. Abed et al., (2014) for instance, used the Belfirst, a Belgian database that lists firms by the attributes of: economic sectors, address, firm size, and annual turnover. From the database samples of firms were drawn. With the detailed information available, they could categorize every firm in the sample into a cross-classification table made of the above attributes in order to initialize their IPU-based model. Similarly, Ryan et al., (2009) used a firm population database of 11,499 firms in Hamilton, Ontario for the year 1990 in one of their models based on IPU; the population of firms was categorized by firm size (by employees), geographic location and a 3-digit classification of industry segments.

ii. Required level of detail is not available. In such a case, IPU is not capable to synthesize the population based solely on aggregate constraints. For an IPU-based model, the best possible solution in this case is to find a proxy variable (indicator) that closely matches the distribution of the firm population and use it to initialize the model.

Therefore, that ideal firm level data we require for IPU for the attributes concerned with this research would look something like below; this would be required for a sample of firms. Abed et al., (2014) tested the goodness of fit between the synthesized data and the observed population for sample sizes of 10%, 30 % and 60% of the population, with the last proving the best fit.

Table 4: Fields of the ideal firm-level sample data

Firm ID	Location	Commodity Type	Economic Sector	Size (employees)	Floor Area (Sq.meters)	Annual Turnover
001						
002						
003						
004						

After a look at these detailed data and with the level of required spatial, commodity or economic sector-based aggregations needed for the FG model intended, we decide on the number of attribute classes for each attribute (e.g. Firm Size [employees]: 0 - 20, 20 - 60, 60 - 150). The modifiable attribute cell problem is relevant here. After a decision on the attribute classes, we need observed population totals along those classifications per attribute. And only this set completes the ideal requirements of an IPU-based firm synthesis model. A consequent question is, then, *what if these levels of detailed data are not available?* This question will be commented on at the end of the next subchapter explaining spatial distribution of firms.

To produce the cell counts within the population, IPU works by iteratively scaling the cell values from the sample in order that their totals match the population totals in the marginals of the table. At this juncture, we note that IPU produces the cell counts within the cross tables. This means, for all practical purposes, the members of a cell in the cross-table are the same agents; therefore, the synthesized data are anonymous. However, if the produced counts are to be replaced by samples of real firms, then the sample used to initialize the IPU can be used again as a pool to draw from. Firms can be drawn repetitively from the sample with Monte-Carlo-based importance samplings. (Axhausen & Zurich, 2010) call this the allocation phase of the IPU.

Combinatorial Optimization

The second method is called combinatorial optimization (CO) (Ryan et al., 2009). While IPU relies on cross classification-based proportions to meet aggregate constraints, CO directly draws from the sample data to fit the population aggregates (Lee & Fu, 2011). Ryan et al., (2009) discuss CO using a numerical example. This numerical example is tailored to this research and elaborated below to enrich the discussion on this method.

Suppose we start from the same data set as in table 2 and table 3. First thing to note is CO uses the same level of detailed input data as IPU for population synthesis. Then CO stats by using the sample of firms to synthesize the populations city-by-city.

Therefore, starting from City-1, 10,000 firms are drawn from the sample. This involves repetition and replacement because the population of City-1 is bigger than the sample. For these 10,000 firms, their distribution of firm size is then evaluated. CO then evaluates whether the match between the observed population and the synthesized population improves by replacing one firm from the synthesized population with one firm from the sample (with replacement). If the match improves, CO retains this new solution. If not, the replacement is discarded.

The same will be done for the 15,000 firms of City-2 and the 5,000 firms of City-3. In such a manner, CO continuously seeks which combination of the sample elements yields the best match to the observed distribution in the population – hence the name combinatorial optimization. This approach by Ryan et al., (2009) explained above has similarity with a brute-force approach to find the solution of an optimization problem with the objective being to minimize the difference between the drawn population and the observed one.

Another problem that CO-based population synthesis is similar to is the bin-packing problem. This problem, (Korf, 2002), involves packing items of certain volume in bins of fixed capacity and utilize the minimum number of bins. Algorithms prepared for solving this problem can also accommodate CO-based population synthesis under certain modifications. One of these modifications would be that the items to be put in a bin (in this case firms) have a size of unity. In addition, the number of bins to be used in a bin-packing problem are minimal and of the same capacity (Alvim et. al., 2001). Meanwhile in CO-based population synthesis, the number of bins (attribute classes) are often determined by the available data and/or the purposes for which the synthesized data are intended (in this case freight demand modeling).

There can be different types of algorithms that can be deployed for optimization ranging from brute force (Ryan et. al.,2009) to simulated annealing (Harland et. al., 2012) and genetic algorithms (Lee & Fu, 2011). All three works agree on the high performance of CO but maintain that CO's performance is not satisfactory in large populations. (Pritchard & Miller, 2012) add to this that CO can fail to reflect the patterns within the initial sample and instead overfits to totals.

Simulation Methods

A separate group are simulation methods. These methods are radically different from the two techniques described above in that they do not use one sample data to detect the joint distribution in the population (Farooq et al., 2013). Instead, keeping the observed aggregate totals as constraints, simulation approaches use specific sampling techniques which generate distributions to draw from. These distributions are based on parameters estimated by combining data at several degrees of aggregation and from different sources. (Hastings, 1970) describes how a basic Monte Carlo sampling algorithm can be modified into dependent random sampling sequences. Advantages of such methods are:

- ability to handle both discrete and continuous attributes (Farooq et al., 2013).
- better catering for multi-dimensional distributions (Moeckel et. al., 2003).
- ability to offset the weakness of IPU method for the zero-cell problem (Farooq et al., 2013).
- better performance than both of the preceding techniques for the same amount of data (Farooq et al., 2013)

Choice of Technique

To choose a population synthesis technique from the above, there first need to be reliable metrics. These metrics are briefly discussed below.

1. **Data availability:** This criterion is related to the availability of the required amount and level of detail of data needed as an input to the model.
2. **Performance:** This criterion measures how accurately a given population synthesis technique has been proven to match observed populations; it is based on literature review of previous work. As much as possible, the evaluation using this criterion adhered to the basic forms of the techniques as opposed to special variants that may not necessarily be reflective of the general category of the techniques.
3. **Computation Time:** Mathematical techniques (analytical or numerical) need a certain amount of run time on computers to render a solution. This time from start of model run to the stop (which could be marked by an inherent process in the technique or a set of stoppage criteria) is a crucial metric to evaluate model efficiency.
4. **Problem (Population) Size:** The population synthesis techniques described above vary by size of the population they can synthesize with a designated level of satisfactory accuracy. This is relevant both in terms of applying the intended model to the case study in this research and to the transferability of the generic model.
5. **Ease of Implementation:** This is an important criterion regarding the time, skills and resources of the project being undertaken; the trade-off between this criterion and the above four represents the practical realization of the project (Lomax & Norman, 2016).

The following table shows a four-scale color rating of the above population synthesis methods to motivate the selection of a population synthesis technique for this research.

Table 5: Evaluation of Synthesis Techniques

	Population Synthesis Techniques			
Evaluation Criteria	Iterative Proportional Updating	Combinatorial Optimization	Simulation-Based Synthesis	Corroboration
Availability of Detailed Data	🟥	🟥	🟩	(Ryan et al., 2009)(Abed et al., 2014)
Performance	🟨	🟨	🟩	(Ryan et al., 2009), (Farooq et al., 2013), (Axhausen & Zurich, 2010)
Computation Time	🟩	🟥	🟨	(Harland et al., 2012), (Ryan et al., 2009), (Farooq et al., 2013)
Problem (population) Size	🟩	🟥	🟩	(Harland et al., 2012), (Ryan et al., 2009), (Farooq et al., 2013)
Ease of Implementation	🟩	🟧	🟨	
Key: Color codes, as seen below, indicate improving ratings from left to right				
🟥	🟧	🟨	🟩	

From the above table IPU and simulation-based methods are found attractive. However, IPU combines satisfactory performance, reasonable convergence time and relatively less involved implementation. Therefore, IPU is the technique of choice in this research.

Performance Evaluation

Having chosen the population synthesis technique, we now take a brief look at the performance evaluation statistics that have been used for population synthesis techniques across literature. This overview is more focused on connecting these statistics to the firm synthesis model and less on a detailed comparative research into the synthesis techniques.

Total Absolute Error

(Harland et al., 2012) have used this evaluation criteria; it measures the difference between the total counts of the observed table and the total counts of the synthetic table.

$$TAE = \sum_{I_1}\sum_{I_2}\cdots\sum_{I_N}\left|\boldsymbol{M_{i_1,i_2,\ldots,i_N}}^{\boldsymbol{S}} - \boldsymbol{M_{i_1,i_2,\ldots,i_N}}^{\boldsymbol{O}}\right|$$

Where I_i are indices for the N-dimensional of the IPU table while $M^S_{I_1,I_2,\ldots,I_N}$ and $M^O_{I_1,I_2,\ldots,I_N}$are the synthesized and observed entries of the table respectively. This measure doesn't anything about how the errors are distributed across the table. Nor is it fit to compare tables of different size.

Standard Root Mean Squared Error (SRMSE)

(Farooq et al., 2013), (Axhausen & Zurich, 2010) and (Pritchard & Miller, 2012) have used this measure. This parameter measures the difference between synthetic distribution and the observed distribution. (Pitfield, 1978) has given the formula as:

$$\text{SRMSE} = \frac{\sqrt{\sum \frac{(\boldsymbol{M_{i_1,i_2,\dots,i_N}}^{\boldsymbol{O}} - \boldsymbol{M_{i_1,i_2,\dots,i_N}}^{\boldsymbol{S}})^2}{\boldsymbol{\varphi}}}}{\sum \frac{\boldsymbol{M_{i_1,i_2,\dots,i_N}}^{\boldsymbol{O}}}{\boldsymbol{\varphi}}}$$

where:

- i_I are attribute classes of N attributes in the population
- $\boldsymbol{M_{i_1,i_2,\dots,i_N}}^{\boldsymbol{O}}$ *and* $\boldsymbol{M_{i_1,i_2,\dots,i_N}}^{\boldsymbol{S}}$ represent observed and synthesized values of the cross tables that follow either the IPU or the simulation synthesis procedures, and
- $\boldsymbol{\varphi}$ *is* the number of corresponding observed and synthesized attribute values respectively.

Based on the formula above we can see that:

- A value of zero is the best fit indicator based on SRMSE.
- The computation can suffer from zero and very small values of the entries for the observed data. Such potentially problematic values are common in IPU however (see IPU pitfalls 1 & 2 above); this requires an exploration of a better parameter.

Coefficient of Determination: R-Squared ($\boldsymbol{R^2}$)

This measure gives the covariance in the synthetic data as a proportion of the covariance in the observed population data. It indicates the utility of the synthesis model in relation to using the covariance in the population (if it were possible). It is taken as the proportion of the explained covariance in using the model (Bartels, 2015).

$$\boldsymbol{R^2} = \frac{\boldsymbol{Variance\ Explained\ by\ Model}}{\boldsymbol{Total\ Variance}}$$

$$\boldsymbol{R^2} = \frac{[n \sum M^S_{I_1,I_2,\dots,I_N} M^O_{I_1,I_2,\dots,I_N} - \sum M^S_{I_1,I_2,\dots,I_N} \sum M^O_{I_1,I_2,\dots,I_N}]^2}{[\boldsymbol{n} \sum M^S_{I_1,I_2,\dots,I_N}{}^{\boldsymbol{2}} - \left(\sum M^S_{I_1,I_2,\dots,I_N}\right)^{\boldsymbol{2}}][\boldsymbol{n} \sum M^O_{I_1,I_2,\dots,I_N}{}^{\boldsymbol{2}} - \left(\sum M^O_{I_1,I_2,\dots,I_N}\right)^{\boldsymbol{2}}]}$$

Where I_i are indices for the N-dimensional of the IPU table while n is the number of corresponding cell entries.

(Farooq et al., 2013) have also used this parameter in conjunction with the SRMSE. This parameter is less likely to face the zero-denominator problem faced by the SRMSE measure. As such, it is the fit measurement of choice in this research.

2.2.2. Firm location in firm synthesis

In this subchapter, spatial distribution is emphasized to attach the firm agents to the physical world in the context of a transportation problem. Furthermore, the discussion will look at spatial distribution as one part of the issue of joint distribution amongst firm attributes; hence, it ends by answering the question of possible indicators to initialize the intended IPU table with in the event of lacking data.

Since transport demand is motivated by spatial separation between the making and usage points of products, location is an important attribute to look at. Given the freight demand predictor variables as number of employees, economic sector and floor area, it is then important to know how values of these predictors (or agents that possess them) are distributed across space.

In the firm synthesis model of this research, location will be looked at in terms of administrative hierarchical levels. The considered levels in The Netherlands are in descending order of geographical size: Nation, province, municipality, neighborhood and address. The postal codes address (PC-6) level data in the Netherlands context can also be a reasonable substitute for exact firm address from a freight transport perspective.

For freight demand modeling at firm level, it is important to locate the firm as precisely as possible because it is the most logical way of knowing the ingress and egress points at which the demanded transport will join the infrastructure network.

Regarding how firms decide to locate, a detailed behavioral model will need to be calibrated to understand location behavior of firms. Nonetheless, valuable insight can be gained from literature which can be adopted to suit this research.

Classical location theories by (Weber & Friedrich, 1929) and (Moses, 1958) classify economic activities in two broad segments as resource-oriented and market-oriented. The former are located close to resources that are essential for their businesses (e.g. agriculture and mining), while the latter are located close to their customer base (e.g. retail). These can serve as more categorical and aggregate indicators of how firms locate.

Delving deeper into disaggregate studies, (De Bok & Van Oort, 2011) posit that in general, rural areas cannot support many firms because they are unable to provide the necessary resource base as compared to urban areas. Hence, most firms tend to locate in urban areas. In his book Urban Economics, (O'Sullivan, 2003) dissects the forces that bring about urban clustering of firms as localization and urbanization economies. He states that localization economy, which is intra-sectoral clustering, of firms happens due to economies of scale and sharing of intermediate inputs by firms. These mono-sectoral clusterings go further to attract cross-sectoral clusterings due to sharing of common inputs such as transportation or financial services. Both the within-sector and cross-sector benefits of agglomeration have been theorized in the works of (Marshall, 1920), (Jacobs, 1969) and (Van Der Panne & Van Beers, 2006).

But what is the implication of this regarding the spatial structure of the urban environment? At this juncture, one can postulate that the firm-supporting capacity of an urban environment can be measured by urban density. Furthermore, as labor is a common input contributing to both Marshallian and Jacobian externalities, measuring urban density using population density is warranted.

The role of accessibility in firm location is also important. (De Bok & Van Oort, 2011) found accessibility (indicated by proximity to transport infrastructure such as train stations or highway entries) to be significant at 95% confidence on relocation choice of firms in the business services sector. The choice parameters for other sectors were not as significant.

We now return to the question raised in the previous subchapter regarding the initialization of an IPU table with indicators. Table-2 provides possible indicators for each of the firm attributes above and the reasoning associated.

From the table, we note the following. First, the utility of each indicator may not be equally adequate in all scales of space or any other attribute. Second, indicators that are adequate for an attribute need to be combined in a function with indicators of other attributes for a correct formulation of an indicator in a joint distribution to initialize the IPU table. This is the added challenge in the event of lack of sample data.

Table 6: Possible indicators for distribution of firm attributes

Firm Attribute	Indicator	Premise/Reasoning
Location	Urban density	Population density can encourage or discourage different firm types
	Accessibility	Access to transportation means an attractive location for potential employees and easier means to transport products, also benefits of agglomeration
	Resource locations	Some firm activities are tied to essential resources
Size (employees)	Age	Size, age and survivability of firms in Japan have shown positive correlation (Yasuda, 2005); (Cabral & Mata, 2003) found similar correlation between size and age for firms in Portugal; the size and age of firms are correlated (Brouwer et. al., 2002)
Floor Area (Sq. meters)	Land rents	In a self-regulatory market, rents govern competition for space; thus, possible negative correlation with floor space per firm
	Available commercial space	In a given geographical area the total available space for an activity determines how many businesses can be sustained
	Urban density	Possible negative correlation with floor area. Mechanism can possibly be via land rents.
Commodity Type	Aggregate economic sector distributions,	There can be aggregate distinctions between firm types
	Resource locations	Some firm activities are tied to essential resources
Economic Sector	Urban density	Population density can be positively correlated as firms seek to utilize labor as a resource
	Regional commodity flows	A matching between commodities and economic sectors can reveal distribution of economic sectors (related to literature review on urban freight demand in Chapter-2)
	Accessibility	In general, accessible locations can be attractive, but there are sectoral variations. Demand-oriented firm types (e.g. a supermarket) are more sensitive to accessibility
Annual Turnover	Age	The growth of a firm correlates to financial robustness
	Land rents	A firm's location in a designated land rent zone can indicate the financial strength of that firm
	Public tax records	Amount of tax is usually correlated to a firm's annual turnover

2.3. Summary

The literature review section has yielded the following important conclusions that serve as a stepping stone for the model specification step.

- Urban freight and the consequent transport demand are generated as a result of basic economic motivations. Hence, the economic sector of a firm and specifically the commodity types it produces or consumes are important attributes in predicting freight transport demand at the firm level.
- The number of employees, floor area, fleet size and annual turnover of the firm are also significant predictors of freight demand for various firms at varying degrees. Therefore, it is worth obtaining firm-level data on these attributes as well.
- Locational difference is the "deterrence" that transportation overcomes in the exchange of goods and services between producers and consumers. It is, therefore, important to know the locations of firms, up to the best possible level of precision, because these are points of ingress/egress for freight into/out of the transport network respectively.

  The above points answer research question 1 stated in section 1.2.
- All the evaluated population synthesis techniques can generate anonymous data for the above pool of firms' attributes: location, employee-size, floor area, commodity type, economic sector, fleet size and annual turnover, as substitutes for private firm data which are normally protected by law.
- The choice of population synthesis techniques depends highly on the data requirements, available data, the expected output and the computational intensity that are involved.
- In events where data requirements of models are not satisfied, substitute indicators can be chosen to replicate the distributions of sample data on original variables. Table-6 gives an overview of possible indicators of firms' attributes related to freight generation.

  The above three points answer research question 2 and 3(a) stated in section 1.2.

3. Model Specification

This section seeks to specify the model by means of joining the literature review on the specification of the firm agents and on population synthesis techniques. It is therefore committed to answering the research questions, 3(a) and 3(b).

3.1. Model Structure

The structure of the intended model stems from the way the firm agent is defined. The literature review concluded with attributes of a firm that were shown to be significant predictors of freight demand at firm level. The commodity type and economic sectors were shown to be important identifiers of firm type while size attributes of employee-size, floor area, annual turnover and fleet size have variant degrees of predictive significance. The firm agent has been specified as:

firm = firm (location, commodity type, economic sector, employee size, floor area, annual turnover, fleet size)

In the above specification, the attributes are a mix of discrete and continuous types. As regards the continuous attributes, we need to classify them into discrete chunks to be able to study significant variations in freight transport demand at the firm level. If we do so, the above specification would describe a firm as being specified by six attributes. Therefore, the entire firm population can be described by a six-way distribution.

We therefore set as an objective of the model that: ***the firm population synthesis model is intended to generate a multi-dimensional distribution of attributes and that it will be based on the IPU algorithm.*** The algorithm is implemented by programming language Python and MS Excel spreadsheets. The following subsection gives a detailed look at the mechanism of the algorithm, the data it requires and the kind of output it produces.

3.2. IPU Algorithm

Here a generic multidimensional algorithm is described. Let N be the number of attributes which form the basis for the model. Each attribute will then have an index $I \in \{1, 2, \dots, N\}$. Let n_I be the number of attribute classes of attribute I. Therefore, $i_I \in \{1, 2, \dots, N_I\}$, i.e., each attribute I is made of $N_I$attribute classes that are counted by the attribute-specific variable i_I. Hence, the N-dimensional cross-table of the IPU model is formed from the column matrices shown below.

$$\begin{array}{cccccccc} I=1 & I=2 & I=3 & & I=I & & I=N & \\ \begin{bmatrix}1\\2\\3\\ \vdots\\ i_1\\ \vdots\\ n_1\end{bmatrix} & \begin{bmatrix}1\\2\\3\\ \vdots\\ i_2\\ \vdots\\ n_2\end{bmatrix} & \begin{bmatrix}1\\2\\3\\ \vdots\\ i_3\\ \vdots\\ n_3\end{bmatrix} & \dots & \begin{bmatrix}1\\2\\3\\ \vdots\\ i_I\\ \vdots\\ n_I\end{bmatrix} & \dots & \begin{bmatrix}1\\2\\3\\ \vdots\\ i_N\\ \vdots\\ n_N\end{bmatrix} & \forall I \in \{1,2,\dots,N\};\ \forall i_I \in \{1,2,\dots,n_I\} \end{array}$$

To generate such a table for the whole firm population, we require the following:

A. *Aggregate data constraints.* These are totals serving as constraints to fit the synthesized data to. They are observed data. They are denoted by T_{i_I}, which are totals for all attribute classes of all attributes.

B. *Prior sample data.* For IPU, a starting sample is necessary to synthesize the population. These samples are necessary to initialize the cross table with the numbers of specific types of firms in a given sample of the N-dimensional joint distribution. Such values are denoted by: $M_{i_1 i_2 \dots i_N}$. In the order they are given in the notation, each of the indices show belongingness to a specific class i_I of a given attribute I. Therefore, when indices i_1 to i_N combine, they show membership in an N-dimensional class. These initial values are then iterated upon to eventually yield the firm population.

Before explaining the procedural workings of IPU, we note one obvious predicament. In terms of visual comprehension, three-dimensional tables are the biggest observable tables and two-dimensional tables are the most visually comprehensible. Therefore, to aid the explanation of the upcoming steps of the algorithm, a two-dimensional table is utilized while the equations for the multidimensional algorithm given are the actual equations of the envisaged model. The iterative steps are enumerated as follows.

1. First the cross table is set up across the attributes and corresponding attribute classes. In the two-dimensional example, these constitute the top and left-hand sides of the cross table.
2. Then elements that signify a joint distribution between the attributes are filled in. These are prior sample data which are the initial cell values that set the distribution pattern of the iterations to come. In table 1 below, they are represented by the values $M_{(I,I)}$ through $M_{(n,n)}$ while for an N-dimensional table, they would be denoted by $M_{i_1=1,i_2=1,\dots,i_N=1}$through $M_{i_1=n_1,i_{2=n_2},\dots,i_N=n_N}$.
3. The marginals of the cross table (shown in table-1 as the right-hand side and the bottom side) are populated by observed aggregate values. These are hereafter referred to as marginals and they are totals per attribute class. They are represented by the values $T_{i_I=1_1}$ through $T_{i_I=n_{N_N}}$ per attribute class (where $n_N$is the number of attribute classes for attribute N)
4. The iterations are then as follows
    a. The cell elements $M_{i_1=1,i_2=1,\dots,i_N=1}$through $M_{i_1=n_1,i_{2=n_2},\dots,i_N=n_N}$. are summed along the direction of each attribute class (rows and columns of table-1 in the two-dimensional example). These sums correspond to $Sum_{i_I=1_1}$ through $Sum_{i_I=n_{N_N}}$ per attribute class.

b. Scale factors $F_{i_I=1_1}$ through $F_{i_I=n_{N_N}}$ are computed as quotients between corresponding totals and sums per attribute class (or per row and column in the example in table-1) as shown in the table below.

c. Starting with any one of the attributes (for instance employee size), each cell element $M_{i_1 i_2 ....i_N}$ is then multiplied by the scale factors $F_{i_I=1_1}$ through $F_{i_I=n_{N_N}}$ generated for that attribute class. This gives rise to new sums $Sum_{i_I=1_1}$ *through* $Sum_{i_I=n_{N_N}}$for all the other attributes based on which the operation was not performed. Along with the new sums follow new scale factors $F_{i_I=1_1}$ through $F_{i_I=n_{N_N}}$.

d. This time, we choose another attribute (for instance floor area) and use its scale factors to multiply all the entries with. This gives rise to new sums and scale factors along the directions of the other attributes. In a similar manner, the iterations will loop through all the attributes to scale the sample with corresponding factors.

The iterative process described thus far, can be shown mathematically as the following set of equations. The equations are based on two-dimensional equations by (Bishop et. al., 2007) but modified to suit the multivariate requirements of the envisaged model.

Iterative adjustments:

$$M_{i_1 i_2 ....i_N}{}^{k+1,K} = (\frac{T_{i_1}}{\sum_{i_2=1}^{n_2}\sum_{i_3=1}^{n_3}\cdots\sum_{i_N=1}^{n_N} M_{i_1 i_2 ....i_N}{}^{k,K}})M_{i_1 i_2 ....i_N}{}^{k,K}$$

$$M_{i_1 i_2 ....i_N}{}^{k+2,K} = (\frac{T_{i_2}}{\sum_{i_1=1}^{n_1}\sum_{i_3=1}^{n_3}\cdots\sum_{i_N=1}^{n_N} M_{i_1 i_2 ....i_N}{}^{k+1,K}})M_{i_1 i_2 ....i_N}{}^{k+1,K}$$

$$M_{i_1 i_2 ....i_N}{}^{k+3,K} = (\frac{T_{i_3}}{\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\cdots\sum_{i_N=1}^{n_N} M_{i_1 i_2 ....i_N}{}^{k+2,K}})M_{i_1 i_2 ....i_N}{}^{k+2,K}$$

$$\vdots$$

$$M_{i_1 i_2 ....i_N}{}^{k+N,K} = (\frac{T_{i_N}}{\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\cdots\sum_{i_N=1}^{n} M_{i_1 i_2 ....i_N}{}^{k+N-1,K}})M_{i_1 i_2 ....i_N}{}^{k+N-1,K}$$

Where:

an index $I \in \{1, 2, \dots, N\}$. Let n_I be the number of attribute classes of attribute I. Therefore, $i_I \in \{1, 2, \dots , N_I\}$, i.e., each attribute I is made of $N_I$attribute classes that are counted by the attribute-specific variable i_I.

- $I \in \{1, 2, \dots, N\}$ is an index for the N attributes.
- $i_I \in \{1, 2, \dots , N_I\}$ are indices indicating each attribute class of attribute I.
- $T_{i_I}$are totals along attribute classes i_I of attribute I.
- $k \in \{1, 2, \dots, N\}$ is an index for the number of sub-iterations, i.e., adjustments for each attribute within the main iteration of fixed index K.
- K is an index for main iterations; it counts how many full sets of k sub-iterations have been carried out.

Note that $\sum_{i_1=1}^{n_1} T_{i_1} = \sum_{i_2=1}^{n_2} T_{i_2} = \cdots = \sum_{i_N=1}^{n_N} T_{i_N}$ holds for the cross table.

e. At this juncture, there need to be set some convergence conditions that mark the termination of iterations. We can safely terminate iterations when values of the previous iterations are the same as values for the current iteration. When this happens, one notes that the scaling factors along the direction of all attribute classes assume a value of unity. This means that setting a tolerance limit (α_{tol}) on the differences between a value of one and the scale factors of the current iteration can serve as a suitable set of convergence criteria. These can be given by:

$$\varepsilon_{i_1}{}^{K} = \left| 1 - \frac{T_{i_1}}{\sum_{i_2=1}^{n_2} \sum_{i_3=1}^{n_3} \cdots \sum_{i_N=1}^{n_N} M_{i_1 i_2 \ldots. i_N}{}^{k,K}} \right|$$

$$\varepsilon_{i_2}{}^{K} = \left| 1 - \frac{T_{i_2}}{\sum_{i_1=1}^{n_1} \sum_{i_3=1}^{n_3} \cdots \sum_{i_N=1}^{n_N} M_{i_1 i_2 \ldots. i_N}{}^{k,K}} \right|$$

$$\vdots$$

$$\varepsilon_{i_N}{}^{K} = \left| 1 - \frac{T_{i_N}}{\sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \ldots \sum_{i_N=1}^{n} M_{i_1 i_2 \ldots. i_N}{}^{k,K}} \right|$$

$$\max\left(\varepsilon_{i_1}{}^{K}, \ldots, \varepsilon_{i_N}{}^{K}\right) < \alpha_{tol}$$

We set the value of α_{tol} to 10^{-6} as it is adequate. One notes how the convergence checks are performed for the main iterations.

5. Once the iterations have converged, the synthesized population is described by the final values of the entries $M_{i_1=1,i_2=1,\ldots,i_N=1}$through $M_{i_1=n_1,i_{2=n_2},\ldots,i_N=n_N}$. .

The results can be validated against corresponding observed data. Here, the goodness of fit between the observed and the synthesized population data can be measured by the chosen fit parameter: R-squared, given as:

$$\boldsymbol{R^2} = \frac{\boldsymbol{Variance\ Explained\ by\ Model}}{\boldsymbol{Total\ Variance}}$$

$$\boldsymbol{R^2} = \frac{\left[\boldsymbol{\varphi \sum M_{i_1,i_2,\ldots,i_N}{}^{S} M_{i_1,i_2,\ldots,i_N}{}^{O}} - \sum M_{i_1,i_2,\ldots,i_N}{}^{S} \sum M_{i_1,i_2,\ldots,i_N}{}^{O}\right]^2}{\left[\boldsymbol{\varphi \sum M_{i_1,i_2,\ldots,i_N}{}^{S^2}} - \left(\boldsymbol{M_{i_1,i_2,\ldots,i_N}{}^{S}}\right)^2\right]\left[\boldsymbol{\varphi \sum M_{i_1,i_2,\ldots,i_N}{}^{O^2}} - \left(\boldsymbol{M_{i_1,i_2,\ldots,i_N}{}^{O}}\right)^2\right]}$$

$$\forall\, \boldsymbol{i_1} \in \{\mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{n_1}\}, \boldsymbol{i_2} \in \{\mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{n_2}\}, \ldots, \boldsymbol{i_N} \in \{\mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{n_N}\}$$

Where:

- $\boldsymbol{M_{i_1,i_2,\ldots,i_N}}^{\boldsymbol{S}}$ **and** $\boldsymbol{M_{i_1,i_2,\ldots,i_N}}^{\boldsymbol{O}}$ are corresponding cell values in the synthesized and observed populations respectively
- $\boldsymbol{\varphi = n_1 * n_2 * n_3 \ldots * n_N}$ corresponds to the number of corresponding cells of the cross-classification table

Table 7: An Iterative Proportional Updating (IPU) table: a two-dimensional example

	Attribute -1					
Attribute – 2	Class-1	…	Class-n	Observed totals per class	Row Sums	Scale factors
Class-1	$\boldsymbol{M}_{(\mathbf{1},\mathbf{1})}$	…	$\boldsymbol{M}_{(\mathbf{1},\boldsymbol{n})}$	Total_1	Sum_1	$\mathrm{f}_1 = \frac{\mathrm{Total}_1}{\mathrm{Sum}_1}$
…	…	…	…	…	…	…
Class-n	$\boldsymbol{M}_{(\boldsymbol{n},\mathbf{1})}$	…	$\boldsymbol{M}_{(\boldsymbol{n},\boldsymbol{n})}$	$\mathrm{Total}_\mathrm{n}$	Sum_n	$\mathrm{f}_\mathrm{n} = \frac{\mathrm{Total}_\mathrm{n}}{\mathrm{Sum}_\mathrm{n}}$
Observed totals per class	Total_1	…	$\mathrm{Total}_\mathrm{n}$	**Grand Total**		
Column Sums	Sum_1	…	Sum_n			
Scale factors	$\mathrm{f}_1 = \frac{\mathrm{Total}_1}{\mathrm{Sum}_1}$	…	$\mathrm{f}_\mathrm{n} = \frac{\mathrm{Total}_\mathrm{n}}{\mathrm{Sum}_\mathrm{n}}$			

Post-IPU Operations

It is relevant to discuss the scenario of data unavailability here again. One aspect of this problem, as described in section 2.2.1 (paragraph on combinatorial optimization), is that the attribute classes (bins) of the available data and the attribute classes for the intended purpose (freight demand modeling) may not match. If the available attribute classes are smaller (finer) than the intended classes, values can be summed to get desired figures for the intended classes. However, if attribute classes are larger (coarser) than the intended classes, methods of unbundling the larger classes into smaller ones are needed.

The following equation describes such an unbundling process. Let the cell entry $\boldsymbol{M_{i_1,i_2,\ldots,i_N}}^{\boldsymbol{S}}$ of an IPU table be the result of the iterative process. Let also that attribute class i_I is to be split into several sub classes q. Then, the number of individuals belonging to each sub-class is given by:

$$\boldsymbol{M_{(q,\ i_1,i_2,\ldots,i_N)}}^{\boldsymbol{S}} = \boldsymbol{M_{(i_1,i_2,\ldots,i_N)}}^{\boldsymbol{S}} * w_q$$

Where:

- $\boldsymbol{M}_{(q,\ i_1,i_2,\dots,i_N)}{}^{S}$ is the synthesized number of firms in sub-class q of class i_I
- w_q is the proportional weight of firms in sub-class q within the bundled class i_I
- $\boldsymbol{M}_{(i_1,i_2,\dots,i_N)}{}^{S}$ is the synthesized number of firms falling under attribute class i_I

Here, we assume that data on w_q, which is the key unbundling statistic, is available.

3.3. Summary

In summary, a generic N-dimensional model has been suggested. Specific to microsimulation of firm attributes as identified in the literature review, the intended model takes the form of a seven-dimensional joint distribution table from which anonymous firm agents of specific attribute sets can be generated. The model is based on the IPU algorithm whose data requirements have been clearly stated. ***Noteworthy is that IPU can disaggregate data only up to the disaggregation level of the prior sample data used and no further.*** Hence, to generate firm data at a certain attribute classification level, a microsample of the same classification level is required.

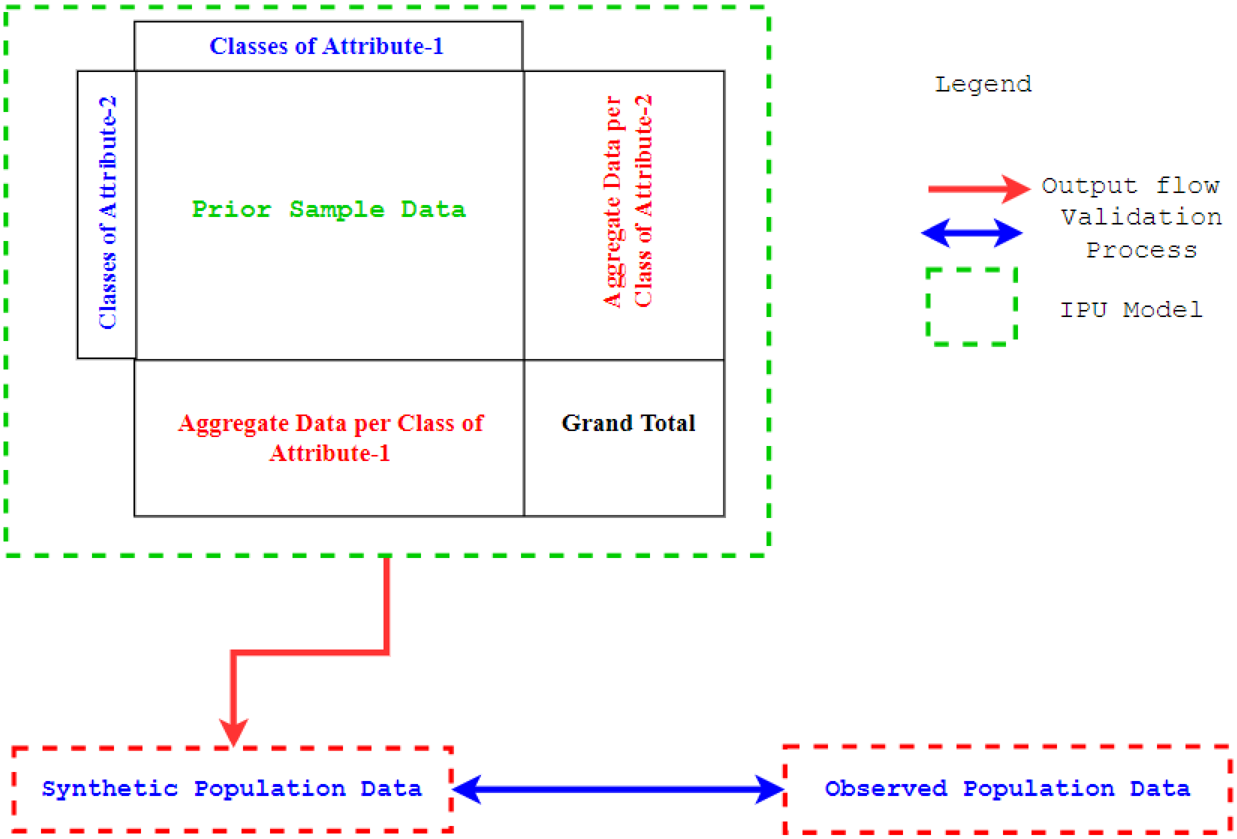


Figure 4: Model Structure (two-dimensional simplification)

4. Case Study: Synthesis of Firm Population in South Holland

The objective of this chapter is to give context to the research by applying the model described in the previous chapter to a specific case study, namely: the firm population of the Province of South Holland in The Netherlands. This chapter sets the geographical limitations to the study area and the levels of data aggregation within the study area. Furthermore, this chapter compares the available data to the ideal data requirements set out in the previous chapter. It then details the impact of the limitations of the available data on the model and the specific workarounds that were used to render as much of the type of outputs as intended in the initial ambitions of the model.

4.1. Study Area

The geographical limit to the case study is set to the province of South Holland. The data used (see below), is collected along spatial hierarchies corresponding to administrative hierarchies. These are: national level, the level of municipalities (Dutch: gemeenten) and neighborhoods (Dutch: buurten). Furthermore, the data used looked at these administrative organizations as they were in 2015; at the time, the study area consisted of 34 municipalities and 1283 neighborhoods.

4.2. Available Data

This sub-chapter explains the sourcing of the data and a comparison to the ideal requirements for a firm population synthesis model. The source of the data used in the research is open data from the Dutch Central Bureau for Statistics (CBS). These are data that CBS offers for public viewing free of charge and they constitute the entirety of the data used in this research.

4.2.1. National Firm Population Statistics

This data set, shown in table 2, gives the distribution of employee size classes across different economic sectors at in The Netherlands. The data sheet also contains the detailed descriptions of the economic activities included in each of the sectors. It is noteworthy that these descriptions have been found adequately similar to the SBI 2008 classification mentioned in the literature review and will therefore be taken as corresponding equivalents thereof.

Table 8: National level data

Economic Sectors	Employment Sizes								Totals
	1	2	3 - 5	5 - 10	10 - 20	20 - 50	50 - 100	> 100	
Agriculture	37990	19675	9300	3125	1205	525	90	50	71960
Mining	260	20	20	25	20	20	15	20	400
Manufactiring	36800	5565	4120	4635	3305	2595	1160	1195	59375
Energy	665	110	80	45	25	25	5	30	985
Water Supply	880	120	90	125	100	90	40	60	1505
Construction	123605	11085	5795	4535	2600	1445	455	305	149825
Wholesale and Trade	140750	30585	22325	16025	7030	4390	1255	1005	223365
Transportation and Storage	22480	4720	2905	2350	1465	1140	465	410	35935
Food and Accomodation	21995	10020	7900	6845	3145	1050	225	150	51330
IT	67720	5320	2745	2215	1410	945	340	255	80950
Financial Institutions	73240	5085	2050	1345	595	290	95	130	82830
Real Estate	17675	3330	1790	810	335	210	100	80	24330
Professional Businesses	262285	18260	8500	6565	3275	1775	500	405	301565
Renting & Leasing	45080	5845	3600	2935	1745	1565	735	805	62310
Public admin. & Government	125	5	10	15	20	45	110	415	745
Education	58020	3505	1095	700	530	450	275	765	65340
Health and Social work	102240	8220	6965	5460	2190	1040	410	915	127440
Culture, sports and recreation	81295	4860	2330	1540	750	470	135	95	91475
Misc. Services	78100	6060	4010	2390	820	370	135	90	91975

Figure 7 shows graphically the sectoral shares of each economic sector in the national firm population given in the dataset above.

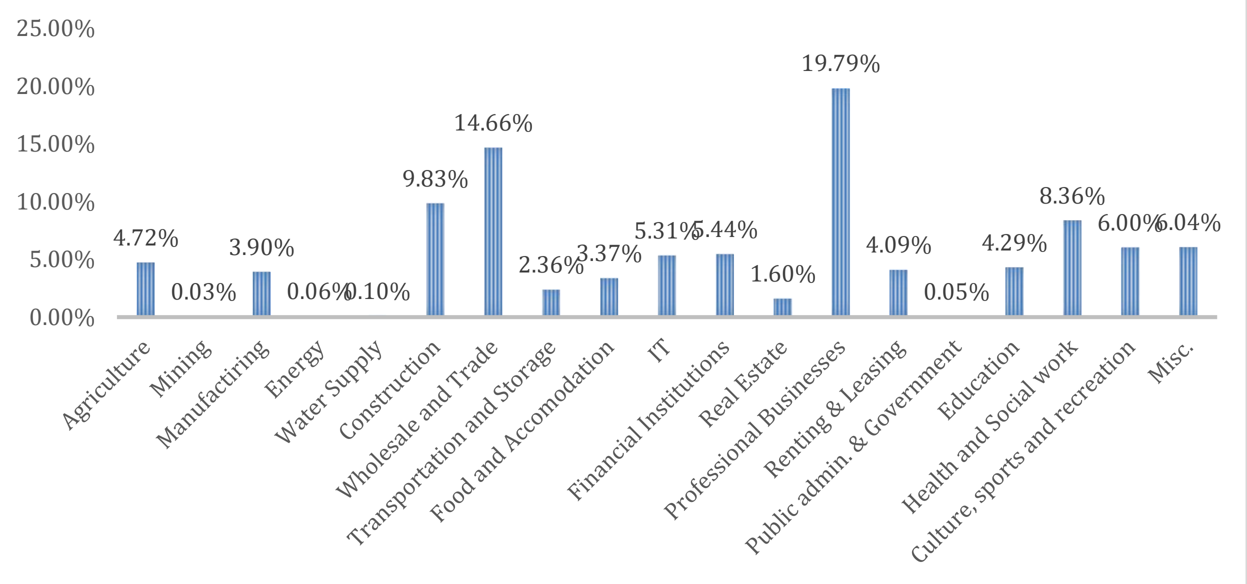
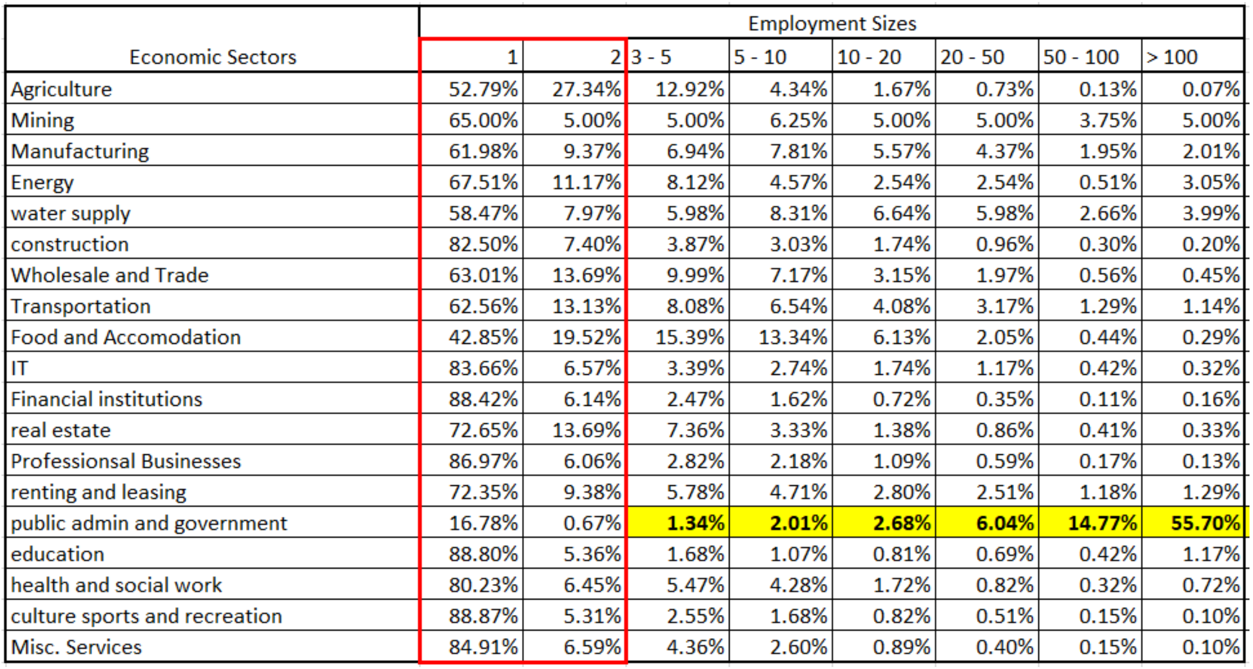


Figure 5: Sectoral shares of economic sectors per SBI 2008 in The Netherlands (Source: National Statistics, CBS (2015))

From the same data set, the intra-sector employment size distributions are shown in the table below. An important insight is that one and two-employee firms (red box) dominate the distributions across sectors (except public and government employment: yellow highlight). For freight demand predictions based on employee size, these firms have very high representation for a comparatively low freight demand.

Table 9: Intra-sector employment distributions (Source: National Statistics, CBS (2015))



Economic Sectors	Employment Sizes							
	1	2	3 - 5	5 - 10	10 - 20	20 - 50	50 - 100	> 100
Agriculture	52.79%	27.34%	12.92%	4.34%	1.67%	0.73%	0.13%	0.07%
Mining	65.00%	5.00%	5.00%	6.25%	5.00%	5.00%	3.75%	5.00%
Manufacturing	61.98%	9.37%	6.94%	7.81%	5.57%	4.37%	1.95%	2.01%
Energy	67.51%	11.17%	8.12%	4.57%	2.54%	2.54%	0.51%	3.05%
water supply	58.47%	7.97%	5.98%	8.31%	6.64%	5.98%	2.66%	3.99%
construction	82.50%	7.40%	3.87%	3.03%	1.74%	0.96%	0.30%	0.20%
Wholesale and Trade	63.01%	13.69%	9.99%	7.17%	3.15%	1.97%	0.56%	0.45%
Transportation	62.56%	13.13%	8.08%	6.54%	4.08%	3.17%	1.29%	1.14%
Food and Accomodation	42.85%	19.52%	15.39%	13.34%	6.13%	2.05%	0.44%	0.29%
IT	83.66%	6.57%	3.39%	2.74%	1.74%	1.17%	0.42%	0.32%
Financial institutions	88.42%	6.14%	2.47%	1.62%	0.72%	0.35%	0.11%	0.16%
real estate	72.65%	13.69%	7.36%	3.33%	1.38%	0.86%	0.41%	0.33%
Professionsal Businesses	86.97%	6.06%	2.82%	2.18%	1.09%	0.59%	0.17%	0.13%
renting and leasing	72.35%	9.38%	5.78%	4.71%	2.80%	2.51%	1.18%	1.29%
public admin and government	16.78%	0.67%	**1.34%**	**2.01%**	**2.68%**	**6.04%**	**14.77%**	**55.70%**
education	88.80%	5.36%	1.68%	1.07%	0.81%	0.69%	0.42%	1.17%
health and social work	80.23%	6.45%	5.47%	4.28%	1.72%	0.82%	0.32%	0.72%
culture sports and recreation	88.87%	5.31%	2.55%	1.68%	0.82%	0.51%	0.15%	0.10%
Misc. Services	84.91%	6.59%	4.36%	2.60%	0.89%	0.40%	0.15%	0.10%

4.2.2. Municipality Data

This is a database file that contains data at the municipality level. It is a vast database of many socio-demographic characteristics. While the all the attribute headers of the data set are given in CBS, (2012), the data headers selected for this research are:

- Municipality name
- Municipality Code: unique identifiers of each municipality
- Municipality area: the area occupied by each municipality.
- Population Size: the number of inhabitants in each municipality.
- Number of firms: these are availed per sector category within each municipality. The numbers are not found for each sector (A - U). Rather, the sectors are found bundled as:
    - Bundle - 1: Agriculture (A),
    - Bundle - 2: Mining and quarrying (B), Manufacturing (C), Electricity, gas, steam and air conditioning supply (D), Water supply; sewerage, waste management and remediation activities (E), Construction (F),
    - Bundle - 3: Wholesale and retail trade; repair of motor vehicles and motorcycles (G), Accommodation and food service activities (I),
    - Bundle - 4: Transportation and storage (H), Information and communication (J),
    - Bundle - 5: Financial institutions (K), Renting, buying and selling of real estate (L),
    - Bundle - 6: Consultancy, research and other specialized business services (M), Renting and leasing of tangible goods and other business support services (N),

- ○ Bundle - 7: Culture, sports and recreation (R), Other service activities (S), Activities of households as employers; undifferentiated goods and service-producing activities of households for own use (T), Extraterritorial organizations and bodies (U)
- ○ Bundle - 8: Public administration, public services and compulsory social security (O), Education (P), Human health and social work activities (Q) as well as all other activities that do not fit within the previous seven bundles.

- Coordinates: these are the centroidal coordinates of the municipalities.

4.2.3. Neighborhood Data

This is a database file containing data at the neighborhood level. It contains all the neighborhoods in South Holland with their unique neighborhood and municipality codes. The full description of the headers is given in CBS, (2012). However, only the headers used for this research are given below.

- Neighborhood name
- Neighborhood Code: unique identifiers of each neighborhood
- Municipality name
- Municipality Code: unique identifiers of each municipality
- Neighborhood area: the average area occupied by each neighborhood.
- Population Size: the number of inhabitants in each neighborhood.
- Number of firms: these are availed per sector categories within the neighborhoods. The sector categories are the same bundles that are found in the municipality data.
- Coordinates: these are the centroidal coordinates of the neighborhoods.

4.3. Usage of the Data

In this sub-chapter we look at how the available data compare to the requirements of the specified population synthesis model. Thereafter, the effects of the available data on the model structure and/or implementation are discussed. The following table shows an audit of the available data per attribute. The headers of the table comprise of the requirements of the IPU model specified in the previous chapter.

Table 10: Inventory of the Dataset Used

Attribute	Aggregate Data (Coarse)			Firm Level Data
	National Population Statistics	Municipality	Neighborhood	
Employee Size	✓	✘	✘	✘
Economic Sector	✓	✓	✓	
Floor Area	✘	✘	✘	
Fleet Size	✘	✘	✘	
Annual turnover	✘	✘	✘	
Commodity Type	✘	✘	✘	
Location	*✓	✓	✓	
****All firms considered are within The Netherlands***				

Looking at the available data, we can make the following observations.

- First, the firm-level data that serve to populate the IPU table by classifying across the possible attribute classes of the six attributes are not available. Therefore, possible workarounds need to be made to initialize the IPU model with values that reflect the joint distribution of firms across attribute classes.
- Second, aggregate data are available for the attribute economic sector at spatial aggregation levels of municipality and neighborhood. These can serve as marginals for disaggregation levels below their respective levels.

Therefore, available data falls short of the requirements of the specified model. This warrants changes in the model structure that reflect the available data. These changes are discussed below while we leave the general impacts of these changes on the research for the conclusions and recommendations chapter.

Given the description of the available data, the IPU level can be applied in the following way to prove its potency in carrying out disaggregation and population synthesis.

- ✓ First, we reformulate the model structure. Figure-6 below shows the IPU model that follows from this new architecture.
- ✓ Recognizing that firm-level data are not available, the model now aims for the finest level of spatial disaggregation attainable with the data, i.e. neighborhood-level. Hence in the new structure, the model can aim for disaggregating the municipal data to neighborhood data. The municipal data are used as bundle-marginals for the IPU table and the neighborhood totals are used for neighborhood marginals (see figure-6). Meanwhile, the observed neighborhood data can be used as validation data.

- ✓ Second, we specify which of the attributes have adequate data to fit into such a model architecture. Here, we observe that only neighborhoods(location) and economic sector bundles have the aggregates to fit into the model. ***Hence, at this point the model downsizes to a bi-variate distribution of economic sector bundles against location (more precisely neighborhoods) leading to the specification:***

$$firm = firm(economic\ sector\ bundles, location)$$

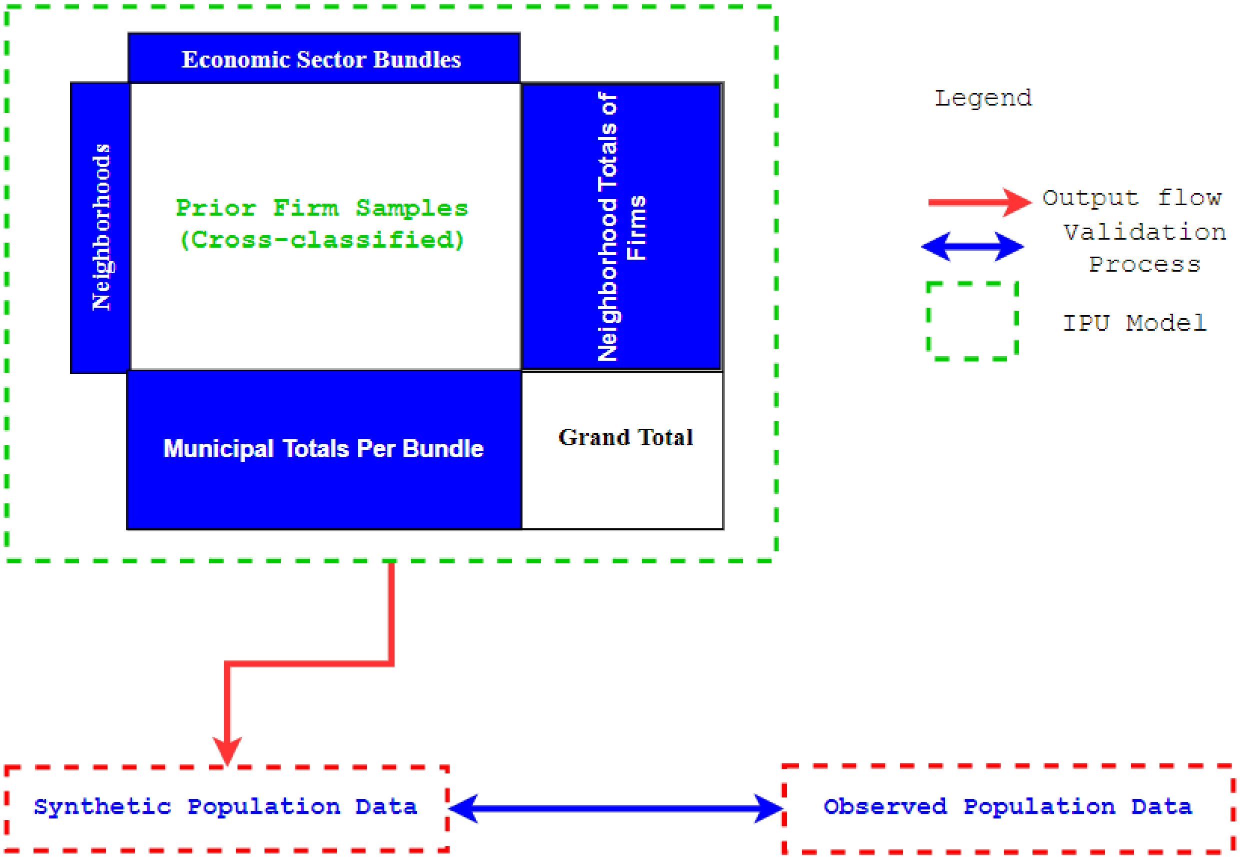


Figure 6: New model architecture (blue fields)

4.3.1. Initializing the IPU Table

As was noted in Chapter-2, IPU sets high requirements on the kind of detailed input it requires. For the data that initialize the inner matrix of the IPU, we need a representative sample of firms with the attribute of economic sector [bundle] at the neighborhood level. However, table-10 of the data inventory shows there is a significant lack of detailed sample data to initialize the algorithm.

Therefore, we turn to table-6 for making a shortlist of candidate parameters that can make for initialization data. The following are formulations of the indicators mentioned in table-6 to fill the data gaps of the specified model.

Urban Density

It has been stated in the literature review that labor force is one of the resources shared by all firms regardless their economic sectors. This makes urban population densities important indicators of firm location distributions. Hence urban density can be one candidate to initialize the IPU matrix with. Important to note here is that not all sectors are positively correlated to urban density. Hence, bundle-1(agriculture) is initialized according to the reciprocal of urban density. ***The implicit assumption here is that firms are***

distributed among neighborhoods of a municipality according to urban density. This is the assumption that a validation against observed neighborhood data can either confirm or refute.

From the description of the data, we recall that economic sectors are grouped into 8 bundles. This bundling of economic sectors puts a restriction in the synthesis architecture of the model because the synthesis would have to furnish the exact same type of population as the validation data for a reasonable comparison. Hence when implementing the IPU model, synthesis takes place not for individual economic sectors but the aggregate number within the bundles. For such data arrangement, the IPU cross table will look like:

Table 11: An Initialized IPU table of Distribution of Economic Sectors (Bundled Structure)

Zone/ Economic sector	Neighborhood-1	Neighborhood-2	…	Neighborhood-n	Observed # firms per Bundle	Row Sums	Scale factors
Bundle-1	$\frac{1}{D_{(1,1)}}$	$\frac{1}{D_{(1,2)}}$	…	$\frac{1}{D_{(1,n)}}$	Total_1	Sum_1	$f_1 = \frac{\mathrm{Total}_1}{\mathrm{Sum}_1}$
Bundle-2	$D_{(2,1)}$	$D_{(2,2)}$	…	$D_{(2,n)}$	Total_2	Sum_2	$f_2 = \frac{\mathrm{Total}_2}{\mathrm{Sum}_2}$
Bundle-3	$D_{(3,1)}$	$D_{(3,2)}$	…	$D_{(3,n)}$	Total_3	Sum_3	$f_3 = \frac{\mathrm{Total}_3}{\mathrm{Sum}_3}$
…	…	…	…	…	…	…	…
Bundle - 8	$D_{(8,1)}$	$D_{(8,2)}$	…	$D_{(8,n)}$	Total_8	Sum_8	$f_8 = \frac{\mathrm{Total}_8}{\mathrm{Sum}_8}$
Column Sums	Sum_1	Sum_2	…	Sum_n	**Grand Total**		
Observed # firms per Zone	Total_1	Total_2	…	Total_n			
Scale factors	$f_1 = \frac{\mathrm{Total}_1}{\mathrm{Sum}_1}$	$f_2 = \frac{\mathrm{Total}_2}{\mathrm{Sum}_2}$	…	$f_n = \frac{\mathrm{Total}_n}{\mathrm{Sum}_n}$			

where:

- ➔ D_{ij} are the urban densities in each zone (neighborhood) j and hence are equal for a fixed j
- ➔ Factors F_i and F_j are row and column scale factors respectively computed as ratios of the actual (observed) sums to the synthesized sums.

The final outputs of the synthesis process are the entries M_{ij}^s, where s stands for synthesized.

Accessibility

Thus, we can take the distances between centroidal neighborhood coordinates and the corresponding nearest highway access points, as elements to initialize the IPU table with. These can be used in conjunction with urban density as multiplied factors. Therefore, each IPU entry will be initialized as:

$$M_{ij} = D_{ij} d_{ij}$$

where:

- M_{ij} is the cell entry of row i and column j
- D_{ij} is the urban density of neighborhood j as specified in table-5.
- d_{ij} is distance from centroid of the neighborhood to the nearest highway access point.

Sector - Specific Factors

Considering the requirements inherent in economic activities can lead to further possibilities in initializing the IPU table. For instance, we can consider agriculture which is a space-intensive economic sector. This makes it ill-suited for dense urban environments. Therefore, for the first row of table-5, one can consider using the reciprocal of urban density values per neighborhood. This can also be used for mining and quarrying activities. In the current format of the table-5, however, this is not possible as mining is found bundled with other sectors in bundle-2.

At times, there can be unique pull factors that strengthen preference for some locations. These could take the forms of: ports, airports, industrial parks, business incubation centers, etc. If a neighborhood contains a pool factor attractive to certain economic sectors, an indicator leading to decreasing preference as a function of distance from the pull factor can be formulated as:

$$M_{ij} = \beta_{ij} \frac{D_{ij}}{d_{ij}}$$

where:

- M_{ij} is the cell entry of row *i* and column *j*
- β_{ij} is a probability amplification term for a pool factor to attractive bundle *i* in neighborhood *j; it takes a value of one for irrelevant bundles*
- D_{ij} is the urban density of neighborhood *j* as specified in table-5.
- d_{ij} is distance from the centroid of neighborhood j to the centroid of the location of the attractive factor

Land Rent

Land rent can be an important factor in a firm's ability to locate (or stay located) at a given available location. This is because firms co-locating seeking the benefits of agglomeration leads to competition for available space which will be dictated by the price of land in the market. In this competitive land market, rent is a disincentive to locate. (Van den Heuvel et. al., 2013) argue in their research that only large firms have the financial robustness to agglomerate in areas of high land rate.

As a simple method of capturing this effect, one can initialize the IPU table with the rent as a disincentive and the financial capacity of the firm as an incentive. Here, average turnover of a firm in a given sector and the land rent can be combined as follows to give:

$$M_{ij} = {}^{T_{ij}}\!/_{R_{ij}}$$

where:

- M_{ij} is the cell entry of row i and column j
- T_{ij} is the average annual turnover of firms in a given sector I (constant in a row)
- R_{ij} is the average land rent in neighborhood j (constant in a column)

One can make the following observations regarding the above initialization. On one hand, it is interesting that the elements of the factor drive the distributions in orthogonal directions as average turnover is constant over a sector and rent is constant per location. On the other hand, both the numerator and denominator are both average values for variables whose values can show immense heterogeneity. This heterogeneity is amplified in the bundled-sectors' arrangement of table-5.

Note on selected initialization:

From the suggested workarounds to bridge the data gap for initializing the IPU, we can see that data on accessibility indicators, annual turnovers and land rent prices are not readily available within the data set used. Meanwhile, urban density is a readily available and practical option in the dataset. Therefore it is used as per the formulation in table-11.

4.3.2. Post-IPU Operations

Once again, we notice that the IPU model structure does not allow for finer disaggregation than what is available in the observed population. Even so, there are some improvements we can make to the synthesis process with the help of some of the data not utilized as yet.

The first improvement is the unbundling of economic sectors. This can be done by using the national level data of the number of firms in each individual sector. These numbers can be used to make proportional weights for each economic sector within a specific bundle. Hence, once the bundles are synthesized, the number of firms in each economic sector across each zone can be computed as:

$$M_{q,ij}^{s} = M_{ij}^{s} w_q$$

where:

- ➔ $M^s_{q,ij}$is the number of firms in economic sector q, with $q \in \{A, B, C, ..., U\}$
- ➔ $w_q$is the proportional weight of firms of economic sector q within bundle i and zone j according to national statistics
- ➔ $M^s_{ij}$is the synthesized number of firms within bundle i and zone j

Evidently, this improvement assumes national and regional distributions of sectors in the firm population are the same.

The second improvement concerns the employment size distributions. Neither the municipality nor the neighborhood-level data contain employee size distributions. These distributions can be synthesized as consequences of economic sectors instead. This is done by means of the national-level data for employee size distributions according to economic sectors. If one assumes national and regional distributions are the same:

$$M^s_{p,q,ij} = M^s_{q,ij} w_p$$

where:

- ➔ $M^s_{p,q,ij}$ is the number of firms of employee-size class p in economic sector q [hence bundle i] and zone j.
- ➔ w_p is the proportion of the number of firms of employee-size class p within sector q as found in the national-level data.
- ➔ $M^s_{q,ij}$is the number of firms in economic sector q as obtained in the prior post-IPU operation.

By such means the available data will be used to furnish the desired distributions of economic sector and employment size. Therefore, at the end of these operations firm agent will have been specified as:

$$firm = firm(employee\ size, economic\ sector, location)$$

4.4. Summary

At this stage, the IPU model has evolved from its initial specification of a six-variable distribution to a bivariate distribution between economic sector and location due to constraints in available data. The level of detail required to initialize the IPU matrix had created a data gap. This has been filled by opting for urban density for a substitute distribution. The model has also downsized in its pool of population attributes to location and economic sector bundles because both aggregate and disaggregate data on floor area, fleet size and annual turnover of firms are not available. To widen the attribute pool of the model, some post IPU operations were employed to synthesize employee size at firm level using national statistics and simplifying assumptions. The model output will be generated in terms of spatial distributions of economic sectors and employment sizes. Thus far, the following schematic describes the input-output relations in the model.

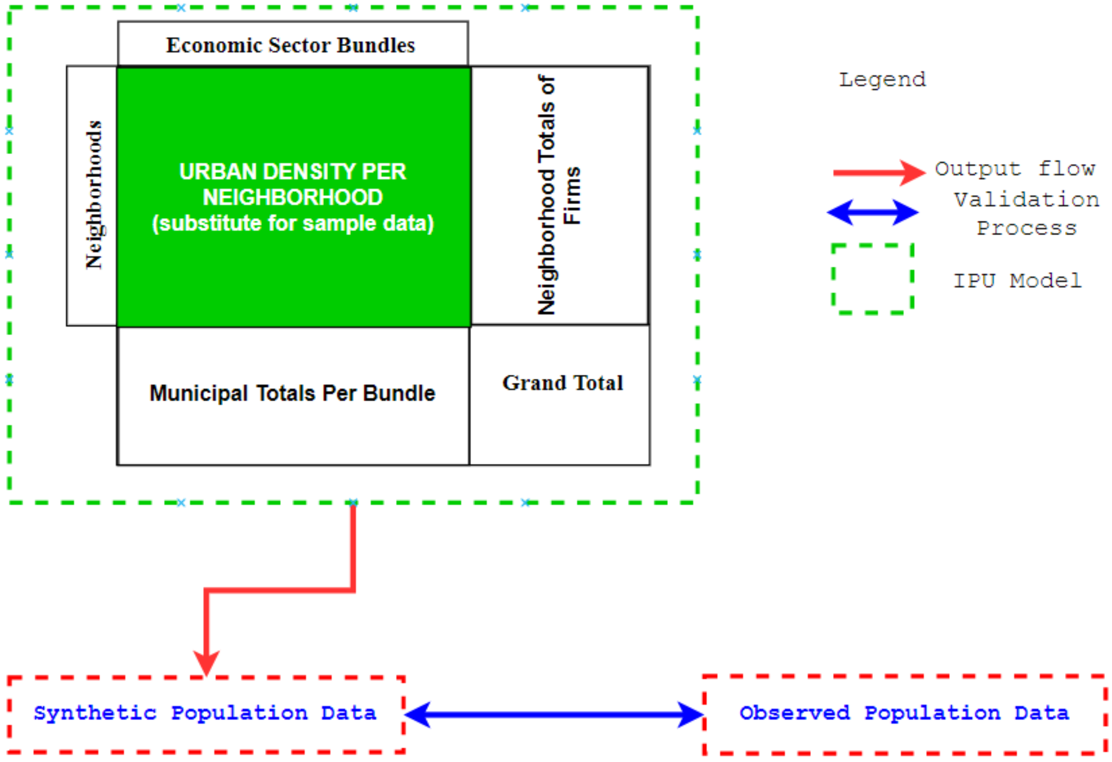

Economic Sector Bundles
Neighborhoods
URBAN DENSITY PER NEIGHBORHOOD (substitute for sample data)
Neighborhood Totals of Firms
Municipal Totals Per Bundle
Grand Total
Legend
Output flow
Validation Process
IPU Model
Synthetic Population Data
Observed Population Data

Figure 7: Model changes due to data constraints (green highlight)

5. Validation

This chapter discusses results gained after implementing the IPU model. It discusses the validation of the model against observed neighborhood firm population data.

As was stated in the case study chapter, free public data from CBS is used to validate the synthesized population. This validation data, however, exists for the distribution of bundles of economic sectors only and not the employment size classes. Hence, only the distribution of the economic sector bundles will be validated. The R-squared parameter (see subsection: IPU algorithm) has been used as a comparative measure for the goodness of fit between synthesized and observed data. This parameter has been computed for each firm agent, for each attribute and at various levels of aggregation. For each kind of computation, the associated modifications on the basic formula for R-squared are given.

Both the synthesized and observed firms have attributes of sector bundle and neighborhood. These data were first arranged in the following indexing format:

${M_{i_1 i_2 i_3}}^O$now indicates the observed number of firms in bundle - i_1 and neighborhood - i_2 of municipality - i_3

${M_{i_1 i_2 i_3}}^S$ now indicates the synthesized number of firms in bundle -i_1 and neighborhood - i_2 of municipality - i_3

Then, the correlation parameter was computed for the firm population of South Holland as:

$$\boldsymbol{R^2 = \frac{[n\sum {M_{i_1 i_2 i_3}}^S {M_{i_1 i_2 i_3}}^O - \sum {M_{i_1 i_2 i_3}}^S \sum {M_{i_1 i_2 i_3}}^O]^2}{[n\sum {M_{i_1 i_2 i_3}}^{S2} - \left(\sum {M_{i_1 i_2 i_3}}^S\right)^2][n\sum {M_{i_1 i_2 i_3}}^{O2} - \left(\sum {M_{i_1 i_2 i_3}}^O\right)^2]}}$$

The overall correlation fit parameter value of **0.76** was obtained for a two-dimensional table of neighborhoods vs sector bundles (size: 1283 rows × 8 columns). However, we need to know more how this overall value is distributed across the rows and columns of the table. Next, we compute correlation parameters for distributions across bundles within each one of the 1283 neighborhoods(rows). The above formula will be used with constant values of i_2 and i_3 per computation. The results are shown in the form of frequency distribution of correlation parameters below.

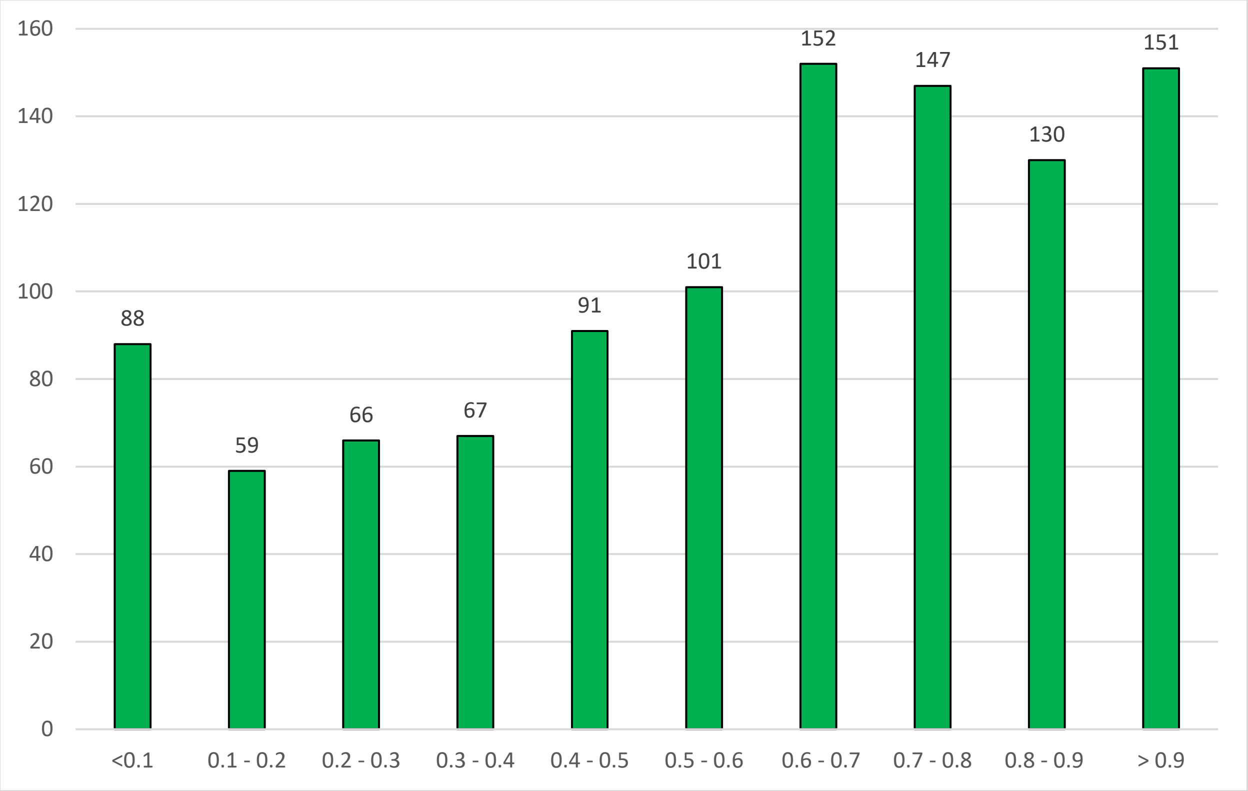


Figure 8: Frequency of R-squared parameters for 1283 neighborhoods of South Holland

The first thing we notice is that the frequency plot above is for 1283 neighborhoods from which the R-squared value is defined; these are 1054 neighborhoods. The remaining 229 neighborhoods have null cell entries across all sector bundles of the IPU table. From the above figure, we can see that the R-squared values are found distributed across the bins from 0 to 1. This variation is better explained by looking at the distribution of correlation values across different bundles of economic sectors. We can use the parameter formula given, keeping the values of the indices i_1 and i_3 constant per computation. Then, we obtain the results in table below.

Table 12: Goodness of fit across bundles

Bundle	**A**	**B -F**	**G & I**	**H & J**	**K & L**	**M & N**	**R & U**
Correlation Parameter (R^2)	0.00	0.60	0.81	0.71	0.70	0.90	0.87

Quick Key:
A = Agriculture, B=Mining, C=Manufacturing, D= Energy, E=Water supply, F = Construction, G = Wholesale and Trade, H= Transportation, I = Food and Accommodation Services, J= IT, K= Finance, L = Real estate, M=Professional business, N = Rental and leasing, R= Recreation and culture,
U = Extraterritorial organizations

We observe that the model underperforms for the first bundle (agricultural firms) while in all other bundles there is more than 60% correlation between synthesized and observed data. Also, all but bundle-2, show more than 70% correlation with observed data. Bundles-6 (professional services and rental services), shows the best fit to the observed population values.

Thus far, we have produced one correlation parameter per row and one per column for the cross table of the synthetic population. Let us now observe a different level of aggregation and compute correlation parameters over municipalities (chunks of rows). To do so, we use the same formula by keeping the value of index i_3 constant per computation. The table below shows the results for the values of the municipal fit parameters. In addition, the percentage shares of each bundle in the population is also given to help infer how dominant bundles have influenced the overall goodness of fit in the municipality.

Table 13: R-squared values Per municipality

No.	Municipality	R-Squared	Percentage share of municipal firm population (obseved data)						
			A	B -F	G & I	H & J	K & L	M & N	R & U
1	Alblasserdam	0.74	0.66%	19.08%	25.33%	10.86%	12.50%	23.36%	8.22%
2	Albrandswaard	0.81	1.62%	12.43%	17.30%	13.51%	14.59%	29.46%	11.08%
3	Bodegraven-Reeuwijk	0.82	6.74%	19.23%	19.80%	6.74%	14.20%	24.53%	8.75%
4	Capelle aan den IJssel	0.36	0.41%	12.02%	21.04%	12.33%	12.85%	29.53%	11.81%
5	Delft	1.00	0.90%	12.60%	19.35%	12.25%	7.72%	34.31%	12.87%
6	Dordrecht	0.68	1.01%	15.96%	24.09%	9.61%	10.92%	26.41%	11.99%
7	Gorinchem	0.84	0.67%	14.98%	24.75%	8.92%	9.93%	28.28%	12.46%
8	Gouda	0.61	0.40%	16.18%	21.54%	9.00%	9.61%	29.42%	13.85%
9	Hardinxveld-Giessendam	0.94	2.28%	28.66%	17.59%	8.79%	15.64%	19.22%	7.82%
10	Hellevoetsluis	0.13	2.29%	16.67%	23.75%	8.13%	11.46%	22.50%	15.21%
11	Hendrik-Ido-Ambacht	0.77	0.96%	15.55%	20.81%	12.68%	15.55%	24.88%	9.57%
12	Hillegom	0.93	4.90%	16.71%	25.36%	5.76%	9.80%	26.22%	11.24%
13	Katwijk	0.53	3.47%	17.57%	30.45%	6.19%	11.51%	20.05%	10.77%
14	Krimpen aan den IJssel	0.86	0.25%	18.25%	21.00%	12.50%	13.50%	24.25%	10.25%
15	Krimpenerwaard	0.83	7.06%	21.07%	19.78%	7.95%	13.02%	20.68%	10.44%
16	Leiden	0.70	0.17%	14.38%	21.05%	9.98%	7.36%	32.23%	14.83%
17	Leiderdorp	0.51	1.10%	15.11%	19.78%	9.34%	10.99%	30.77%	12.91%
18	Lisse	0.33	4.62%	13.33%	25.64%	6.41%	11.79%	26.67%	11.54%
19	Maassluis	0.73	2.60%	16.67%	23.44%	9.90%	11.72%	22.92%	12.76%
20	Nieuwkoop	0.59	13.42%	22.12%	17.58%	6.62%	9.83%	20.98%	9.45%
21	Noordwijk	0.87	3.35%	11.81%	24.80%	6.89%	13.58%	28.15%	11.42%
22	Noordwijkerhout	0.95	12.96%	15.28%	27.24%	5.32%	7.31%	22.59%	9.30%
23	Oegstgeest	0.95	0.60%	7.78%	16.17%	7.49%	13.47%	43.11%	11.38%
24	Papendrecht	0.91	0.46%	16.09%	22.76%	9.66%	13.79%	24.83%	12.41%
25	Ridderkerk	0.17	2.15%	17.31%	26.32%	10.59%	13.73%	19.74%	10.16%
26	Rijswijk	0.82	0.77%	14.67%	20.72%	10.55%	10.04%	31.27%	11.97%
27	Rotterdam	0.69	0.59%	12.88%	22.18%	11.22%	8.99%	29.33%	14.81%
28	Schiedam	0.44	1.23%	21.90%	24.08%	8.25%	8.82%	22.94%	12.80%
29	's-Gravenhage	0.67	3.34%	16.26%	20.04%	8.15%	7.96%	29.96%	14.29%
30	Sliedrecht	0.69	0.56%	20.62%	24.86%	9.60%	14.12%	21.19%	9.04%
31	Westvoorne	0.84	8.68%	11.70%	21.51%	5.66%	17.74%	23.40%	11.32%
32	Zoetermeer	0.82	0.83%	12.85%	23.19%	10.73%	8.03%	31.34%	13.04%
33	Zoeterwoude	0.56	7.50%	19.38%	21.25%	8.75%	11.25%	21.88%	10.00%
34	Zwijndrecht	0.53	0.95%	17.85%	21.48%	16.43%	11.85%	21.48%	9.95%
Color Codes									
>20%			5 % - 20%				<5%		

From the above results, we can see that the goodness of fit and thereby the use of urban density as a proxy variable, is more satisfactory at the municipal level. However, there are still very low correlation values (highlighted in blue) that draw attention. These are the municipalities of Capelle aan den Ijssel, Hellevoetsluis and Lisse. These municipalities have low correlation values because a significant number of their neighborhoods have no inhabitants (hence zero population density); Capelle aan den Ijssel (30/79), Hellevoetsluis (4/36), Lisse (2/21) neighborhoods have zero densities respectively. Therefore, they present the zero-cell problem for an IPU model. As indicated in the literature review, there has not been a satisfactory remedy for this problem in IPU.

Then a deeper look was taken at municipal-level goodness of fit in each economic sector bundle to better interpret the results in the table above. To this end, each correlation parameter was computed as using the same parameter formula, but with indices i_1 and i_3 constant per computation. The results are given both in table and frequency graph formats below.

Table 14: Distribution of R-Squared Parameters per Municipality

Municipality	Bundles						
	Agriculture	Mining, Water Supply, Energy, Construction	Wholesale and Food Services	Logistics & IT firms	Real Estate and Finances	Professional Services & Rental	Recreational Activities
Alblasserdam	*	0.80	0.81	0.71	0.63	0.79	0.25
Albrandswaard	0.01	0.78	0.73	0.31	0.91	0.94	0.76
Bodegraven-Reeuwijk	0.04	0.86	0.89	0.84	0.88	0.93	0.89
Capelle aan den IJssel	0.00	0.17	0.19	0.11	0.19	0.32	0.67
Delft	0.99	1.00	1.00	1.00	1.00	1.00	1.00
Dordrecht	0.00	0.45	0.64	0.60	0.56	0.83	0.70
Gorinchem	0.00	0.72	0.89	0.84	0.78	0.94	0.92
Gouda	0.00	0.41	0.61	0.61	0.47	0.70	0.88
Hardinxveld-Giessendam	0.24	0.90	0.87	0.82	0.99	0.84	0.92
Hellevoetsluis	0.00	0.28	0.19	0.47	0.53	0.57	0.75
Hendrik-Ido-Ambacht	0.00	0.59	0.58	0.86	0.94	0.95	0.85
Hillegom	0.18	0.96	0.94	0.90	0.93	0.97	0.98
Katwijk	0.01	0.39	0.71	0.72	0.83	0.84	0.89
Krimpen aan den IJssel	*	0.72	0.90	0.85	0.87	0.65	0.42
Krimpenerwaard	0.05	0.90	0.84	0.69	0.91	0.93	0.90
Leiden	0.00	0.35	0.55	0.69	0.50	0.63	0.69
Leiderdorp	0.01	0.46	0.10	0.71	0.77	0.68	0.69
Lisse	0.04	0.40	0.77	0.25	0.47	0.65	0.81
Maassluis	0.04	0.71	0.66	0.67	0.79	0.87	0.84
Nieuwkoop	0.01	0.74	0.71	0.81	0.74	0.91	0.91
Noordwijk	0.31	0.82	0.81	0.86	0.79	0.91	0.93
Noordwijkerhout	0.43	1.00	0.99	0.98	0.98	0.97	0.94
Oegstgeest	0.30	0.45	0.86	0.65	0.93	0.98	0.88
Papendrecht	*	0.79	0.86	0.93	0.89	0.93	0.91
Ridderkerk	0.02	0.51	0.26	0.47	0.70	0.81	0.78
Rijswijk	0.02	0.82	0.65	0.78	0.82	0.89	0.76
Rotterdam	0.02	0.37	0.77	0.47	0.62	0.89	0.79
Schiedam	0.02	0.57	0.60	0.51	0.31	0.61	0.72
's-Gravenhage	0.02	0.33	0.73	0.67	0.41	0.76	0.80
Sliedrecht	*	0.80	0.57	0.55	0.59	0.72	0.23
Westvoorne	0.00	0.86	0.91	0.87	0.88	0.94	0.87
Zoetermeer	0.05	0.85	0.61	0.85	0.86	0.90	0.84
Zoeterwoude	0.00	0.93	0.98	0.99	0.99	0.90	0.70
Zwijndrecht	*	0.53	0.46	0.50	0.67	0.70	0.36

Quick Key: A = agriculture, B=mining, C=manufacturing, D= energy, E=water supply, F = construction, G = wholesale and trade, H= transportation, I = food and accommodation services, J= IT, K= finance, L = real estate, M=professional business, N = rental and leasing, R= recreation and culture, U = extraterritorial organizations

Color Codes	<0.3	0.3 - 0.4	0.4 - 0.5	0.5 - 0.6	0.6 - 0.7	0.7 - 0.8	0.8 - 0.9
	> 0.9						

* both observed and synthesized populations show null values; R-squared is undefined.

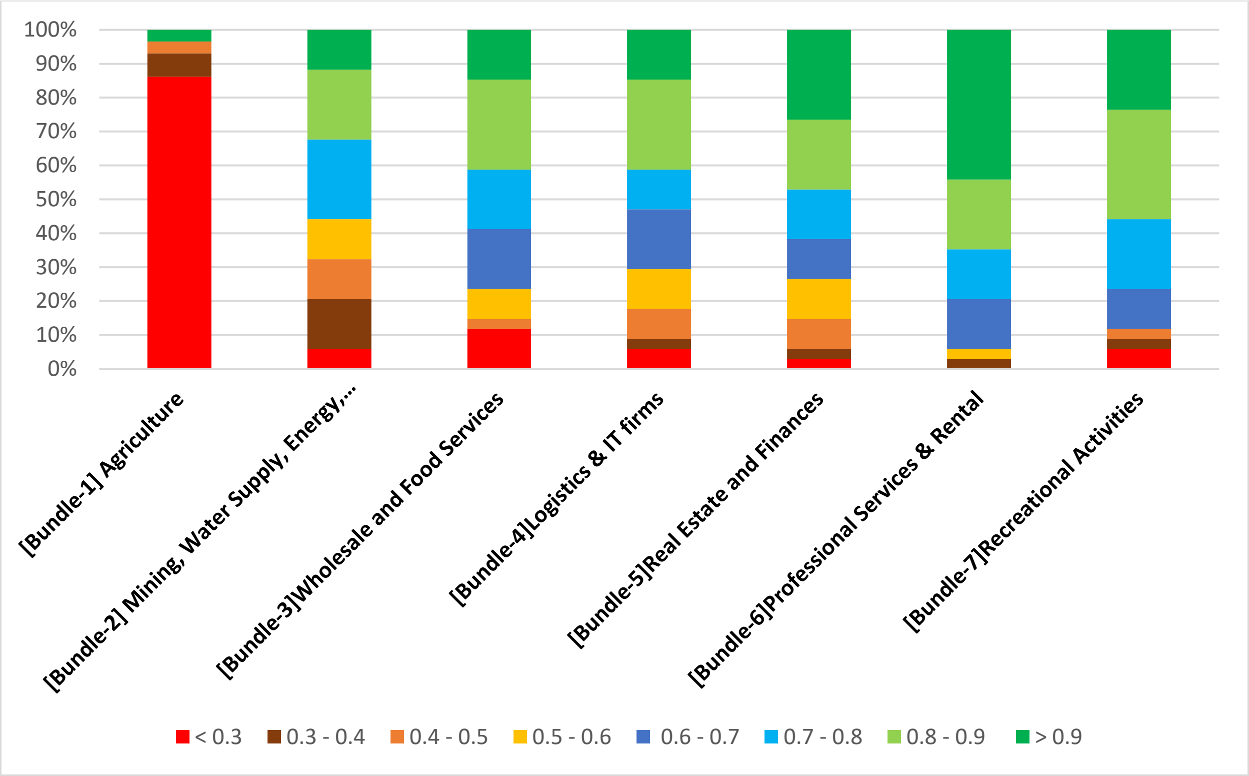


Figure 3: Summary of Table 14: Frequency of correlation parameters at municipal level

Here, we can see that the distribution of agricultural sector is not explained at the municipal level either, with 32 instances of $R^2 < 0.3$. This low performance of the model relates to the necessity of a better indicator for the agricultural sector. As mentioned in section 4.3.1., the IPU table was initialized to the reciprocal of the urban density for each neighborhood. Although this is sound in principle, we see that the reciprocals of densities are small numbers which give agriculture a far small share of the neighborhood marginals. Likewise, very sparsely populated neighborhoods have very high values of the reciprocals; this gives them exceedingly high shares of the bundle marginals. Therefore, in both dimensions of the table, the cells for agriculture are not set up to fit to observed data.

Bundle – 2 performs better than agriculture, however, its contribution o freight demand is high. Hence a better goodness of fit would have been satisfactory. The municipal-level also reveals significant variation (see figure-4) within the demand-oriented bundles (3,5,6,7,8) that performed well in overall bundle-wise fitting.

For all the bundles, urban density was the only indicator used in the initialization of the IPU table. This urban density-based firm location can bunch firms in city centers and residential areas for all sectors. This is not necessarily correct even for firm types that are not negatively correlated with urban density; e.g. manufacturing firms are rarely located in the city center. This can possibly explain the lower performance of the model at the finer (neighborhood) level but a more satisfactory performance at the coarser (municipality) level.

Another related discussion topic is the effect of bundled sector data on the results. The firms in a bundle are constrained to the same bundle totals and their distributions have the same fate regardless their sectors. But some bundles are more diverse than others (e.g. bundle-2 vs bundle-8). This makes it important to know composition of the bundles in order to get an idea of the distribution of errors in fitting amongst sectors. Since within-bundle proportions are not known from the observed data, the nuances of the results cannot be revealed.

Thus far, we have mainly discussed the interpretations from the figures seen in the results as regards spatial distribution and economic sectors. We now turn to an issue concerning the IPU method itself via an interesting case. In Chapter 2, it was stated that IPU needs modifications in dealing with rare instances, i.e., small totals which when proportioned between many cells of non-zero initializations, yield null distribution values when integerized. Here, we need initializations pointing exactly to which attribute class(es) the instances within the totals belong to and setting other cells null. This is different from selecting the right indicator or formulating a proper function for it.

So, what does the above mean for the results specifically? Example case: for the municipalities of *Capelle aan den Ijssel, Dordrect, Gouda* and *Gorninchem,* observed data show small numbers of agricultural firms confined to one or two neighborhoods. Meanwhile, proportional synthesis gives null[integer] values across all corresponding neighborhoods. Here, the ideal indicator that initializes the IPU would have been null for all but the concerned neighborhoods. Therefore, in this case, synthesized and observed data do not fit because both method (IPU) and initialization fail.

6. Conclusions and Recommendations

This chapter summarizes the findings of the research as they answer the research questions. It also reflects on the assumptions and limitations of the research. It finalizes by making recommendations for further research.

6.1. Conclusions

The research topic had been introduced by formulating the following main research question.

How can a synthetic firm population for the Province of South Holland be developed for microsimulation model?

Afterwards, a series of sub-questions were formulated to steer the research direction and facilitate answering the main research question. The sub-questions were:

1. ***What are the required attributes in the definition of a firm agent needed for a microsimulation model?***
2. ***How will different producer and consumer firm agents be characterized in the synthesis model?***
3. ***How will the spatial distribution of the firm population be characterized? I.e., what theory(ies) will be used to locate a firm of specific attribute set?***
4. ***What technique of population synthesis will be selected to synthesize the firm population in the research?***
5. ***How will a firm synthesis model be specified and implemented?***

The sub-questions have been answered as follows. The first two questions served the purpose of answering what type of firm-level attributes are required to understand a firm's freight transport demand. *The literature review section answers the first sub question by stating the pool of attributes that define a firm as: location, employee size, floor area, commodity type, economic sector, annual turnover and fleet size*. Regarding the second sub-question, differentiating between different types of firms, *the economic sectors of firms mainly decide their freight transport demand. Meanwhile commodity type will better qualify freight demand because implicit in it are: the economic sector of the firm, the inputs required, the market network for the commodity and important for logistics modeling, the type of transport the commodity can use.* (Comi et al., 2012) *refer to these as "the underlying mechanisms of demand" and market that commodity-based freight models can capture. Meanwhile, firm size attributes (number of employees & floor area) can indicate variation in freight demand amongst firms in the same sector.*

To assess sector-level importance of firms, aggregate, commodity-based statistics of generated freight and financial transactions based on input-output tables were used. These tables revealed the need to look at all sectors of the firm population and not the main producers in isolation, because the origin-destination network forms the economic ties that are the underlying drivers of regarding freight demand.

It was further established that the spatial distribution of attributes is crucial because difference in the location between firms is what necessitates transport. *The location of firms, as a subject of the third sub-question, was related to urban economics*. The thesis of urban economic theory that is focused on in this research is that urban areas can provide the labor force, the services markets and infrastructural amenities to support a large population of firms. In addition, firms simultaneously choosing the same urban location results in further agglomeration benefits. in the case study urban density was found to explain firm distributions at the municipal level. The findings on the neighborhood level are also important because some sectors are well located but better indicators of firm distribution at neighborhood and finer levels are required. This answers the third sub-question.

The fourth sub-question inquires what population synthesis technique will be used to make the firm synthesis model. Three categories of synthesis techniques were explored. *For reasons of ease and practicality as well as the availability of spatially distributed data, iterative proportional updating (IPU) was chosen.* This method is a scaling algorithm that transforms sample population into a synthetic population.

The fifth sub-question was answered by specifying the firm synthesis model based on the specification of the agent. The attribute list for this model was downsized to location, economic sector and employee size as they were the only data points that could be synthesized from openly available data using the IPU algorithm. Furthermore, the spatial distribution of economic sectors was the only data point per agent for which there was observed data to validate against.

The model has been validated with promising results apart from the agricultural sector that has negative correlation with urban density. The intricacies of firms' spatial distribution are mentioned above in relation to the third research question. Considering these nuances, it is likely that more detailed data can reveal better heterogeneity among firm types thereby showing more gaps in the indicators that can reflect the capacity of a neighborhood to support and sustain a firm of a certain attribute set.

In an ideal case where a sample of agents were available to initialize the model, these very same sample agents will be used to synthesize the population after the IPU fitting is complete; i.e., the IPU method gives cell counts of joint distributions. But they are not actual firm agents. Monte – Carlo simulation techniques that use IPU results for importance sampling must be used to generate the synthetic population of agents (Axhausen & Zurich, 2010).

Attempt was made to expand the number of attributes by synthesizing the employee size distributions from national statistics. This is likely to be unrealistic as the spatial scales are too varied; however, it is a demonstration that data from different aggregation levels can be used to generate proportions.

As a concluding remark on model usage, the model can be used to inform policy regarding freight transport demand at the firm level. Since the results can also be aggregated at any desired level, they can inform aggregate estimations of freight demand as it relates to land use. When coupled with decision behavior models, the synthesized population of firms can be used to build grouped profiles of the population regarding several logistics choice decisions.

6.2. Recommendations

The model produced in this research is founded on a set of conceptual and scoping assumptions. Extension and/or relaxation of these assumptions opens opportunities for improvement of this model and for further research. Accordingly, here are some recommendations.

- Actors and Interactions

The research was scoped such that the actors of the model are only examined from a freight transport demand perspective. Their logistics decisions on vehicle type, shipment size, choice of logistics service provider, etc. which dictate their interactions amongst each other are not studied. Including these decisions can further qualify the firm agent and can expand the pool of firm attributes mentioned in this study. This will be particularly evident if we compare the ability of the listed firm attributes in this research to predict freight generation to their ability to predict freight trip generation, as the latter is strongly reliant on the interactions of logistic actors (Holguín-Veras et. al., 2011). However, (Samimi et al., 2014) warn of the computational intensity that comes following a wide attribute set; thus, a broader study on a list of firm attributes will finally have to prioritize attributes and downsize.

- Limitations of open data

As was discussed in the data section, the openly available data are not found at firm level, but they are rather found at the neighborhood level. This has limited the structure of the model to disaggregation at neighborhood level both from synthesis and validation aspects. This is also closely tied to the IPU method in that its synthetic data is disaggregated as finely as the available microsample. Disaggregation can be extended further by deriving urban density values for smaller geographic units. Furthermore, the author recommends future research on firm-level freight transport demand be conducted using firm-level data which can be used after anonymization by a population synthesis technique of choice.

- Capabilities of the IPU method

This research on firm synthesis aimed to produce a population of firms for a firm-level prediction of freight transport demand. The method of choice for the synthesis model is IPU. This method is a simple *scaling algorithm* that iteratively *inflates a sample into a population* based on aggregate constraints. Thus, a micro-sample of a given disaggregation level yields a population at that disaggregation level. As all the necessary data points are not openly available at the firm level, special databases which contain firm-level information must be used to truly synthesize the population; e.g. a payable registry database on businesses in The Netherlands provides economic activities of each firm (KVK, 2019). Afterwards, the synthesized population is in effect anonymous and hence, can be used for agent-based freight transport studies.

- Use of a Simulation Approach

Simulation based population synthesis techniques utilize selected sampling techniques to match aggregate constraints of the observed population data. These synthesis techniques do not necessarily need samples although they can benefit from them. Simulation methods are, hence good alternatives to IPU where data requirements are not satisfied. This is with the caveat on computation time.

- Urban density and economic sectors

In this research, urban density has been used as a proxy variable to replace the distributions that could be obtained from a sample of firms; hence, implicitly, it was assumed firms and urban populations are spatially distributed in the same way. This choice is backed by assertions in urban economic theory that firms prefer urban locations and furthermore, they experience positive externalities by agglomeration. However, not all economic sectors have a positive correlation with urban density. Agriculture is a good example. Furthermore, there is also notable variation amongst firm types that prefer populated areas. The age of the firm and land rent are important factors that can create these variations, for instance. It is also important to note interactions between firms that have competitive or collaborative effects that either attract or push firms to/from a given location.

As for the IPU, the best way to initialize it remain observed microsamples instead of exogenous variables such as urban density. However, if further research is also conducted based on openly available data, finely disaggregated land use data (e.g. distribution of agricultural land) can augment urban densities.

- Bundling of Economic Sectors

The IPU model was estimated on a set of bundled economic sectors. This bundling was not made by the researcher but rather was present in the available data. As the bundling was not done, to the researcher’s best knowledge, with the purpose of investigating freight transport in mind, it posed undue constraints and consequent simplifications to render a proper sectoral distribution. It is recommended, however, that if data is available, to estimate models per sector. If there is utility in any bundling or unbundling of sectors, it must be done upon proper argumentation in relation to estimating in freight transport demands.

Bibliography

Abed, O., Bellemans, T., Cho, S., Janssens, G. K., Janssens, D., & Wets, G. (2014). A Bottom up Approach to Estimate Production-consumption Matrices from a Synthetic Firm Population Generated by Iterative Proportional Updating. *Transportation Research Procedia*, *1*(1), 49–56. https://doi.org/10.1016/j.trpro.2014.07.006

Alvim, A. C. F., Aloise, D. J., & Glover, F. (2001). A Hybrid Improvement Heuristic for the Bin Packing Problem. *4th Metaheuristics International Conference*, *10*(2), 63–68.

Axhausen, K. W., & Zurich, E. (2010). *Population synthesis for microsimulation: State of the art*. Retrieved from https://www.researchgate.net/publication/228973867

Bartels, R. (2015). *Re-interpreting R-squared , regression through the origin , and weighted least squares Re-interpreting R 2 , regression through the origin , and weighted least squares Robert B ARTELS University of Sydney Business School N OVEMBER 2015 Author ' s footnote*. (October).

Bastida, C., & Holguín-Veras, J. (2009). Freight generation models: Comparative analysis of regression models and multiple classification analysis. *Transportation Research Record*, (2097), 51–61. https://doi.org/10.3141/2097-07

Beckman, R. J., Baggerly, K. A., & McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*. https://doi.org/10.1016/0965-8564(96)00004-3

Ben-Akiva, M., & de Jong, G. (2008). The Aggregate–Disaggregate–Aggregate (ADA) Freight Model System. In *Recent Developments in Transport Modelling* (pp. 117–134). https://doi.org/10.1108/9781786359537-007

Bishop, Y. M., Holland, P. W., & Fienberg, S. E. (2007). Discrete multivariate analysis theory and practice. In *Discrete Multivariate Analysis Theory and Practice*. https://doi.org/10.1007/978-0-387-72806-3

Brouwer, A., Mariotti, I., & van Ommeren, J. (2002). www.econstor.eu. *42nd Congress of the European Regional Science Association: "From Industry to Advanced Services - Perspectives of European Metropolitan Regions", August 27th - 31st, 2002, Dortmund, Germany*. Retrieved from http://hdl.handle.net/10419/115677

Cabral, L. M. B., & Mata, J. (2003). On the evolution of the firm size distribution: Facts and theory. *American Economic Review*, *93*(4), 1075–1090. https://doi.org/10.1257/000282803769206205

CBS. (2012). *Toelichting kerncijfers wijken en buurten*. 1–36. Retrieved from https://opendata.cbs.nl/statline/#/CBS/nl/dataset/84286NED/table?dl=18E0D

Choupani, A. A., & Mamdoohi, A. R. (2016). Population Synthesis Using Iterative Proportional Fitting (IPF): A Review and Future Research. *Transportation Research Procedia*, *17*, 223–233. https://doi.org/10.1016/j.trpro.2016.11.078

Comi, A., Site, P. D., Filippi, F., & Nuzzolo, A. (2012, August). Urban freight transport demand modelling: A state of the art. *European Transport - Trasporti Europei*, pp. 1–17.

Davydenko, I. Y., Tavasszy, L. A., & Blois, C. J. De. (2013). *Modeling interregional freight flow by distribution systems*. 1–21.

De Bok, M., & Tavasszy, L. (2018). An empirical agent-based simulation system for urban goods transport (MASS-GT). *Procedia Computer Science*, *130*, 126–133. https://doi.org/10.1016/j.procs.2018.04.021

De Bok, M., Tavasszy, L., Bal, I., & Thoen, S. (2018). *ScienceDirect The incremental development path of an empirical agent-based simulation system for urban goods transport (MASS-GT)-review under responsibility of WORLD CONFERENCE ON TRANSPORT RESEARCH SOCIETY*. Retrieved from www.sciencedirect.comwww.elsevier.com/locate/procedia2352-1465

De Bok, M., & Van Oort, F. (2011). Agglomeration economies, accessibility and the spatial choice behavior of relocating firms. *Journal of Transport and Land Use*, *4*(1), 5. https://doi.org/10.5198/jtlu.v4i1.144

de Oliveira, L. K., Nóbrega, R. A. de A., Ebias, D. G., & Corrêa, B. G. e. S. (2017). Analysis of freight trip generation model for food and beverage in Belo Horizonte (Brazil). *Region*, *4*(1), 17–30. https://doi.org/10.18335/region.v4i1.102

Eastman, C. R. (1980, January 2). A review of freight modelling procedurest. *Transportation Planning and Technology*, Vol. 6, pp. 159–168. https://doi.org/10.1080/03081068008717186

Estimation Methods of Urban Freight Travel Demand Background and Goals Background. (2015).

Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological*, *58*, 243–263. https://doi.org/10.1016/j.trb.2013.09.012

Fienberg, S. E. (1970). An Iterative Procedure for Estimation in Contingency Tables. *The Annals of Mathematical Statistics*, *41*(3), 907–917. https://doi.org/10.1214/aoms/1177696968

Gerber, P., Caruso, G., Cornelis, E., & de Chardon, C. M. (2018). A multi-scale fine-grained luti model to simulate land-use scenarios in Luxembourg. *Journal of Transport and Land Use*, *11*(1), 255–272. https://doi.org/10.5198/jtlu.2018.1187

Harland, K., Heppenstall, A., Smith, D., & Birkin, M. (2012). *Creating realistic synthetic populations at varying spatial scales : A comparative critique of population synthesis techniques*.

Hastings, W. K. (1970). Monte carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109. https://doi.org/10.1093/biomet/57.1.97

Holguín-Veras, J., Xu, N., de Jong, G., & Maurer, H. (2011). An Experimental Economics Investigation of Shipper-carrier Interactions in the Choice of Mode and Shipment Size in Freight Transport. *Networks and Spatial Economics*, *11*(3), 509–532. https://doi.org/10.1007/s11067-009-9107-x

Iding, M. H. E., Meester, W. J., & Tavasszy, L. A. (n.d.). *Advanced Services-Perspectives of European Metropolitan Regions*. Retrieved from http://hdl.handle.net/10419/115840www.econstor.eu

Jacobs, J. (1969). *The Economy of Cities*.

Korf, R. E. (2002). *A New Algorithm for Optimal Bin Packing Introduction and Overview*. 731–736. Retrieved from www.aaai.org

KVK. (2019). Bedrijfsprofiel. Retrieved September 9, 2019, from https://www.kvk.nl/producten-bestellen/bedrijfsproducten-bestellen/bedrijfsprofiel/

Lee, D. H., & Fu, Y. (2011). Cross-entropy optimization model for population synthesis in activity-based microsimulation models. *Transportation Research Record*, (2255), 20–27. https://doi.org/10.3141/2255-03

Lomax, N., & Norman, P. (2016). Estimating population attribute values in a table: “Get me started in” iterative proportional fitting. *Professional Geographer*, *68*(3), 451–461. https://doi.org/10.1080/00330124.2015.1099449

Ma, J., Mitchell, G., & Heppenstall, A. (2015). Exploring transport carbon futures using population microsimulation and travel diaries: Beijing to 2030. *Transportation Research Part D: Transport and Environment*, *37*, 108–122. https://doi.org/10.1016/j.trd.2015.04.020

Macgill, S. M. (1977). Theoretical Properties of Biproportional Matrix Adjustments. *Environment and Planning A: Economy and Space*, *9*(6), 687–701. https://doi.org/10.1068/a090687

Marshall, A. (1920). *Principles of Economics*.

Moeckel, R., Wegener, M., & Spiekermann, K. (2003). Creating a synthetic Population. *Proceedings of the 8th International Conference on Computers in Urban Planning and*

Urban Management (CUPUM), (May), 1–18. Retrieved from http://www.spiekermann-wegener.com/pub/pdf/CUPUM_2003_Synpop.pdf

Moses, L. N. (1958). Oxford University Press. *The Quarterly Journal of Economics*, *72*(5794), 259–272. Retrieved from https://www.jstor.org/stable/1880599

Novak, D. C., Hodgdon, C., Guo, F., & Aultman-Hall, L. (2011). Nationwide Freight Generation Models: A Spatial Regression Approach. *Networks and Spatial Economics*, *11*(1), 23–41. https://doi.org/10.1007/s11067-008-9079-2

O'Sullivan, A. (2003). *[Arthur_O'Sullivan]_Urban_Economics(z-lib.org)*.

Otani, N., Sugiki, N., Vichiensan, V., & Miyamoto, K. (2012). Modifiable attribute cell problem and solution method for population synthesis in land use microsimulation. *Transportation Research Record*, (2302), 157–163. https://doi.org/10.3141/2302-17

Pitfield, D. E. (1978). Sub-optimality in freight distribution. *Transportation Research*. https://doi.org/10.1016/0041-1647(78)90028-X

Pritchard, D. R., & Miller, E. J. (2012). Advances in population synthesis: Fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation*, *39*(3), 685–704. https://doi.org/10.1007/s11116-011-9367-4

Rich, Jeppe, & Mulalic, I. (2012). Generating synthetic baseline populatio.. *Transportation Research Part A: Policy and Practice*, *Volume 46*(Issue 3), Pages 467-479.

Ruschendorf, L. (1995). Convergence of the Iterative Proportional Fitting Procedure. *The Annals of Statistics*, *23*(4), 1160–1174. https://doi.org/10.1214/aos/1176324703

Ryan, J., Maoh, H., & Kanaroglou, P. (2009). Population synthesis: Comparing the major techniques using a small, complete population of firms. *Geographical Analysis*, *41*(2), 181–203. https://doi.org/10.1111/j.1538-4632.2009.00750.x

Samimi, A., Mohammadian, A., Kawamura, K., & Pourabdollahi, Z. (2014). An activity-based freight mode choice microsimulation model. *Transportation Letters*, *6*(3), 142–151. https://doi.org/10.1179/1942787514Y.0000000021

Tavasszy, L. A. (2006). *Freight Modeling An Overview of International Experiences*.

Tavasszy, L., & de Jong, G. (2013). Modelling Freight Transport. In *Modelling Freight Transport*. https://doi.org/10.1016/C2012-0-06032-2

Van den Heuvel, F. P., de Langen, P. W., van Donselaar, K. H., & Fransoo, J. C. (2013). Spatial concentration and location dynamics in logistics: The case of a Dutch province. *Journal of Transport Geography*, *28*, 39–48. https://doi.org/10.1016/j.jtrangeo.2012.10.001

Van Der Panne, G., & Van Beers, C. (2006). *On the Marshall-Jacobs controversy: it takes two to tango*.

Weber, A., & Friedrich, C. (1929). *Alfred Weber's Theory of Location of Industries*.

Xin Ye, Karthik Konduri, Ram M. Pendyala, Bhargava Sana, P. W. (2009). *PopulationSynthesizerPaper_TRB*. *9600*(206).

Yasuda, T. (2005). Firm growth, size, age and behavior in Japanese manufacturing. *Small Business Economics*, *24*(1), 1–15. https://doi.org/10.1007/s11187-005-7568-y

Appendix

Appendix A: SBI2008 Economic Sectors and Commodities

Sector Category (SBI 2007)	NST 2007 Commodities/Economic activities
A - Agriculture, forestry and fishing	Crop, animal production, hunting and related activities, Forestry and logging, Fishing and aquaculture
B - Mining and quarrying	Extraction of crude petroleum and natural gas, Other mining and quarrying, except petroleum and gas,
C - Manufacturing	Manufacture of food products, Manufacture of beverages, Manufacture of tobacco products, Manufacture of textiles, wearing apparel and leather, Manufacture of wood and products, except furniture, Manufacture of paper and paper products, Printing and reproduction of recorded media Manufacture of coke and refined petroleum products Manufacture of chemicals and chemical products Manufacture of pharmaceutical products and preparations, Manufacture of rubber and plastic products, Manufacture of other non-metallic mineral products, Manufacture of basic metals, Manufacture of metal products, except machinery, equipment, Manufacture of computer, electronic and optical products, Manufacture of electrical equipment Manufacture of machinery and equipment n.e.c. Manufacture of motor vehicles, trailers and semi-trailers Manufacture of other transport equipment Manufacture of furniture, Other manufacturing, Repair and installation of machinery and equipment
D - Electricity, gas, steam and air conditioning supply	Electricity, gas, steam and air conditioning supply
E - Water supply; sewerage, waste management and remediation activities	Water collection, treatment and supply, Sewerage, waste management, materials recovery activities
F - Construction	Construction of buildings, Civil engineering, Specialized construction activities
G - Wholesale and retail trade; repair of motor vehicles and motorcycles	Trade and repair of motor vehicles and motorcycles, Wholesale trade, except of motor vehicles and motorcycles, Retail trade, except of motor vehicles and motorcycles
H - Transportation and storage	Land transport and transport via pipelines, Water transport, Air transport, Warehousing and support activities for transportation, Postal and courier activities
I - Accommodation and food service activities	Accommodation, Food and beverage service activities
J - Information and communication	Publishing activities, Audiovisual production, programming and broadcasting, Telecommunications, Computer programming, consultancy and related activities, Information service activities
K - Financial institutions	Financial service activities, except insurance and pension funding, Insurance and pension funding, except compulsory social security, Activities auxiliary to financial services and

	insurance activities, Real estate activities, Legal and accounting activities
L - Renting, buying and selling of real estate	Real estate activities
M - Consultancy, research and other specialised business services	Activities of head offices; management consultancy activities, Architectural, engineering activities; technical testing and analysis Scientific research and development, Advertising and market research, Other professional, scientific and technical activities, Veterinary activities
N - Renting and leasing of tangible goods and other business support services	Rental and leasing activities
O - Public administration, public services and compulsory social security	Security and investigation activities, Office administrative and other business support activities, Public administration and defence; compulsory social security
P - Education	Education
Q - Human health and social work activities	Human health activities, Residential care and social work activities
R - Culture, sports and recreation	Arts, entertainment, cultural, gaming and betting activities, Sports activities and amusement and recreation activities, Activities of membership organisations
S - Other service activities T - Activities of households as employers; undifferentiated goods and service-producing activities of households for own use ***[These are coupled for reasons of similarity]***	Services to buildings and landscape activities, Employment activities, Travel agency, tour operator reservation and related activities, Repair of computers and personal and household goods, Other personal service activities
U - Extraterritorial organisations and bodies	N/A

Appendix B: Matching of SBI2008 Economic Sectors and NST 2007 Commodity Groups

NST 2007 Group	Description	Freight Volume (1000 tons)	Transported Freight (million ton-km)	SBI2008 Producer (researcher's opinion)	SBI2008 Consumer (researcher's opinion)
1	Products of agriculture, hunting, and forestry; fish and other fishing products	57,714	8,919	A	Several: C, G, I
2	Coal and lignite; crude petroleum and natural gas	0	0	B	Several: B, C, D
3	Metal ores and other mining and quarrying products; peat; uranium and thorium	90,379	3707	B	C
4	Food products, beverages and tobacco	124106	13630	C	Several: G, I
5	Textiles and textile products; leather and leather products	4158	505	C	Several: C, G
6	Wood and products of wood and cork (except furniture); articles of straw and plaiting materials; pulp, paper and paper products; printed matter and recorded media	26720	3888	A, C	G, J
7	Coke and refined petroleum products	9195	910	C	All sectors
8	Chemicals, chemical products, and man-made fibers; rubber and plastic products; nuclear fuel	60920	7292		All sectors
9	Other nonmetallic mineral products	60686	5148	C	F
10	Basic metals; fabricated metal products, except machinery and equipment	24904	3496	C	C
11	Machinery and equipment n.e.c.; office machinery and computers; electrical machinery and apparatus n.e.c.; radio, television and communication equipment and apparatus; medical, precision and optical instruments; watches and clocks	24934	2890	C	All sectors
12	Transport equipment	8514	1953	C	C, G, H

13	Furniture; other manufactured goods n.e.c.	3336	544	C	All sectors
14	Secondary raw materials; municipal wastes and other wastes	43941	2827	C	C, E
15	Mail, parcels	3174	257	***[categories here are either ubiquitous across sectors or too generically named to assign to specific sector categories]***	
16	Equipment and material utilized in the transport of goods	35350	3253		
17	Goods moved in the course of household and office removals; baggage and articles accompanying travellers; motor vehicles being moved for repair; other non-market goods n.e.c.	680	48		
18	Grouped goods: a mixture of types of goods which are transported together	46621	8476		
19	Unidentifiable goods: goods which for any reason cannot be identified and therefore cannot be assigned to groups 01-16.	16231	1517		
20	Other goods n.e.c.	0	0	This is a miscellaneous null category	

Appendix – B: I/O Tables of 2015: The Netherlands

	Shares of Transactions by Inputs																		
Sectors in SBI 2008	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S+T
A	32.37%	0.05%	9.63%	0.03%	0.28%	0.24%	0.51%	0.64%	2.03%	0.04%	0.01%	0.06%	0.00%	0.06%	0.65%	0.26%	1.15%	0.61%	0.98%
B	2.75%	31.01%	1.52%	7.30%	0.26%	0.51%	0.54%	0.20%	1.56%	0.03%	0.12%	0.24%	0.23%	0.20%	0.33%	1.32%	0.62%	1.32%	0.46%
C	39.96%	7.89%	52.73%	9.61%	14.62%	24.46%	8.48%	11.45%	30.44%	11.48%	2.24%	2.18%	3.49%	11.78%	7.49%	5.96%	12.79%	9.25%	9.22%
D	1.97%	23.53%	2.52%	34.22%	2.72%	0.14%	1.69%	0.83%	4.01%	0.56%	0.39%	0.49%	0.78%	0.23%	0.97%	2.15%	0.93%	2.95%	0.58%
E	3.14%	0.48%	1.59%	0.85%	40.19%	0.83%	0.34%	0.41%	0.65%	0.19%	0.19%	0.08%	0.41%	0.14%	6.82%	0.73%	2.31%	0.97%	0.16%
F	2.54%	2.03%	1.01%	14.15%	2.25%	53.18%	1.40%	3.07%	1.12%	2.07%	0.11%	19.86%	0.78%	0.82%	16.77%	9.36%	4.45%	2.72%	0.55%
G	3.20%	1.45%	2.99%	2.14%	6.08%	1.89%	14.30%	3.72%	1.67%	3.87%	0.73%	0.26%	1.67%	49.46%	1.42%	1.22%	1.61%	2.02%	1.05%
H	1.86%	0.43%	3.52%	0.82%	8.71%	0.35%	9.40%	44.00%	1.09%	2.01%	2.45%	0.42%	1.59%	1.49%	3.71%	2.68%	4.43%	1.94%	10.63%
I	0.03%	0.18%	0.24%	0.28%	0.13%	0.13%	0.82%	1.46%	3.62%	0.45%	0.33%	0.05%	0.76%	0.17%	0.72%	0.86%	6.10%	1.94%	1.06%
J	0.96%	1.93%	2.25%	3.71%	2.95%	2.15%	7.98%	2.95%	4.31%	39.62%	7.77%	1.17%	8.86%	4.90%	7.54%	10.24%	6.82%	6.25%	5.73%
K	3.91%	6.06%	4.14%	5.07%	3.38%	3.42%	11.18%	5.28%	5.80%	5.78%	57.51%	63.20%	14.13%	8.74%	16.52%	9.13%	9.09%	13.39%	7.29%
L	0.73%	0.43%	1.30%	1.03%	1.50%	2.34%	9.25%	2.93%	11.96%	2.98%	4.31%	6.54%	11.77%	2.85%	1.33%	3.58%	5.27%	3.10%	3.57%
M	3.27%	8.80%	7.43%	10.56%	4.71%	5.39%	21.09%	5.81%	10.24%	15.69%	9.46%	2.83%	31.54%	6.43%	11.02%	5.93%	2.89%	8.60%	19.74%
N	0.01%	7.66%	0.69%	0.15%	0.41%	1.05%	0.96%	4.19%	0.48%	1.40%	0.56%	0.14%	2.89%	4.23%	0.52%	0.73%	0.74%	2.92%	0.91%
O	1.82%	0.89%	1.30%	1.54%	3.45%	0.50%	3.32%	3.19%	3.15%	2.86%	1.77%	0.50%	2.81%	1.89%	8.15%	6.01%	5.63%	3.08%	2.39%
P	0.02%	0.18%	0.25%	0.44%	0.32%	0.21%	0.36%	0.33%	0.40%	0.47%	1.66%	0.05%	7.09%	0.23%	3.45%	9.22%	1.81%	2.95%	0.74%
Q	0.07%	0.13%	0.14%	0.05%	0.06%	0.13%	0.52%	0.58%	0.28%	0.16%	0.23%	0.03%	0.38%	0.11%	3.35%	0.85%	16.52%	0.29%	0.67%
R	0.43%	0.13%	0.35%	0.33%	0.68%	0.27%	1.05%	0.27%	2.74%	2.79%	0.98%	0.23%	1.70%	0.56%	1.35%	2.76%	2.45%	26.22%	2.13%
S + T	0.94%	6.77%	6.43%	7.75%	7.32%	2.83%	6.82%	8.69%	14.44%	7.53%	9.18%	1.69%	9.11%	5.72%	7.89%	27.01%	14.40%	9.50%	32.15%

	Shares of Transactions by Outputs																		
Sectors in SBI 2008	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S+T
A	30.77%	0.01%	61.21%	0.01%	0.09%	0.66%	1.46%	1.21%	0.95%	0.05%	0.01%	0.14%	0.00%	0.01%	1.03%	0.11%	1.14%	0.30%	0.82%
B	8.40%	26.14%	30.99%	9.55%	0.26%	4.49%	4.92%	1.22%	2.35%	0.15%	0.51%	1.94%	0.19%	0.15%	1.69%	1.73%	1.99%	2.07%	1.26%
C	6.57%	0.36%	58.00%	0.68%	0.79%	11.63%	4.18%	3.76%	2.47%	2.66%	0.51%	0.94%	0.16%	0.48%	2.06%	0.42%	2.20%	0.78%	1.35%
D	3.37%	11.13%	28.84%	25.11%	1.52%	0.70%	8.66%	2.83%	3.39%	1.34%	0.94%	2.19%	0.37%	0.10%	2.79%	1.58%	1.67%	2.59%	0.89%
E	6.59%	0.28%	22.25%	0.76%	27.57%	5.06%	2.13%	1.73%	0.68%	0.56%	0.56%	0.47%	0.23%	0.07%	23.99%	0.66%	5.06%	1.04%	0.29%
F	0.92%	0.20%	2.44%	2.20%	0.27%	55.76%	1.52%	2.22%	0.20%	1.06%	0.05%	19.00%	0.08%	0.07%	10.19%	1.46%	1.69%	0.50%	0.18%
G	2.93%	0.36%	18.23%	0.84%	1.82%	4.99%	39.14%	6.80%	0.75%	4.97%	0.92%	0.62%	0.42%	11.22%	2.17%	0.48%	1.53%	0.95%	0.85%
H	1.05%	0.07%	13.28%	0.20%	1.61%	0.57%	15.93%	49.70%	0.30%	1.60%	1.93%	0.63%	0.25%	0.21%	3.51%	0.65%	2.61%	0.56%	5.34%
I	0.17%	0.24%	7.72%	0.57%	0.20%	1.85%	11.83%	14.08%	8.59%	3.07%	2.22%	0.61%	1.01%	0.20%	5.83%	1.79%	30.69%	4.78%	4.55%
J	0.59%	0.33%	9.24%	0.97%	0.59%	3.82%	14.68%	3.62%	1.30%	34.20%	6.61%	1.89%	1.50%	0.75%	7.75%	2.70%	4.37%	1.97%	3.12%
K	0.96%	0.41%	6.78%	0.53%	0.27%	2.42%	8.20%	2.58%	0.70%	1.99%	19.53%	40.83%	0.96%	0.53%	6.77%	0.96%	2.32%	1.68%	1.58%
L	0.72%	0.12%	8.49%	0.43%	0.48%	6.62%	27.15%	5.73%	5.77%	4.11%	5.86%	16.91%	3.19%	0.69%	2.18%	1.51%	5.39%	1.56%	3.10%
M	1.28%	0.95%	19.42%	1.77%	0.60%	6.09%	24.70%	4.54%	1.97%	8.63%	5.13%	2.92%	3.41%	0.62%	7.22%	1.00%	1.18%	1.72%	6.85%
N	0.05%	6.89%	15.00%	0.21%	0.43%	9.90%	9.37%	27.29%	0.78%	6.41%	2.55%	1.23%	2.60%	3.42%	2.83%	1.03%	2.51%	4.88%	2.64%
O	2.73%	0.37%	13.00%	0.99%	1.69%	2.16%	14.93%	9.56%	2.33%	6.03%	3.69%	1.96%	1.16%	0.70%	20.47%	3.88%	8.82%	2.37%	3.18%
P	0.09%	0.21%	7.00%	0.80%	0.45%	2.53%	4.62%	2.80%	0.83%	2.83%	9.80%	0.60%	8.34%	0.24%	24.64%	16.90%	8.08%	6.44%	2.80%
Q	0.24%	0.12%	3.25%	0.07%	0.07%	1.25%	5.32%	4.02%	0.48%	0.77%	1.11%	0.24%	0.36%	0.10%	19.34%	1.25%	59.43%	0.51%	2.05%
R	1.18%	0.09%	6.38%	0.38%	0.61%	2.11%	8.52%	1.48%	3.66%	10.67%	3.68%	1.63%	1.27%	0.38%	6.15%	3.23%	6.95%	36.49%	5.13%
S + T	0.45%	0.89%	20.50%	1.58%	1.14%	3.91%	9.75%	8.28%	3.40%	5.06%	6.08%	2.13%	1.20%	0.68%	6.31%	5.55%	7.17%	2.32%	13.62%