Kurzfassung

Name:	Guangxi Shi
Thema:	Vervollständigung logistischer Standortdaten für die Modellierung von Transportketten
Betreuer:	Prof. Dr.-Ing. Manfred Boltze
	M.Sc. Kevin Rolko

Seit 1995 wächst die Logistikbranche und ihr Marktvolumen hat sich bis 2011 im Vergleich zum Jahr 1995 fast verdoppelt. Dabei nehmen die Güterverkehrsflüsse, insbesondere der Anteil des Straßengüterverkehrs, stetig zu. Für die weitere Entwicklung bis 2025 oder sogar 2030 sind viele Prognosen verschiedener Organisationen somit sehr optimistisch. Diese Tendenz kann die wirtschaftliche Konjunktur sowohl bundesweit als auch interkontinental anreizen. Aber andererseits bringt sie der menschlichen Gesellschaft auch Nebenwirkungen wie umweltschädliche Immission, eine Zunahme der Staus, vermehrte Unfälle u. s. w.

Um die verkehrlichen Auswirkungen der Logistikbranche zu kontrollieren und dementsprechend geeignete verkehrliche Maßnahmen zu ermitteln, werden häufig Güterverkehrsmodelle angewandt, durch die die verkehrlichen Auswirkungen auf das Verkehrsnetzwerk, etwa bei einem neuen Bauprojekt oder einer Neuerung der Verkehrspolitik, mathematisch berechnet werden.

Nach Köhler 2001 können die Güterverkehrsmodelle in 2 Typen kategorisiert werden:

- Güterbezogene Modelle
- Fahrtenbezogene Modelle

In den güterbezogenen Modellen werden die Güterflüsse (in Tonnen) zwischen ihrer Quelle und ihrem Ziel modelliert. Danach werden sie durch Ermittlung der ausführlichen Logistikentscheidungen in die Verkehrsflüsse umgerechnet. Solche Modelle fordern detaillierte Daten über die involvierten Logistikakteure, um die vielfältigen Logistikaktivitäten logisch zu beschreiben. Demgegenüber überspringen die fahrtenbezogenen Modelle die Modellierung der Güterflüsse. Die Verkehrsflüsse werden über unterschiedliche Regressionsmodelle anhand der logistischen Attribute direkt ermittelt. Trotzdem müssen die Eingangsdaten für solche Modelle im Voraus gut differenziert werden, sodass das Bias aus der Aggregation unverträglicher Daten vermieden werden kann. Zusammenfassend kann man sagen, dass die modernen Güterverkehrsmodelle immer mehr Eingangsdaten fordern, folglich erhöht sich auch das Risiko, dass der Modellierer, auch nach intensiver Datenerhebung, noch eine lückenhafte Datenbank behandeln muss, bevor sie als Eingangsgröße ins Modell eingeleitet wird.

Um die verkehrlichen Auswirkungen von Logistikstandorten zu untersuchen, will das Institut für Verkehrsplanung und Verkehrstechnik an der TU Darmstadt (im Folgenden als IVV abgekürzt wird) auch ein eigenes Modell aufbauen. Dazu wird eine Logistikstandort-Datenbank mit 47 Attributen für momentan bundesweit 3757 Standorte durch sekundäre Datenerhebung aufgebaut. Da die öffentlichen Datenquellen sehr schwierig zu finden sind, leidet die Datenbank aktuell enorm unter fehlenden Daten. Deswegen muss eine Lösung gefunden werden, um die fehlenden Stellen in der Datenbank logisch einzuschätzen.

Zu einem ähnlichen Thema haben bereits viele Statistiker verschiedene Annahmensysteme konstruiert, basierend auf der Grundlage, die Robin 1976 angelegt hat. Die meisten davon beziehen sich auf die Beziehung zwischen den Modellparametern, die die Größen der Variablen kontrollieren, und den Parametern, die die Fehlenswahrscheinlichkeit kontrollieren. In dieser Arbeit muss der Fehlensprozess für ignorierbar gehalten werden. Die Modellparameter bleiben also in allen Zeilen unverändert, weil die Nachreichung echter Daten zur Überprüfung des Imputationsergebnis wegen Zeitknappheit unmöglich ist.

Basierend auf dem Annahmensystem können die fehlenden Stellen in der Datenbank durch die Maximierung der Wahrscheinlichkeitsfunktion der Datensätze eingeschätzt werden, indem zuerst die Modellparameters durch Lösung der Gleichung:

$$\frac{\partial L(\theta)}{\partial \theta} = 0$$

ermittelt und danach die fehlenden Werte mittels der Modellparameter berechnet oder ausgelost werden.

Da die Wahrscheinlichkeitsfunktion für die Datensätze mit fehlenden Stellen sehr umständlich aufzubauen und dementsprechend zu lösen ist, werden verschiedene vereinfachte Missing-Data-Methoden erfunden.

Im listenweisen und paarweisen Fallausschluss werden keine fehlenden Stellen eingeschätzt. Stattdessen wird die Korrelation zwischen den Variablen direkt aufgrund der schon beobachteten Daten ermittelt. Obwohl beide keine Imputation erzeugen können, stellen sie in den meisten Imputationsalgorithmen eine grundlegende Phase zur Schätzung des Anfangsparameters dar.

Die spaltenweisen Methoden bieten die einfachsten Wege, die fehlenden Stellen auszufüllen. Aber die Korrelation zwischen den Variablen wird in solchen Methoden vernachlässigt.

Um die Korrelation zwischen Variablen besser zu beschreiben, wird der EM-Algorithmus (Expectation-Maximization) angewendet, indem ein Zyklus zur Ermittlung des Modellparameters und der fehlenden Stellen wiederholt durchgeführt wird, bis eine Konvergenz oder ein Schwellenwert erreicht ist, der die Iteration stoppt. Einen Fortschritt, basierend auf dem EM-Algorithmus, macht die sequentielle allgemeine Regression durch ihre spaltenweise Regression und flexible Auswahl des Schätzmodells für jede einzuschätzende Variable.

Neben dem arithmetischen Schätzmodell spielt auch das Auslosungsmodell eine wichtige Rolle in Bezug auf die Imputation. Die PMM-Methode (Predictive Mean Matching) ist ein sehr beliebtes Beispiel dafür. Aufgrund des Schätzwertes für eine fehlende Stelle wird ein beobachteter Wert in derselben Variable direkt als Imputationsergebnis für diese Stelle ausgewählt.

Für der Imputationsaufgabe für die Logistikstandort-Datenbank des IVV wird die sequentielle allgemeine Regression als Imputationsmethode ausgewählt. Nach der Beschränkung der einzuschätzenden Attribute werden die 4 nachfolgenden Imputationsvarianten untersucht:

- Alle Attribute durch lineare Regression einschätzen (V1)
- Alle Attribute durch die PMM-Methode einschätzen (V2)
- Die Stammattribute direkt durch die Summe ihrer Subattribute berechnen, die anderen durch lineare Regression einschätzen (V1.1)
- Die Stammattribute direkt durch die Summe ihrer Subattribute berechnen, die anderen durch die PMM-Methode einschätzen (V2.1)

Die 4 Varianten werden durch 4 Experimente über ihre Leistungen in den nachfolgenden Feldern bewertet:

- Robustheit des einzelnen Imputationswertes
- Robustheit der Korrelation zwischen den Attributen
- Genauigkeit
- Vergleich mit schon aggregierten Daten aus anderen Datenquellen

Die Robustheit des einzelnen Imputationswertes wird über den Verlauf des Durchschnitts und des Standardfehlers der einzuschätzenden Attribute in Abhängigkeit von der Anzahl der Iterationen bewertet. Die Variante 2, die sowohl eine schnelle Stabilität als auch einen dünneren konvergierten Wertbereich erweist, erbringt die beste Leistung in dem ersten Experiment.

Im zweiten Experiment ergibt die Variante 2 auch die Datensätze, die eine relativ stabile Korrelation ermitteln können. Darüber hinaus können die verschiedenen Submengen von der ursprünglichen Datenbank auch die stabilste Korrelation ermitteln. Deswegen wird die Variante 2 in diesem Experiment auch als die beste bewertet.

Die Genauigkeit wird durch einen Vergleich zwischen den Imputationswerten für die Amputationsdatenbank und ihren ursprünglichen Werten bewertet. Die Variante 2, die in diesem Experiment am häufigsten die kleinste durchschnittliche Änderungselastizität ermittelt, hat auch hier den ersten Platz inne.

Zudem ermittelt die Variante 2 immer noch das Ergebnis, welches den aggregierten Daten des Fraunhofer Instituts am nächsten liegt..

In der Gesamtheit erbringt die Variante 2 die besten Leistungen in allen 4 Experimenten. Das Imputationsergebnis durch das gesamte Behandlungsverfahren (Vorbereitung, Imputation und Nachbearbeitung) mit der PMM-Methode als Schätzmodell für alle einzuschätzenden Attribute wird als das Endprodukt für die Imputationsaufgabe ausgewählt. Der Vergleich der Größenverteilung der Attribute von der ursprünglichen und der vervollständigten Datenbank sowie die Kartendarstellung wird darauf basiert visualisiert.

Abstract

Name:	Guangxi Shi
Theme:	Missing Data Analysis for the completion of logistics location data for modeling transport chains
Mentor:	Prof. Dr.-Ing. Manfred Boltze
	M.Sc. Kevin Rolko

The logistics industry has been growing since 1995, and its market volume has almost doubled by 2011 compared to 1995. From this, the freight traffic flows, in particular the share of road freight traffic, are steadily increasing. Many forecasts from different organizations are therefore very optimistic for their further increasing till 2025 or even 2030. On the one hand, this tendency can stimulate the economic situation nationwide even intercontinentally. But on the other hand, it also brings side effects such as environmental pollution, traffic jams, accidents, etc. to human society.

In order to measure the transport impact of the logistics industry and to identify appropriate transport measures, freight transport models are often built up by mathematical assessing the impact on the transport network from eg. a new construction project or new transport policy.

According to Köhler 2001, the freight transport models can be categorized into 2 types:

- Goods-related models
- Trip-related models

In goods-related models, flows of goods (in tonnes) are modeled between their source and destination. Thereafter, they are converted into the traffic flows by determining the detailed logistics decisions. Such models require detailed data about the involved LSPs in order to logically describe the diverse logistics activities. In contrast, the trip-related models skip the modeling of the flows of goods. The traffic flows are determined directly using different regression models on the basis of the logistical attributes. Nevertheless, the input data for such models must be well-differentiated in advance so that the bias of aggregation of incompatible data can be avoided. In summary, modern freight transport models demand more and more input data, which also increases the risk that the modeler, even after intensive data collection, still has to deal with a patchy database before it is introduced as input to the model.

In order to investigate the traffic effects of the logistics locations, the Institute of Transport Planning and Traffic Engineering at the TU Darmstadt (subsequently abbreviated as IVV) also wants to set up its own model. Before that, a logistics site database with 47 attributes for currently nationwide 3757 locations will be built by secondary data collection. Since the public data source is very difficult to find, the database currently suffers from enormous missing data. Because of this, a solution must be found to logically evaluate the missing parts in the database.

For the similar proposes, many statisticians have constellated various assumption systems based on the basis that Robin founded in 1976. Most of these relate to the relationship between the model parameters that control the quantity of variables and the parameters that control the likelihood of missing. In this work, the missing process must be considered ignorable so that the model parameters remain unchanged

in all records, because the replenishment of real data to check the imputation result due to time constraints is impossible.

Based on the assumption system, the missing parts in the database can be estimated by maximizing the probability function of the datasets. Firstly the model parameter should be determined by solving the equation:

$$\frac{\partial L(\theta)}{\partial \theta} = 0$$

Then the missing values are calculated or drawn according to the model parameter.

Since the probability function for the records with missing parts is normally very cumbersome to build and to solve, various simplified Missing Data methods are invented.

In the listwise and pairwise deletion, no missing digits will be estimated. Instead, the correlation between the variables is determined directly on the basis of the already observed data. Although the two can not produce an imputation, in most imputation algorithms they represent a fundamental phase for estimating the initial model parameter.

The column-wise methods are the easiest ways to fill in the missing places. But the correlation between the variables is neglected in such methods.

To describe the correlation between variables better, the EM (Expectation-Maximization) algorithm is applied by repeatedly performing a cycle of determining the model parameter and then the missing locations until convergence or until a threshold for stopping the iteration. And based on the EM algorithm the sequential general regression makes a progress by its column-wise regression and flexible selection of the estimation model for each variable to be estimated.

Apart from the arithmetic estimation model, the draw models also play an important role in imputation. The PMM method (Predictive Mean Matching) is a very popular example of these. Based on comparison between all the estimated values of the variable to be filled in, an observed value is directly selected as the imputation result for a blank location in the same variable.

For the imputation task for IVV's Logistics Location Database, the sequential general regression is selected as the imputation method. And after restricting the attributes to be estimated, the 4 following imputation variants are examined:

- Estimate all attributes by linear regression (V1)
- Assess all attributes by PMM method (V2)
- Calculate the base attributes directly by adding the sum of their subattributes, estimate the others by linear regression (V1.1)
- Calculate the master attributes directly by the sum of their sub-attributes, estimate the others by PMM method (V2.1)

The 4 variants are evaluated by 4 experiments on their performance in the following fields:

- Robustness of the individual imputation value
- Robustness of the correlation between attributes
- Accuracy
- Comparison with already aggregated data from another data source