6-2014

Continuous Data Integration for Land Use and Transportation Planning and Modeling

Liming Wang
Portland State University

Kihong Kim
Portland State University

Recommended Citation

Wang, Liming and Kihong Kim. Continuous Data Integration for Land Use and Transportation Planning and Modeling. NITC-RR-581. Portland, OR: Transportation Research and Education Center (TREC), 2014. https://doi.org/10.15760/trec.37

NITC
NATIONAL INSTITUTE for TRANSPORTATION and COMMUNITIES

FINAL REPORT

Continuous Data Integration for Land Use and Transportation Planning and Modeling

NITC-RR-581 ■ *June 2014*

NITC is the U.S. Department of Transportation's national university transportation center for livable communities.

TREC
Transportation Insight for Vibrant Communities

CONTINUOUS DATA INTEGRATION FOR LAND USE AND TRANSPORTATION PLANNING AND MODELING

Final Report

NITC-RR-581

by

Liming Wang and Kihong Kim
Portland State University

for

National Institute for Transportation
and Communities (NITC)
P.O. Box 751
Portland, OR 97207

June 2014

Technical Report Documentation Page

1. Report No. NITC-RR-581	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Continuous Data Integration for Land Use and Transportation Planning and Modeling		5. Report Date June 2014 6. Performing Organization Code
7. Author(s) Liming Wang and Kihong Kim Portland State University		8. Performing Organization Report No.
9. Performing Organization Name and Address		10. Work Unit No. (TRAIS) 11. Contract or Grant No.
12. Sponsoring Agency Name and Address National Institute for Transportation and Communities (NITC) P.O. Box 751 Portland, Oregon 97207		13. Type of Report and Period Covered 14. Sponsoring Agency Code
15. Supplementary Notes		

16. Abstract

There is an urgent need for improved models that address the interdependencies between land use and transportation, and considerable new work is underway to develop such models in Oregon and elsewhere. These models and planning practices to integrate land use into the process, however, require the integration of massive amounts of land use data that is messy and incomplete. There have been considerable advances in the treatment of such data problems in other domains, drawing on data mining and machine-learning techniques to address issues in various domains. To date, however, little systematic effort has applied these technological advances to the problem domain of land use and transportation data. Experience suggests that as much as 70% of the total effort in developing integrated land use and transportation models is directly or indirectly associated with data development, integration and cleaning, yet there is remarkably little systematic research focused on the development of reusable methods and tools to support this problem domain.

In coordination with an ongoing project of similar theme funded by University of California Transportation Center (UCTC), this project will focus on bringing interdisciplinary methods to develop land use datasets for integrated land use and transportation planning and modeling, with special attention on preserving temporal dimension of the data and monitoring data quality through indicators. Utilizing statistics and machine-learning techniques, we will develop reusable tools that will create a harmonized and coherent land use database from various public and private sources. These tools will be made available to the public and can be used by cities, counties, metropolitan planning agencies, state agencies, universities or anyone else needing to develop a usable database for use in integrated planning and modeling. All of the resulting data, with the exception of proprietary or confidential input data where there is no viable alternative, would be public and reusable for planning and research.

17. Key Words Transportation and Land Use Data, Transportation and Land Use Modeling, Data Integration		18. Distribution Statement No restrictions. Copies available from NITC: www.NITC.us	
19. Security Classification (of this report) Unclassified	20. Security Classification (of this page) Unclassified	21. No. of Pages 67	22. Price

ACKNOWLEDGEMENTS

The author gratefully acknowledges the National Institute for Transportation and Communities (NITC) and National Institute for Transportation and Communities (NITC) for sponsoring this project. In addition, the author is thankful to research assistant Kihong Kim and Eugenio Arriaga for their contributions and efforts. Any errors or omissions are the sole responsibility of the authors.

DISCLAIMER

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

Since the 1990s, federal legislation and local and state politics have changed the landscape for metropolitan planning of land use and transportation. There is an urgent need for improved models that address the interdependencies between land use and transportation, and considerable new work is underway to develop such models by Metropolitan Planning Organizations (MPOs). These models and planning practices to integrate land use into the process, however, require the integration of massive amounts of land use and transportation data that is messy and incomplete. There have been considerable advances in the treatment of such data problems, drawing on data mining and machine-learning techniques to address issues in various domains. To date, however, little systematic effort has applied these technological advances to the problem domain of land use and transportation data. Experience suggests that as much as 70% of the total effort in developing integrated land use and transportation models is directly or indirectly associated with data development, integration and cleaning, yet there is remarkably little systematic research focused on the development of reusable methods and tools to support this problem domain.

This report presents the results of a data integration project aiming to address this challenge. It focuses on bringing interdisciplinary methods to make the best use of available data in land use and transportation. Utilizing statistics and machine-learning techniques, we develop reusable tools that will create a harmonized and coherent land use database from various public and private sources.

This project develops a continuous approach and reusable tools for data integration. As data sources increasingly update more frequently and even in real time, and policy decisions demand timely data and more rigorous analysis, data collection and updating increasingly become a continuous process. The traditional practice adopted by most planning agencies updates datasets as a one-off effort at long intervals, and is very costly and increasingly challenged. We pay special attention to preserving temporal dimension of the data and monitoring data quality through automated data-quality indicators following continuous software testing procedures, so that the quality information is available at all times as data sources are being integrated on a continuous basis. In addition, the historical dimension also provides useful information to better clean data and enable an understanding of historic trends.

These tools will be reusable for different time periods and for different regions. We report the experiences of testing and applying the approach and tools to our test metropolitan regions, including the Portland, OR, and San Francisco Bay Area.

The data structure, tools for data processing and quality monitoring, and documentation are available to the public and can be used by cities, counties, metropolitan planning agencies, state agencies, universities or anyone else needing to develop a usable database for use in integrated planning and modeling.

1.0 INTRODUCTION

Since the 1990s, federal legislation and local and state politics have changed the landscape for metropolitan planning of land use and transportation, and Metropolitan Planning Organizations (MPOs) are scrambling to react. There is an urgent need for improved models that address the interdependencies between land use and transportation at state and regional levels, and considerable new work is underway to develop such models (See, for example, ULTRANS ITS UC Davis and HBA Specto Incorporated, 2011; Waddell et al., 2010; Weidner et al., 2009). These models and planning practices to integrate land use and transportation, however, require the integration of massive amounts of land use and socio-economic data that is messy and incomplete. There is a suggestion that as much as 70% of the total effort in developing integrated land use and transportation models is directly or indirectly associated with data development, integration and cleaning (Waddell et al., 2005).

Even with such excruciating efforts, current practice of data development in most planning agencies is largely ad-hoc and un-reusable. This practice is increasingly challenged, as data sources increasingly update more frequently (for example, the Household and Population Census moved from a decennial survey to the annual rolling American Community Survey) and even real time (such as traffic counts data). In addition, the quality issues coming with the new data sources also make the agencies scramble to cope.

In recent decades, there have been considerable advances in techniques in computer science and statistics, such as Bayesian statistics, data mining and machine-learning techniques, which have been applied to address such data problems in a wide range of domains. These domains are as varied as cleaning of web data, detecting fraud in credit data, reconciling medical records, mining the vast streams of email and web content for targeted advertising, and many others (See, for example, Hu et al., 2012). To date, most attempts to tackle the problem in the modeling communities (Abraham et al., 2009, 2005; for example, Waddell et al., 2005) are tied to a specific model system and a chosen study area. Few systematic efforts have applied these technological advances to produce re-usable tools in the problem domain of land use and transportation data.

The data integration project aims to address the challenging data problems by leveraging interdisciplinary techniques to develop reusable methods and tools. Specifically, we focus on making the data preparation process for modeling more systematic and reproducible with a harmonized data scheme and re-usable tools. In light of the challenges faced by the modeling communities, we envision data development as a continuous process instead of an ad-hoc, one-off one. Instead of solely focusing on constructing base-year data sets, we pay equal attention to preserve historical data as well as current and future data as they become available. Bringing these data in equal footing not only enables new possibilities for data analysis and modeling, but also creates new opportunity for data cleaning and imputation. To manage the challenges of a massive amount of data that is constantly changing, we propose a data-quality monitoring framework with indicators. The ultimate goal is to enable the best use of available data to inform

policy making and facilitate development of integrated land use and travel demand models at various geographical scales.

This chapter is organized as follows: The next section describes the background of challenging data issues and the current practice of data development for land use and transportation modeling. The Methodology section then proposes an alternative approach to the data development process.

1.1 BACKGROUND

Almost every integrated land use and transportation model system, in particular the land use model component of such systems, has its own information requirements, which are usually extracted and processed from diverse data sources. Table 1.1 shows an incomplete list of the main data sets, their sources and updated frequencies.

Table 1.1 Common Data Sets

Data Set	Main Attributes	Data Sources	Update Frequency
Parcels	Lot size, property value, units, building sq ft, year built, land use type, geometries	County Assessors	Annually/varies
Buildings	Units, building sq ft, stories, footprint	County Assessors, LIDAR	Varies
Census Geographies	Block group, Census tract, PUMA	Census Bureau	Every decade
Households and Persons	Income and housing characteristics (age, relation, sex, race/ethnicity, education)	Census Bureau	Annually
Businesses	Employment, NAICS code, location/address	Census Bureau, BLS, BEA, private sources	Quarterly
Political Boundaries	Geometries	Various jurisdictions	Infrequently
Land Use Plans (and other land use controls and environmental protection areas)	Zoning code (restrictions, fees and subsidies, geometries)	Municipal, county, regional, and state agencies	Annually/varies
Real Estate Transactions	Property identification/address, sale/rental price, sale/rental date	County Assessors, private sources	Varies
Traffic Counts	Screen line location, time and duration, counts by mode	DOT, MPO, municipality	Varies

Travel and Activity Survey	Household and person attributes, travel diary, activities, vehicle fleet	DOT and MPO	Annually for NHTS; Varies for local data

Table 1.1 Common Data Sets (Continued)

Transportation Networks	Classification, lanes, speed, direction, geometries	DOT, MPO	Varies
Traffic Analysis Zones	Geometries, other attributes aggregated to TAZ	DOT, MPO	Infrequently

Because of the diverse sources of data and extensive attributes used in integrated land use and transportation models, each region developing such models has its own unique data issues, such as missing data sets and values, questionable data quality, etc. For regions facing such challenges, the seemingly least-cost path of developing data for a model application is to process, impute and transform the available data directly into inputs required by a specific chosen modeling platform. This process is usually ad-hoc, as it depends on the data needs of the platform, the data available at the time for the region, and particular problems presented in the data sets. Because of asynchronous update cycles, different data issues occurring at different data snapshots, and data needs for the model system keeping evolving, the data development process is also usually a one-off – it can hardly be re-used for historical data or new crops of data.

Figure 1.1 illustrates the ad-hoc approach to data development, in which an individual data set is directly processed for specific models. It may be the fastest and least costly to get the selected models up and running in the short term, but it has a few shortcomings, including:

- The process and resulting data are tied to the chosen modeling platform and don't lend themselves well to other model platforms or extra analysis outside the chosen platform.
- Information from different data sources may contradict each other, which causes an issue in model outcomes; there are lost opportunities to utilize the relationship between data sets to create consistent inputs.
- Data processing efforts are not re-usable, as most changes; however small they may be, the raw data or the model inputs would require changing the data development process.

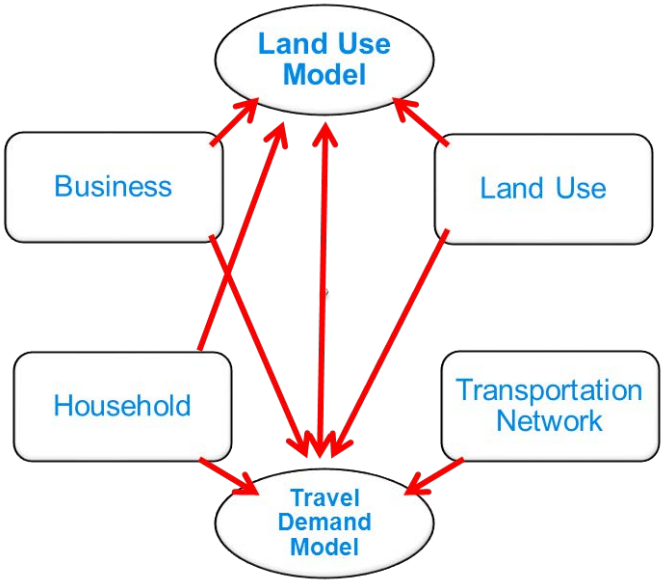


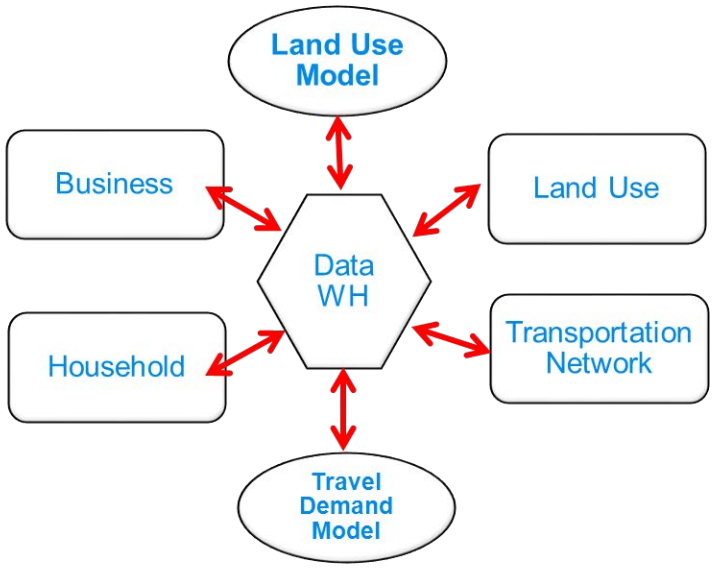
Figure 1.1 Ad-hoc Approach to Data Development

In the practice of model development, the ad-hoc approach is the most frequently used approach. The data development for the UrbanSim application for the Puget Sound region (Waddell et al., 2005) exemplifies such an approach, in which data from original sources are processed into data tables for UrbanSim. Abraham et al (2005) tested three rule-based approaches to synthesize real-estate inventory data for modeling applications according to data available in three different regions. Each approach is tied to the specific model requirements and geographical resolution (TAZ, parcel, or gridcell) specified for one of the regions. While these documents and algorithms are useful hints to other regions adopting these models, each region would have to re-create its own process and tools to prepare data, not to mention if it wants to switch modeling platforms or compare two different platforms.

1.2 METHODOLOGY

Land use and transportation modeling is very data intensive, requiring data reflecting various aspects of land use and transportation systems. The data requirement understandably varies by model system, but even for the same model system the data required could evolve over time as the model system evolves. To shift data preparation for land use and transportation models from an ever ad-hoc effort to a re-usable one, we propose these principles as the methodology for data development:

- Reproducible: All changes to the data are done through documented scripts, ideally through literate programming (Knuth, 1984) and with necessary version control and provenance. The end data should be able to be reproduced from raw data by re-running these scripts if necessary.
- Re-usable: Open sourced and adaptable scripts are preferred to non-repeatable data editing (e.g., with Excel); Save data in non-proprietary format (e.g., Excel format).
- Consistent and harmonized: A harmonized data warehouse (WH) as an intermediary between data processing and modeling inputs so that the processed and cleaned data is not tied to a specific model system, but can be transformed to feed various model systems as needed (Figure 1.2). Such a model-system, agnostic data warehouse would isolate the region-specific data issues and system-specific data needs and requirements from processes and algorithms that are generic to all regions and model systems.
- Constant quality monitoring: Quality of the data should be constantly monitored with data-quality indicators as the raw data is being loaded and processed. This provides some assurance of the quality of the data used in the model.



Data WH – Data Warehouse
Figure 2.2 Proposed Approach of Data Development

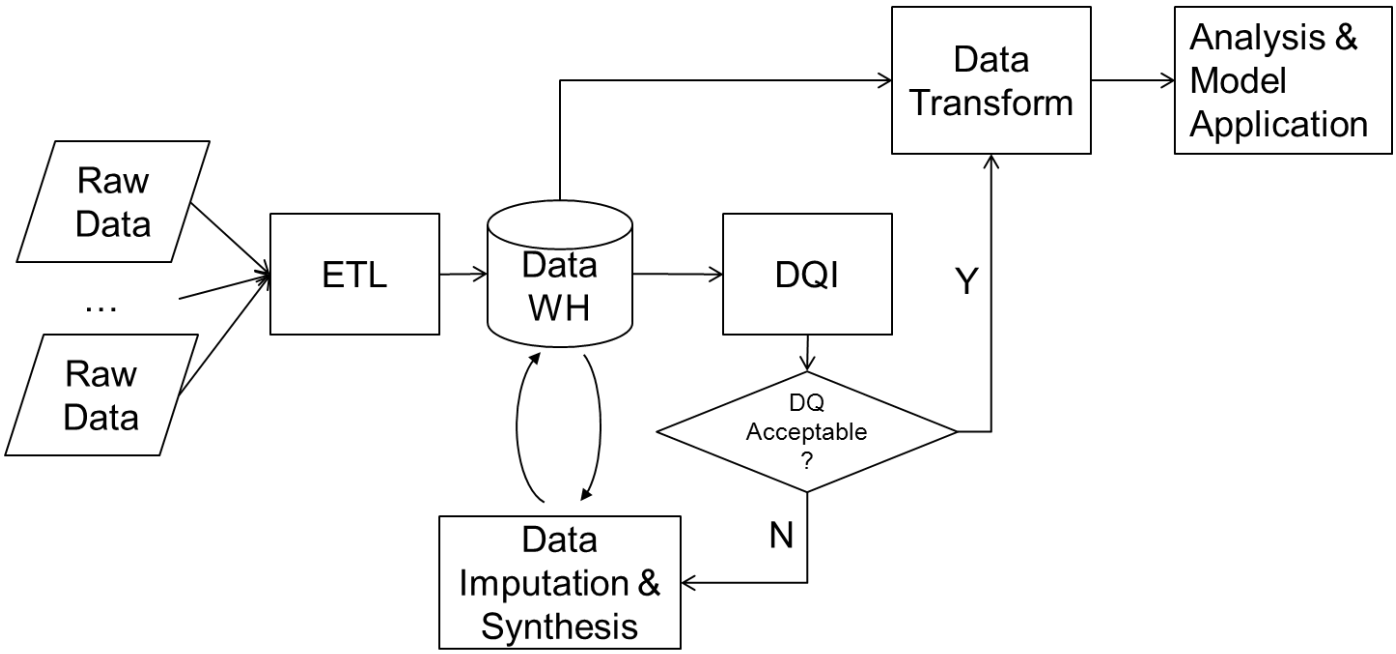
Following these principles, we propose a data development work flow (Figure 1.3). Each step in the work flow is discussed in turn.

1.2.1 Extract, Transform and Load (ETL)

Originated in the data management and business information field, the Extract, Transform and Load (ETL), or Extract, Load and Transform (ELT), pattern is able to isolate variations in

aspects of data sources (for example, location, format and table structure), and required processing from later steps of data management. Such patterns implemented in high-level scripting language such as R and Python provides a versatile infrastructure to read and process data from various sources quickly without the need to alter other steps, even when data sources change.

The Extract step extracts data from outside sources. As shown in Table 1.1, these sources can be very diverse, and this step can thus vary from downloading datasets from websites and unzipping compressed files to scraping information from unstructured sources. The Transform step then transforms the extracted data so that it conforms to the defined data schemes. The final Load step loads the transformed data into the end target, a data warehouse in our case.



Data WH – Data Warehouse; DQ – Data Quality; DQI – Data Quality Indicators
Figure 1.3 Data Development Work Flow

The difference between ETL and ELT is largely a matter of preference, in particular, of whether it is preferable to have the raw information extracted in the E step stored in the target data storage (Data Warehouse in our case). If it is, then an ELT pattern is preferred; otherwise an ETL one is. Both patterns achieve the goal of isolating diverse data sources from later processing equally well.

A special type of data that is commonly used in transportation and land use modeling is geometry (GIS) data. The geometry data is most frequently processed with ArcGIS, and the process is hardly reproducible or re-usable. A more reproducible and re-usable alternative we propose is PostGIS, a spatial Database Management System (DBMS) based on open sourced Postgresql (Obe and Hsu, 2011). Not only can geometry data like shape files be stored in PostGIS along with attribute data; it also provides a large set of spatial operations that satisfy most data processing needs with reproducible and reusable scripts.

1.2.2 Consistent and Harmonized Data Schemes

Instead of molding data into the requirement of a specific model system, we propose to work towards a harmonized data warehouse with pre-defined table schemes, including commonly used data tables. The input tables for a specific model can then be generated from the harmonized data warehouse with translator scripts. Although this seems to be a detour from directly preparing input tables from raw data for the model system, we believe this approach avoids the pitfall of the current direct but ad-hoc practice.

Two successful data projects have adopted the harmonized data schemes to facilitate data use from diverse sources: the Metropolitan Travel Survey Archive project (Levinson, 2004; Levinson and Deshpande, 2006, 2003) and the Integrated Public Use Micro-data Series (IPUMS) project (Ruggles et al., 2010). The Metropolitan Travel Survey Archive project archives historical travel surveys done at metropolitan areas across the U.S., while the IPUMS project collects available historical micro data from Census counts around the world and harmonizes them with uniform code books to facilitate use of the data for research. As far as we know, there is no research so far aiming to develop tools and procedures to create consistent and harmonized land use and socio-economic data.

One of the advantages of going through harmonized data schemes is that it makes it easy to monitor data quality and offer opportunities to improve consistency between data sets, arguably one of hardest issues in developing data for integrated modeling.

1.2.3 Data-quality Monitoring

Once we have harmonized data schemes, the data-quality monitoring can then be automated following continuous software testing procedures so that the data quality from various sources is constantly monitored. We envision constant data-quality monitoring with data-quality indicators, which provides information of the quality of the harmonized data warehouse and informs the data imputation and synthesis process. In addition to data-quality dimensions described in the Data-quality Indicators chapter, data-quality indicators can also be adapted to identify outliers. Singleton (2013) recently reviewed a series of indicators for outlier identification and applied a few of them to clean walk trips in travel survey data. Such indicators can be implemented in the monitoring framework, both for quality assurance and for assisting data cleaning and imputation.

Besides quantitative assessment of data quality with indicators, interactive visualization and diagnostic tools such as the OpenRefine software and linked graphs can assist visual inspection of individual data observation for potential outlier detection, especially for data sets that are linked to each other by associations.

1.2.4 Data Imputation and Synthesis

More often than not, data loaded to the warehouse may need to go through an iterative process of cleaning, imputation and synthesis, although the extent of the imputation process depends on whether the data-quality monitoring step indicates an acceptable quality. Data cleaning removes invalid and inaccurate values identified by data-quality indicators or via interactive data

inspection, while data imputation fills in incomplete observations and missing values. In the case where there are more extensive gaps in the data, data synthesis may be needed to fill these gaps.

Data cleaning, imputation and synthesis are commonly done through undocumented manual programs and processes, such as Excel and text editing software, which creates issues including un-reproducible results and un-reusable process. We propose to do these processes with documented scripts, loosely following the ETL pattern. Documented scripts working with harmonized data schemes make it possible to re-use the data imputation and synthesis process across regions, even when each of them face different data issues.

There has been an expanding list of methods for imputing and synthesizing land use and socio-economic data that have been tried and proposed by modelers to tackle data issues. From the published literature, there are generally four types of imputation and synthesis methods: rule-based approach, Iterative Proportional Fitting (IPF)-based approach, model-based approach, and machine-learning approach.

The rule-based approach generally imputes data following a deterministic or heuristic rule, such as balancing demand and supply. Abraham et al (2005) provide examples of rule-based data imputation, in which the authors assign floor space to geography to meet pre-determined demand. IPF is the underlying algorithm for population synthesizers that are commonly used in transportation demand modeling (See Müller and Axhausen, 2011 for a recent review of popular population synthesizers). Abraham et al. (2009) applied IPF for synthesizing real-estate supply (built form). In a model-based approach, the data to be imputed or synthesized are outcomes of one or more models in the model platform. Since the data itself is from models, there is likely less inconsistency issues in data imputed with the model-based approach. Weidner et al (2007) uses PECAS to calibrate the base-year demand/supply of floor space; Wang, Waddell and Outwater (2011) applied the workplace choice model to synthesize data that is consistent with workers' commute patterns. The machine-learning approach borrows algorithms from computer sciences. Waddell (2009) identifies two algorithms that are particularly promising for data cleaning and imputation: K-nearest neighbors and Support Vector Machine. A promising new approach proposed by Farooq et al (2013) using Markov-Chain Monte Carlo Bayesian Simulation for population synthesis; it has the advantage of being able to capture multi-dimensional dependence among attributes.

While there is now an expanding list of methodologies for data imputation, it is beyond the scope of this project to provide an exhaustive review of such methodologies. We instead focus on one promising approach – Multiple Imputation with Chained Equations (MICE) – for data imputation.

1.2.5 Data Transformation and Application

Once the data in the data warehouse is of acceptable quality, the final step of the data development work flow is to transform it to the format required by the chosen model. The step may require recoding of variable values, aggregating or disaggregating attributes from joined tables, etc. Transform and Export scripts similar to those for the ETL step can be developed and applied for this step.

1.3 CHOICE OF SOFTWARE

In our choice of software for implementing the data development process, we prefer software that is free and open source, which is more readily available to the public and more transparent. We also prefer command line tools and programming scripts to software with Graphic User Interface (GUI), as processes developed over GUIs are less re-producible or re-usable although they may have an easier learning curve.

We eventually selected R as our main scripting language for implementing most of the processes in the proposed work flow. R is a free and open-source statistical programming language, which is widely used in statistical and computational research and practice. It has also gained wide use in the land use and transportation modeling community. There are now both land use and transportation models implemented in R by researchers and practitioners, such as MetroScope (Metro, 2009); LUDSR; GreenSTEP (Gregor, 2012); TravelR; and JEMnR. Another advantage of R is the thousands of packages made available by researchers that extend R to do various sophisticated analysis.

We use PostgreSQL and PostGIS, its extension for geographical data support, as the backend for the data warehouse. Postgresql is a free and open source Object-Relational Database Management System (ORDBMS) that is one of the most popular and robust systems. PostgreSQL supports standard Structural Query Language (SQL) that provides functions for queries and data management. In addition, R can directly interact with PostgreSQL via DBI APIs. PostGIS adds supports for geographic objects to the PostgreSQL database. Similar to its commercial counterparts, such as Oracle Spatial and Graph and ArcSDE Geodatabase, PostGIS enables not only storing and retrieving geographic data into and from PostgreSQL databases, but also supports of basic spatial queries and operations.

The report is organized as follows: Chapter 2.0 describes the Extract, Transform, and Load (ETL) process. Chapter 3.0 describes data-quality assurance and diagnoses, including data-quality indicators and interactive inspection tools. Chapter 4.0 demonstrates the MICE tool for data imputation. Chapter 5.0 discusses a discrepancy analysis approach for revealing activity patterns in travel diaries data. Chapter 6.0 concludes the report and discusses topics for future research. Several appendices provide details of scripts and tools used in the report.

2.0 EXTRACT, TRANSFORM AND LOAD (ETL)

Extract, Transform and Load (ETL) is a process in data warehousing that:

- Extracts data from varied data sources;
- Transforms the data for storing it in a proper structure for querying and analysis; and
- Loads the data into a data warehouse (DW).

In the business intelligence domain, the ETL process commonly integrates data coming from multiple applications or systems that may be managed and operated by different employees and on different hardware (than the data warehouse). Similarly, the data for land use and transportation modeling, in particular the land use data, comes from various sources that are arguably even more diverse than that in typical business intelligence (BI) application.

Even though it requires working with diverse data sources, in general the ETL process is easy to standardize. There is software specifically implementing the ETL function. There is at least one existing R package, ETLUtils, which implements the ETL operations in R. For our purpose, it is possible to use the functions in standard R and a few packages for ETL operations. Appendix A-1 exemplifies the ETL operation in R. In application, we will mix and match snippets of R code according to the situation and requirements, and combine them to create an ETL process for a particular data set.

2.1 EXTRACT

The first step of an ETL process involves extracting the data from sources. The land use and transportation data usually consolidate data from multiple sources, and each source may use a different data organization and/or format. Common data source formats include:

- relational databases
- dbf files
- csv, tab, or other flat files
- Excel spreadsheet
- fixed format
- SAS, SPSS or other statistical data archive

In general, the goal of the extraction phase is to convert the source data into a single format appropriate for transformation processing. In our case, we will extract data from various sources into R data.frame, as almost all data sets for land use and transportation modeling can fit into the memory of a high-end desktop computer. If the data cannot fit into the memory, it is possible to use R package ff that handles files too large to read into the computer memory all at once.

2.2 TRANSFORM

The data transformation step applies a series of rules or functions to the data extracted from the source to the desired structure before loading into a data warehouse. In some cases, the data may not require any transformation at all. In other cases, one or more of the following transformations may be required:

- Selecting a subset of columns to load: For example, ignore object id column in a dbf file;
- Selecting only certain rows: For example, ignore all those records where a certain column is not present (= null);
- Recoding attribute values: For example, recode 1 for male and 2 for female to M for male and F for female;
- Calculating a derived value: For example, population density = population / area;

- Joining data from multiple sources (e.g., lookup, merge);
- Removing duplicated data;
- Aggregation: For example, summarizing multiple rows of data;
- Disaggregation of repeating columns into a separate look-up table: For example, moving a series of addresses in one record into single addresses in a set of records in a linked address table;
- Transposing or pivoting: For example, turning multiple columns into multiple rows or vice versa; and
- Splitting a column into multiple columns: For example, converting a comma-separated list, specified as a string in one column, into individual values in different columns.

An important function of data transformation is cleansing and validating data according to a set of rules before passing it on to the target. For example, it needs to verify that columns from two different sources are using the same data dictionary. These cases must be handled correctly or eventually lead to a number of data-quality issues.

2.3 LOAD

The load phase loads the data into a data warehouse. Our ideal is that new data in an historical form are appended to the data warehouses at regular intervals - for example, annually.

As the load phase interacts with a database, the constraints defined in the database schema — discussed in the next chapter — apply (e.g., uniqueness, referential integrity, mandatory fields), which also contribute to the overall data-quality performance of the ETL process.

3.0 DATA-QUALITY ASSURANCE

Despite being a prominent issue in many disciplines and with recent efforts for an integrated view of data quality (Devillers et al., 2007; Yang et al., 2013), data quality is difficult to define, as it is domain- and problem-specific (Pipino et al., 2002). Data quality is usually described with hyper-dimensions and the data-quality literature has generally identified these dimensions (Hsu, 2014):

- Completeness whether each observation in a data is complete row-wise, or whether any observation has missing values for certain columns. The completeness can be easily measured, for example, as the percentage of observations containing non-null values.
- Accuracy whether the value recorded in the data reflects the true information of the object represented by the observation. For example, whether a household reportedly

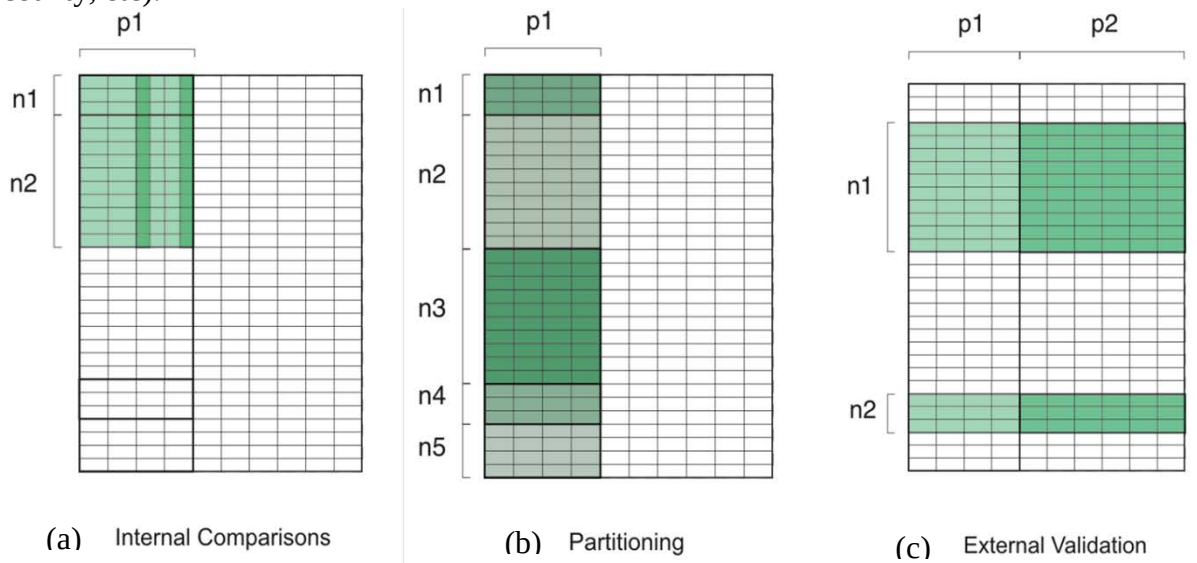
having 10 vehicles does, in fact, have that many. Usually external data is required to cross check the information before an accuracy dimension can be assessed.

- Validity whether the data value is reasonable in terms of magnitude and units. For example, a building cannot have a character as a number of floors or 99 floors if the data is not for cities with skyscrapers. Validity is a weaker and easily measured form of accuracy.
- Consistency whether recorded information in one data set is consistent with relevant information from another data set. For example, the x, y coordinates of a household's residence location recorded by a GPS may be inconsistent with a misspelled street address.

The data-quality indicators or metrics are quantitative measurement of data quality on one or more of these dimensions. They provide a gauge of the quality of individual data sets and consistency between different data sets.

Hsu (2014) suggests a series of operations for computing data-quality indicators. We adopted three of them for our data-quality indicators, namely internal comparisons, partitioning and external comparisons (Figure 4.1). Internal comparisons examine the internal consistency between columns of a data set; partitioning contrasts different subgroups (n1, n2, …) in a data set, while external comparisons compare information from one data set with exogenous information (p2).

We adopted his methodology and demonstrate some of the indicators in the next section. The measurement of completeness and validity is straightforward, while accuracy and consistency have multiple potential metrics. We put more focus on how accuracy and consistency changes over time (annually) and over geographical scale (e.g., parcel, block group, tract, municipality, county, etc).



p1 are columns of a dataset, while p2 are columns in an external dataset, n1-n5 are subgroups of dataset. Source: (Hsu, 2014)

Figure 3.1. Three Strategies for Deriving DQ indicators

3.1 DQIS VS. INTERACTIVE EXPLORATION

The DQI usage should also follow the re-usable principle. The code snippet below demonstrates definition of DQIs functions in R that are re-usable across attributes, data sets or study areas. The code defines functions for Standardized Root Mean Square Error (SRMSE) and R^2 that are useful for internal comparison and external validation. Appendix A-2 showcases an application of the DQI functions on housing units between data from the American Community Survey and Metro's Regional Land Information System (RLIS)

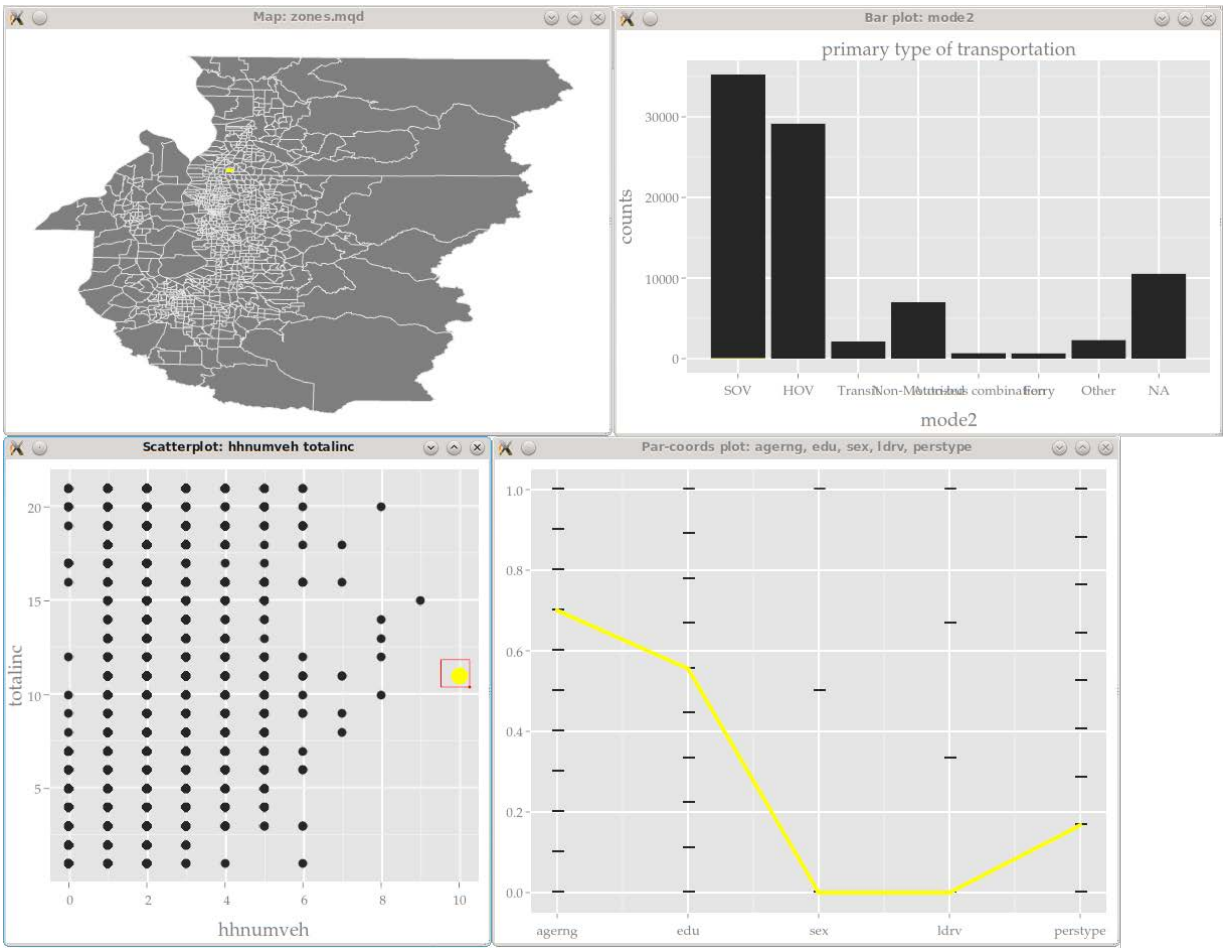
```
##R code defining DQI functions
rmse <- function(x=x, y=y) sqrt(mean((x-y)^2))
r2 <- function(x=x, y=y, ...) {summary(lm(y~x,...))$r.squared}
intercept <- function(x=x, y=y, ...)
                     {coef(lm(y~x,...))["(Intercept)"]}
slope <- function(x=x, y=y, ...) {coef(lm(y~x,...))["x"]}

##DQI example
require(plyr)
require(data.table)
units <- read.table("housing_units.csv", sep=",", header=T)
units.dt <- data.table(units, key = "Year")

dqi <- ddply(units.dt, .(Year), summarise,
      mfh.rmse=rmse(x=RLIS_MFH, y=ACS_MFH),
      mfh.r2=r2(x=RLIS_MFH, y=ACS_MFH),
      mfh.intercept=intercept(x=RLIS_MFH, y=ACS_MFH),
      mfh.slope=slope(x=RLIS_MFH, y=ACS_MFH)
      )
print(dqi)
```

In addition to quantitative DQIs intended for automatically monitoring data quality, interactive tools are necessary to assist the assessment of DQ and potentially fix data issues. Two examples of such tools are OpenRefine and CRANVAS. OpenRefine (http://openrefine.org/) is an open source tool for data exploration, cleaning and transformation. It is particularly well suited for identifying and processing invalid values. OpenRefine demos and tutorials are readily available online.

CRANVAS provides a linked graph function and is helpful for data exploring and identifying inconsistency between data sets. Since inconsistency between data sets is a major difficulty in data development for land use and transportation modeling, it is beneficial to be able to dig into data points that may be inconsistent across multiple linked data sets. A common example of linked data sets is the household travel and activity survey data sets, which usually includes individual data sets for households, persons, trips and activities. Figure 4.2 shows an example of using CRANVAS, a visualization tool developed in R, for interactively exploring the travel survey data sets for the Puget Sound region.

CRANVAS can simultaneously visualize attributes from multiple data sets, each in any type of graph supported by R. The panels in Figure 4.2 show household residence location in a map, distribution of household member's primary mode of transportation in a bar chart, number of vehicles versus income category in a scatter plot, and a line graph for various personal attributes. What makes CRANVAS useful is that these charts are linked by the data sets' foreign keys. In

the example shown in Figure 4.2, a household with an exceptionally high number of vehicles (hhnumveh=10) is selected in the lower left panel, and the residence location is then automatically highlighted in the top left panel while personal attributes for household members are highlighted in the lower right panel.



Clock-wise from top left: household residence TAZ; Primary mode of transportation for household members; number of vehicles and total household income (category); Line graph of person attributes of age range, education, sex, is licensed driver, type of person.

Figure 3.2. Linked Graph for Travel Survey Data Sets

3.2 DQI CASE STUDY

In this paper, we focus on a case study in the Portland metropolitan area. The state of Oregon, particularly the Portland area, is a pioneer in developing and applying integrated modeling. Metro has been systematically collecting, documenting and archiving land use and transportation data via its Regional Land Inventory System (RLIS). This case study is thus an easy case study to demonstrate the data-quality indicator framework.

Since measuring completeness and validity dimensions of data quality are easy, we focus the case study on the far more complex accuracy and consistency measures. Below we show only a few examples of data-quality indicators applied in the context of land use and socio-economic data sets. They can be applied more extensively to many more attributes and relationships in the data.

3.2.1 Internal Comparisons

The internal comparisons example looks into the number of households from the Census versus those in the synthesized population by geography. Although they do not exactly conform to Figure 4.1(a), the synthesized population originated from the former and should have excellent internal consistency between the two, unless there is an error in the synthesis process. Figure 4.3 shows that the synthesized data fits the original census data perfectly, indicating the population synthesizer run was successful by this measure.

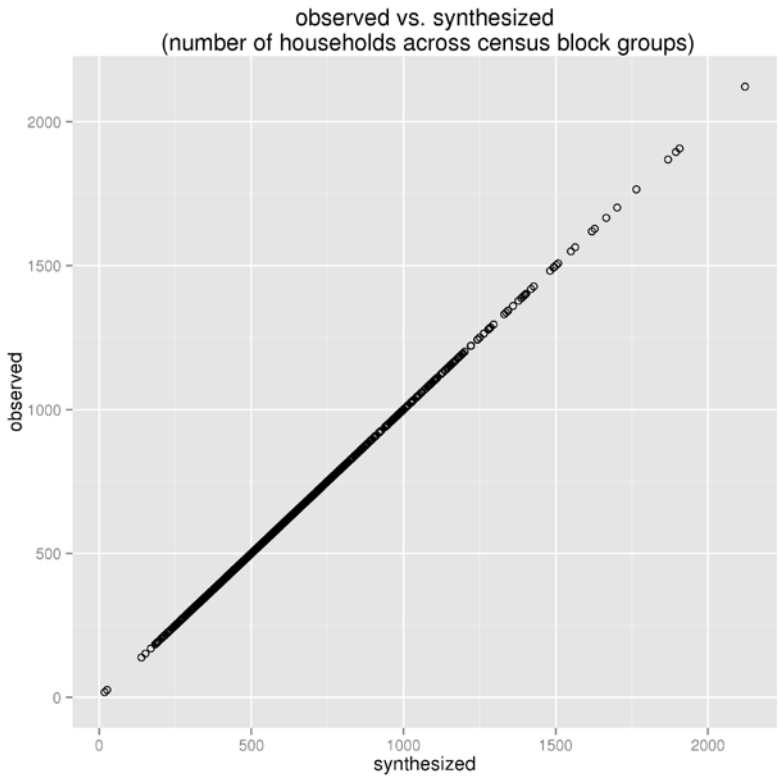


Figure 3.3 Household Counts (Census) vs Household Counts in Synthesized Population by Census Tract

3.2.2 Partitioning

The partitioning example (Figure 4.4) shows housing units partitioned by unit type and year in a box plot. It is clear that, although the mean and mass of the housing units are stable for both unit types over years, there are more extreme values in 2011-2012 and more outlier tracts for multifamily housing than single-family housing.

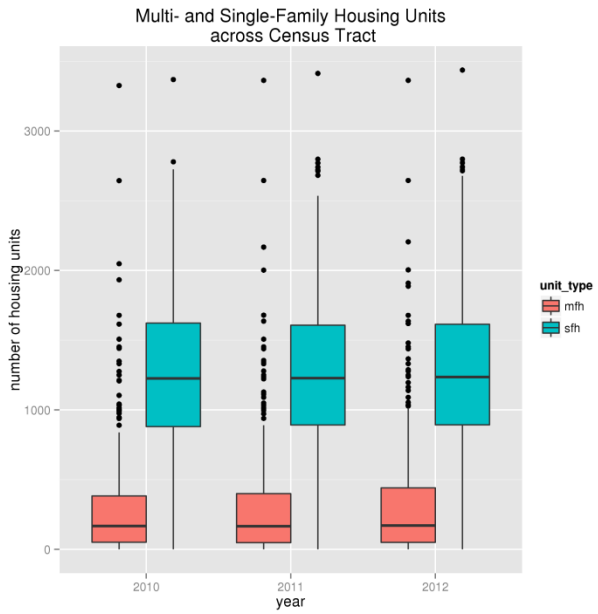
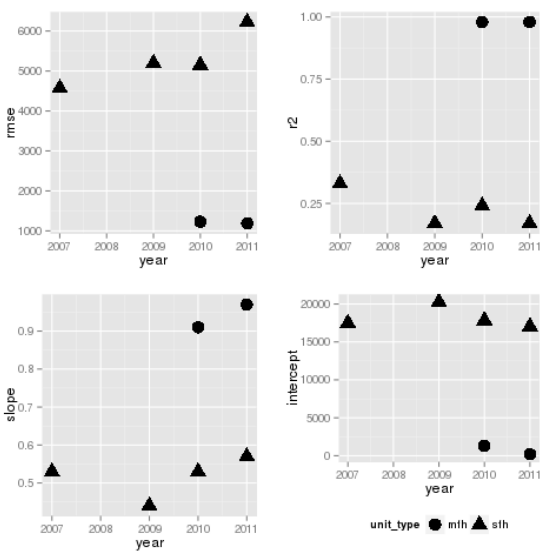


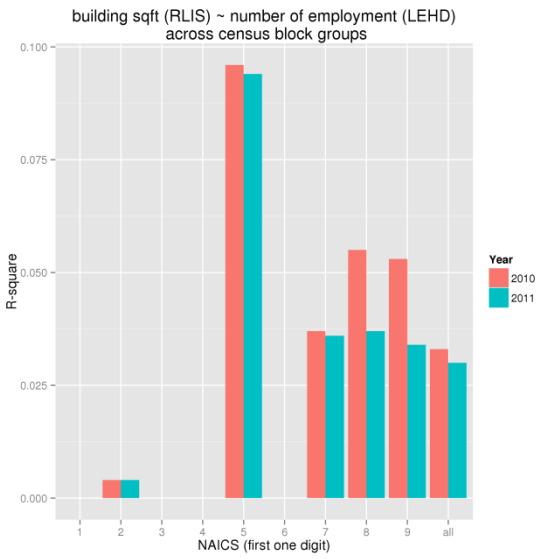
Figure 3.4. Box Plot of Housing Units by Unit Type and Year

3.2.3 External Comparisons

The inconsistency between real estate demand and supply are among some of the major data issues in land use and socio-economic data sets. Figure 4.5 presents indicators measuring the consistency via external comparisons of real estate inventory (from tax lots) and demand (from the Census and LEHD).



(a)



(b)

(a) RMSE and R^2 between housing units (tax lot) and households (census) by housing type (Single Family Housing and Multi-Family Housing) and year; (b) R^2 between building sqft (tax lot) and number of employment (LEHD) by NAICS sector and year.

Figure 3.5. External Comparisons

Figure 4.5(a) plots the consistency at Public Use of Microdata Area (PUMA) level between housing units (from parcels) and households (from Census) for two housing types for the years the housing units information are available (2007, 2009-2011), measured by RMSE and R^2. The higher the RMSE number goes, the less consistent the two data sets are; while the higher the R^2, the higher the consistency is. Surprisingly, the multifamily housing (MFH) and its residents have better consistency than that between the single-family housing (SFH) and its residents. Further exploration of the data (not shown here) shows that there are a few PUMAs that have large gaps between SFH units and household reported in ACS; future investigation is needed to identify which source produces inaccurate numbers. Also note that the indicator values vary by year, which indicates there may be an opportunity to utilize historical data to examine and impute data for later years. As simple metrics, these indicators provide an overall view of consistency between two key data sources. Figure 4.5(b) presents similar consistency metrics between non-residential real estate demand and supply. Unsurprisingly, the consistency measure for the non-residential real estate market is much lower than that for the residential real estate market. The variation in space occupied by each employee across sector and location is one factor contributing to the larger inconsistency; the lower accuracy in the LEHD data series caused by jittering could be the other.

4.0 DATA IMPUTATION

4.1 INTRODUCTION

As analysis and models for land use and transportation planning gradually move to the micro level because of policy requirements and theoretical and technical advantages (Waddell, 2009), missing data becomes an increasingly common problem in micro-level land use and transportation modeling and policy analysis. According to Waddell et al's (2005) estimate, up to 70% of the efforts in land use and transportation modeling are spent on data processing, and handling missing data is a major piece of these efforts.

Missing values in land use and transportation data sets, particularly those in land use data sets, are generally handled on an ad-hoc basis with either manual inspection and cleaning, or rule-based or heuristic methods. While these methods can solve one type of missing data problem at a time in some user-specified sequence, they are hard to be adapted for different applications and there is no systematic way to assess their quality (Waddell, 2009).

More sophisticated statistical modeling and machine-learning techniques have been developed in statistics and computer sciences, and tested and applied to tackle missing data problems in many fields such as industrial engineering (Lakshminarayan et al., 1999) and forestry (Eskelson et al., 2009) etc. Waddell (2009) suggests that k-nearest neighbors and support vector machines may be two promising techniques for imputing missing land use data, particularly parcel-level data. However, to our knowledge, there is no systematic research assessing the applicability and quality of these techniques in imputing missing land use data.

Furthermore, most applications of data imputation in land use data generate a single imputation of the missing value – that is, substitute a missing value with a single best guess. However, single imputation masks the uncertainty in the missing values. In other words, the imputed values in the data are treated as if they are real observed values, without appropriately addressing the uncertainty associated with the substitutions. To this end, Rubin (2004) proposes a multiple imputation framework. Instead of producing a single best guess, multiple imputation uses a Monte Carlo simulation and replaces each missing entry with a set of *m* plausible values in which m is typically small, say 3 to 10. The advantage of multiple imputation is that each of the *m* imputed data sets can still be analyzed with standard complete-data methods, but the analysis results can be combined by simple arithmetic calculations to produce mean estimates, standard errors, and p-value for quantities of interest that incorporate uncertainty due to imputation of missing data.

A popular approach to implementing multiple imputation, especially for multivariate missing data, is multiple imputation by chained equations (MICE), also known as fully conditional specification or sequential regression multiple imputation. It has been widely applied in psychology (Azur et al., 2011) and medicine and epidemiology (Burgette and Reiter, 2010; White et al., 2011), etc. One of the reasons for the popularity of MICE is its flexibility. MICE can handle missing continuous and discrete variables because each variable with missing data is imputed based on its own imputation model (Azur et al., 2011; Burgette and Reiter, 2010; White et al., 2011). Recent studies attempt to implement recursive partitioning within MICE in order to automatically preserve interaction effects in the data ((Burgette and Reiter, 2010; Doove et al., 2014; Shah et al., 2014)).

In this study, we apply the non-parametric MICE approach to impute missing values in parcel data. Recently, more and more land use and transportation modeling work and analysis moves to use parcel data (Waddell, 2009; Waddell et al., 2005). However, parcel data are prone to problems of incomplete or missing values. In addition, in land use and transportation modeling, base year data are usually used to project future land use and travel patterns. So, even though the number of missing values may be small, it is still necessary to impute them. We aim to test different methods in imputing missing parcel data and assess the quality of the imputation results, taking the uncertainty into consideration.

The remainder of this paper is organized as follows. The next section overviews the background of missing data patterns and mechanisms, multiple imputation, MICE, and the implementation of recursive partitioning in MICE. In the Parcel Data Imputation section, we discuss multiple imputations for parcel data, along with missing data description, MICE set-up and visual

diagnosis. In the section that follows, we discuss the validation of non-parametric MICE. Finally, we conclude the paper with some suggestions for future research.

4.2 BACKGROUND

4.2.1 Patterns and Mechanisms of Missing Data

For practical reasons, it is useful to distinguish three different patterns of missing data: univariate, monotone and arbitrary patterns (Buuren, 2007; Schafer and Graham, 2002). First, a missing data pattern is univariate if only one variable has missing values and the remaining variables are all completely observed. For the univariate missing data, a variety of parametric or non-parametric methods can be used for conducting multiple imputations. Second, in the monotone pattern, more than one variable are missing and the missing variables can be ordered such that if a variable is missing for a subject, then other variables are also missing for that subject. Such a monotone pattern is often observed in a longitudinal data set as one leaves the longitudinal study in the middle of entire waves of data collection. In this case, it is possible to impute the multivariate missing data by a series of univariate methods for multiple imputations. Lastly, an arbitrary pattern is observed from multivariate missing variables, in which their missing values can occur in any set of variables for any subjects. Then, we need a truly multivariate method for multiple imputations. The arbitrary pattern is the focus of this research.

On the other hand, it is important to understand different mechanisms through which data are missing, because different missing mechanisms may cause different risks of bias when missing data are excluded in analysis. The missing mechanisms are commonly classified as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). First, MCAR assumes that the probability of data being missing depends on neither observed nor unobserved variables. In this case, the set of subjects that are completely observed is also a random sample from the source population. Thus, a complete case analysis in which all missing data are excluded gives unbiased results, even if such a simple method is less efficient because the entire data set is not used. Second, in the MAR mechanism, the probability that data are missing depends on observed variables, but not on unobserved ones. In this case, a complete case analysis is no longer based on a random sample, and selection biases are likely to occur. Such a bias can be overcome through data imputation, in which a missing value for a subject is replaced by a predicted value from the subject's other known characteristics. Lastly, a mechanism for missing data is MNAR if the probability of missing data depends on both observed and unobserved variables. In this case, there is no universal method. We may only address the MNAR biases by conducting sensitivity analyses in which the effects of different mechanisms are compared (Donders et al., 2006). As in many other studies, the MAR mechanism is assumed in this study to apply a data imputation approach to the missing data problems.

4.2.2 Multiple Imputation

Multiple Imputation is a statistical approach to handling missing data, which was first proposed in the 1980s by Rubin (2004). It aims to overcome the limitation of single imputation by allowing for the uncertainty introduced by missing data. Multiple imputation function is

increasingly available in common statistical software, such as R, SAS, Stata, and SPSS. In general, multiple imputations consist of three steps. The first step is to construct an imputation model, which is sometimes referred to as an imputer or an imputation engine. For a single missing variable *z*, the imputer regresses *z* on a set of non-missing variables among individuals with observed *z* values. Now each missing value of z is replaced *m* times by a plausible value that is a random draw from the predictive distribution of the imputation model.

The second step is usually simple. Each of the multiple imputed data sets is analyzed separately but identically by a complete-data method (e.g., a linear regression) to estimate quantities of interest (e.g., the regression coefficients). Since the multiple data sets are identical for the complete data but not for the imputed data, this step generates the *m* different estimates for each quantity.

The third step is to combine the *m* estimates using Rubin's rules (Rubin, 2004). Suppose that $\hat{\theta}_j$ is an estimate of a scalar quantity of interest obtained from imputed data set *j* (*j* = 1, 2,…, *m*) and W_j is the variance of $\hat{\theta}_j$. The overall estimate $\hat{\theta}$ is simply the average of the individual *m* estimates:

$$\hat{\theta} = \frac{1}{m}\sum_{j=1}^{m}\hat{\theta}_j$$

The overall variance of the overall estimate $\hat{\theta}$ is formulated with two components: the within-imputation variance (W) and the between-imputation variance (B):

$$var(\hat{\theta}) = W + \left(1 + \frac{1}{m}\right)B$$

The within-imputation variance explains the variation of the estimate in one imputed data set, which is calculated by averaging the *m* variances:

$$W = \frac{1}{m}\sum_{j=1}^{m}W_j$$

The between-imputation variance measures how the estimates vary from imputation to imputation to reflect the uncertainty due to missing data, which can be given:

$$B = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{\theta}_j - \hat{\theta})^2$$

The greater the variation of the estimates across the imputations, the higher the uncertainty introduced by missing data. The overall standard error is the square root of $var(\hat{\theta})$. A significance test of the null hypothesis $\theta = 0$ is performed in a usual way. Since single imputation omits the between-imputation variance, the standard error is always too small (7).

4.2.3 Multiple Imputation by Chained Equations (MICE)

In large data sets with missing values such as the parcel data set, the missing values often occur in more than one variable and follow an arbitrary pattern. When the pattern of the multivariate missing data is arbitrary, two general multivariate approaches are available for multiple imputation: joint modeling (JM) and multiple imputation by chained equations (MICE). The JM approach assumes a multivariate normal distribution for all variables in the imputation model. Imputed values are obtained from the estimated joint distribution. JM was first proposed by Schafer (Schafer, 1997) and was widely used in the earlier applications of multiple imputation to multivariate missing data. However, the parametric method of JM may lack flexibility due to its normality assumption. In case there are both continuous and discrete variables, the assumption of multivariate normality is often violated. The MICE approach is more flexible in that instead of assuming multivariate normality, MICE models a series of conditional distributions, one for each missing variable, based on the other variables in the imputation model. The semi-parametric method of MICE is increasing popular and there are many statistical packages available for MICE (Buuren, 2007).

To describe the procedure of generating multiple imputed data sets in MICE, suppose a set of variables, $y_1, \ldots, y_j$, where some of them are missing. The following steps are involved (6):

- Step 1: Fill in every missing value by a random draw from the observed values, which will serve as a placeholder.
- Step 2: For the first missing variable, say y_1, return the placeholders to missing, and then construct an imputation model that regresses y_1 on the other variables, say $y_2, \ldots, y_j$, only among individuals with the observed y_1.
- Step 3: Replace every missing value in y_1 by a random draw from the posterior predictive distribution of the imputation model for y_1.
- Step 4: For the second missing variable, say y_2, return the placeholders to missing, and then construct an imputation model that regresses y_2 on the previously imputed variable(s) y_1 and the other variables $y_3, \ldots y_j$, only among individuals with the observed y_2.
- Step 5: Replace every missing value in y_2 by a random draw from the posterior predictive distribution of the imputation model for y_2.
- Step 6: For all other missing variables, repeat Steps 4 and 5.
- Step 7: To stabilize the results, repeat Steps 2 through 6 *l* times (e.g., 10 to 20), which produces one imputed data set.
- Step 8: Repeat Steps 1 through 7 *m* times (e.g., 5 to 10) to generate *m* imputed data sets.

The most significant advantage of using MICE is its ability to handle different types of variables because each missing variable is imputed from its own imputation model. For example, a linear regression imputer can be used to impute missing values in a continuous variable, while a logistic regression imputation model may be constructed to impute a discrete missing variable. When a missing variable is skewed and its transformation to normality is impossible, Predictive Mean Matching (PMM) is usually suggested as an imputer. PMM can be seen as a type of random k-nearest-neighbor method (Waddell, 2009). In a PMM model, imputed values are simply sampled from the observed values of the missing variable, so that the distribution of the imputed values is similar to that of the observed values. Especially, PMM is desirable when the

sample size is large and a missing variable is “semi-continuous” for which many values are equal (White et al., 2011).

4.2.4 Recursive Partitioning in MICE

In MICE, it is required to build an imputation model for each of the variables with missing values. In practice, however, specifying the imputation models is not an easy task for at least two reasons. First, missing variables to be imputed may have complex distributions. In this case, standard parametric models are not appropriate for the imputers. Also, data transformations to normality are not always possible. Second, interactive and nonlinear relationships among the variables in the imputers may exist. The nature of these interactions and nonlinearity is usually unknown a priori. It is very laborious to add the interaction and nonlinear effects to the imputation models. To mitigate these two challenges, recent studies suggest the use of recursive partitioning as an imputation engine within MICE.

One of the most popular recursive partitioning techniques is Classification And Regression Trees (CART) (Breiman et al., 1984). CART is called either classification trees or regression trees, depending on the response variable of interest being categorical or continuous, respectively. In CART, a predictor space is partitioned so that observations are clustered into several groups that are relatively homogeneous with respect to the response variable. The best partitions are found by recursive binary splits of the predictors. The resulting series of splits are represented in a tree structure. The leaves of the tree represent the groups of observations, and the values in each leaf approximate the conditional distribution of the response variable based on the predictors. Since all splits are conditional on the previous splits, complex interactions are automatically detected in the tree. The grown tree is usually pruned to avoid overfitting and for easier interpretation. However, when a tree is built for imputation, this pruning procedure is not desirable (Burgette and Reiter, 2010).

However, there are two major disadvantages of CART because of its hierarchical nature. First, the trees of CART may be suboptimal. A “best” split is determined only locally in each leaf regardless of future splits. Second, CART may produce unstable trees because trees may vary markedly from sample to sample. In other words, small changes in a sample may lead to different initial splits and then yield quite different final trees. Alternatively, another popular recursive portioning method is Random Forests. While CART builds a single tree, Random Forests generates multiple trees and averages them (Breiman, 2001). Variation in the individual trees is produced by bootstrapping and random variable selection. Instead of using all observations, a bootstrap sample is used to build each tree. Rather than using all variables, a subset of variables is randomly selected to find the best split at each leaf (Doove et al., 2014).

4.3 CASE STUDY: PARCEL DATA IMPUTATION

Integrated land-use and transportation models are trending towards using disaggregated spatial resolution, such as parcel. Although some concerns for using micro-level data have been pointed out, such as large data requirements, long computing times, and stochastic variations (Wegener, 2011), the advantages of using parcels as the spatial unit of analysis are enormous. First, parcels

are more behaviorally realistic in modeling the built environment than uniform grid cells as well as coarse zones. In the real world, transaction, regulation and development usually occur at the parcel level. Second, using the parcel geography allows us to measure walking-scale accessibility by representing people's activity locations on parcels. Walking access to transit stops or grocery stores is increasingly important from public health and policy perspectives (Waddell, 2009).

In the U.S., parcel data are usually managed by local governments that assess property taxes. There is huge variation in data completeness and data quality across jurisdictions, and missing values are common in variables that are necessary for developing integrated models, including building square feet, building type, number of stories, year built, building value, land value, land use, and so on. Especially for parcels that are exempt from property taxes but nevertheless important for models, there tend to be more missing values on the attributes. Therefore, there is a significant need for handling missing values in parcel data as the research and practice of integrated models move to use parcel as the spatial boundary.

For this study, we picked Multnomah County in Oregon as our study area. Multnomah County includes Oregon's largest city, Portland. The parcel data are obtained from Metro's Regional Land Information System (RLIS), which provides a variety of geographic information for the Portland metropolitan area. Metro and the regional partners update the parcel data every quarter. For this study, we focus on one parcel data set updated in the fourth quarter of 2011. The parcel data set provides information such as parcel area in square feet, building floor area in square feet, building value in dollar, land value in dollar, land use type, and so on. We detect a modest amount of missing entries from the parcel data set. The missing data pattern is shown in Table 5.1. Among the total of 271,977 parcels, 83.7% are completely observed for all of the variables. The remaining parcels have at least one missing entry for any of the variables. The largest number of missing entries (i.e., 37,037 or 13.6%) is observed in the land-value variable, while the smallest number of missing entries (i.e., 2,097 or 0.8%) is in the land-use variable. Among the parcels with missing entries, 60.6% are missing for only one variable, 37.3% for two variables, and 2.2% for three variables. We also notice that the missing entries are scattered among the variables. Therefore, we conclude that the missing values in the parcel data set follow an arbitrary pattern, which makes MICE a suitable method for imputing the multivariate missing values in the parcel data set.

Table 4.1 Missing Data Pattern of 271,977 Parcels

	Parcel Area	Land Use	Building Value	Building Floor Area	Land Value	Frequency
	1	1	1	1	1	227,591
	1	0	1	1	1	2,097
	1	1	1	0	1	2,825
	1	1	0	1	1	13
	1	1	1	1	0	21,959
	1	1	0	0	1	2,414
	1	1	1	0	0	14,122
	1	1	0	1	0	1
	1	1	0	0	0	955
Total	*none*	*2,097*	*3,383*	*20,316*	*37,037*	*271,977*

(1 = observed & 0 = missing)

4.3.1 Setting Up MICE: Choice of Predictors

Assuming the MAR mechanism and arbitrary missing pattern, we now set up MICE to impute all the multivariate missing values on the Multnomah County parcels. The first step is to determine which variables are included as predictors in the imputation models. Table 5.2 is a list of the variables. We include all of the four missing variables: building floor area, building value, land value, and land use. Since the first three continuous variables include huge values from which it is hard to draw meaningful interpretations, we scale down by dividing each of the three variables by parcel area. As a result, we use three new missing variables, namely, building floor area per parcel area (i.e., FAR; floor area ratio); building value per parcel area (i.e., BVPA); and land value per parcel area (i.e., LVPA). It is also important to note that these three missing variables are semi-continuous because a large portion of their values is zero and the remaining portion is continuous. The other missing variable (i.e., LU) is categorical with nine levels, including agriculture, commercial, forest, industrial, multifamily residential, public/semi-public, rural, single family residential, and vacant. Additionally, four non-missing variables are included to increase the prediction power of the imputation models: Area, District, POP10 and JOB11. Area is a continuous predictor, representing the parcel size in square feet, while District is a categorical predictor with 10 polygons covering Multnomah County among 20 districts that divide the Portland metro area for a variety of planning and analysis purposes. The other two non-missing variables are obtained from outside data sources. POP10 represents the 2010 population size of a Census block group the parcel belongs to, which is obtained from U.S. Census data. JOB11, or the number of jobs in a Census block group the parcel belongs to, is extracted from the 2011 Workplace Area Characteristics (WAC) file of Longitudinal Employer-Household Dynamics (LEHD). We use these two outside variables as a proxy for neighborhood characteristics of parcels.

Table 4.2 A List of Predictors Included in the MICE Imputers

Variable Name	Description	Variable Type	Number of Missing Parcels	Min	Max	Mean	Median
FAR	building floor area per parcel area (in square feet)	semi-continuous	20,316 (7.5%)	0	4,457	1.22	0.25
BVPA	building value per parcel area (in dollar)	semi-continuous	3,383 (1.2%)	0	588,178	278.36	20.05
LVPA	land value per parcel area (in dollar)	semi-continuous	37,037 (13.6%)	0	2,232	19.42	16.47
LU	9 land use types	categorical	2,097 (0.8%)	-	-	-	-
Area	parcel area (in square feet)	continuous	no missing	2	30,956,896	35,772.12	5,433.68
Disctrict	10 districts covering Multnomah County	categorical	no missing	-	-	-	-
POP10	2010 population size by Census block group	continuous	no missing	8	4,746	1,657.36	1,407.00
JOB11	2011 number of jobs by Census block group	continuous	no missing	4	30,390	1,173.45	263.00

4.3.2 Setting Up MICE: Choice of Imputers

The aim of this study is to impute missing values in four variables, namely, FAR, BVPA, LVPA, and LU using MICE. The key to the success of a MICE application is how well the imputation model is specified for each missing variable. In the parcel data set described previously, missing values are observed in a mix of continuous and categorical variables with complex distributions. Especially, three of them (i.e., FAR, BVPA, and LVPA) are continuous but inflated with zeroes. In addition, there may exist a variety of interactions between the predictors. For example, we can easily think of the interaction effects of land use type and district geography on the building floor area ratio on a parcel. Multifamily residential parcels located in CBD district tend to have a higher FAR (i.e., a higher building on a smaller size of parcel) than the same land use type of

parcels in other districts. Manually testing such interaction effects in the imputation model is a very time-consuming task with no guarantee of success. For these reasons, we initially consider two recursive partitioning methods (i.e., CART and Random Forests) as an imputation model of MICE. We generate five imputed data sets (m = 5 in Step 8) after five iterations (l = 5 in Step 7).

4.3.3 Visual Diagnosis

Because we never observe the missing values, we cannot quantitatively assess how well each imputation technique performs. An important step in multiple imputation is to diagnose whether imputations are plausible or not. As a simple visual way of checking the plausibility of the imputations, it is useful to compare the distributions of original data and imputed data. Figure 5.1 shows the distributions of the three continuous variables as individual (jittered) points, in which blue points are observed data and red points are imputed data. The zeroth imputation of all the panels contains blue points only, representing the original distribution. The red points follow the blue points reasonably well without showing any data points that are clearly impossible. However, it should be noticed that the CART-based MICE did not produce proper imputes for extreme values of FAR and BVPA. They are severely right skewed; the skewness of FAR, BVPA, and LVPA is 265, 92 and 21, respectively. For more graphical diagnostic tools, refer to (Abayomi et al., 2008). To be able to quantitatively assess the performance of each imputation model, we create a validation data set and compare how well each imputation recovers the true values, which will be discussed in the next section.

Figure 4.3 Visual Diagnosis of the CART-based MICE

4.4 ASSESSMENT OF MICE

To evaluate the performance of MICE, we conduct cross validation for the three different MICE specifications (i.e., MICE via PMM, MICE via CART, and MICE via Random Forests). The combination of MICE with PMM is widely used because, as a semi-parametric approach, it can handle any variable types. Especially, PMM is often used to impute a semi-continuous missing

variable. It is expected that CART and Random Forests can automatically capture complex interaction effects on missing variables with little tuning effort needed in developing the imputation models.

To this end, we first create a validation set by removing parcels with any missing values from the original parcel data, which produces a validation data set of 227,591 parcels with the eight complete variables. Then, from the validation data set, we artificially make 5% of the values in the four variables (i.e., FAR, BVPA, LVPA, and LU) to be missing at random (MAR), yielding a training set of the same number of parcels but with missing values. As shown in Table 5.3, the artificial missing values are evenly scattered across all the missing variables, indicating an arbitrary pattern of missing data.

Table 4.3 Missing Data Pattern of 227,591 Parcels in the Training Data Set

	Area	District	POP10	JOB11	FAR	LVPA	LU	BVPA	Frequency
	1	1	1	1	1	1	1	1	185,350
	1	1	1	1	1	1	0	1	9,745
	1	1	1	1	0	1	1	1	9,711
	1	1	1	1	1	1	1	0	9,883
	1	1	1	1	1	0	1	1	9,729
	1	1	1	1	0	1	0	1	502
	1	1	1	1	1	1	0	0	532
	1	1	1	1	0	1	1	0	505
	1	1	1	1	1	0	0	1	512
	1	1	1	1	0	0	1	1	502
	1	1	1	1	1	0	1	0	516
	1	1	1	1	0	1	0	0	19
	1	1	1	1	0	0	0	1	33
	1	1	1	1	1	0	0	0	26
	1	1	1	1	0	0	1	0	26
Total	*none*	*none*	*none*	*none*	*11,298*	*11,344*	*11,369*	*11,507*	*227,591*

With this training data set, we perform MICE with three different imputation models (i.e., PMM, CART and Random Forests). The "mice" package of R is used for the PMM-based MICE and the CART-based MICE (Abayomi et al., 2008), while we use the "CALIBERrfimpute" package of R for the MICE imputation based on Random Forests (Buuren and Groothuis-Oudshoorn, 2011). Each of the three MICE imputation runs produces five imputed data sets as designed. It is often suggested that small numbers of imputed data sets (e.g., 3 or 5) are sufficient unless missing rates are unusually high, say 50% (White et al., 2011). For each imputed data set, we compare the imputed values of the training set with the original complete values of the validation set. We calculate root mean square error (RMSE) and R-squared for the continuous variables, where an accuracy rate is computed for the categorical variable.

Table 5.4 shows the validation results. We find that both of the non-parametric MICE imputations perform much better than the semi-parametric MICE for the categorical missing variable (i.e., LU), indicated by a higher average accuracy rate for MICE via Random Forest (0.934) and MICE via CART (0.927) than MICE via PMM (0.780). However, it is found that there are noticeable differences in performance between the two non-parametric imputations for the continuous missing variables. The MICE with Random Forests performs best for LVPA; the average R-squared for LVPA is 0.450 with Random Forests, 0.426 with CART, and 0.234 with PMM. As for FAR and BVPA, the MICE with CART performs best with the highest average R-square (0.665 for FAR and 0.506 for BVPA). Surprisingly, the Random Forests within MICE shows a similar performance to the PMM within MICE (0.568 and 0.569 for FAR; 0.405 and 0.406 for BVPA). The lower performance of Random Forests in this study can be explained by a different level of skewness for the three continuous missing variables. As mentioned earlier, as the three variables are semi-continuous, they are all severely positive or right skewed. However, there is a significant gap in the skewness level among the semi-continuous missing variables. The skewness of FAR, BVPA and LVPA in the training data set is 65, 222, and 16, respectively. The relatively high skewness, especially of BVPA, results from some extreme data points or outliers as shown in the blue dots of Figure 5.1. Since we build only 10 trees in Random Forests due to computational burden, it is possible to easily miss the effects of such extreme values during the tree building process of drawing bootstrap samples and selecting random input variables. Therefore, we conclude that the CART-based MICE is the most appropriate for imputing missing values in a large data set, such as the parcel data in this study.

Table 4.4 Validation Results of MICE (PMM vs. CART vs. Random Forests)

Imputation Model	Missing Variable	Validation Measure	Multiple Imputed Data Set					Average
			m=1	m=2	m=3	m=4	m=5	
PMM	FAR	RMSE	0.270	0.286	0.268	0.294	0.283	0.280
		R-squred	0.590	0.558	0.586	0.547	0.565	0.569
	BVPA	RMSE	50.871	38.446	51.229	41.432	34.979	43.391
		R-squred	0.285	0.457	0.181	0.453	0.656	0.406
	LVPA	RMSE	19.157	17.104	17.184	17.657	16.662	17.553
		R-squred	0.179	0.250	0.236	0.232	0.276	0.234
	LU	Accuracy Rate	0.780	0.783	0.781	0.776	0.782	0.780
CART	FAR	RMSE	0.247	0.222	0.248	0.238	0.238	0.239
		R-squred	0.684	0.710	0.637	0.655	0.641	0.665
	BVPA	RMSE	43.843	34.327	30.682	34.442	45.009	37.661
		R-squred	0.331	0.571	0.670	0.563	0.393	0.506
	LVPA	RMSE	14.484	15.269	14.306	13.801	13.955	14.363
		R-squred	0.422	0.384	0.427	0.454	0.443	0.426
	LU	Accuracy Rate	0.927	0.925	0.930	0.924	0.928	0.927
Random Forest	FAR	RMSE	0.284	0.313	0.282	0.281	0.229	0.278
		R-squred	0.543	0.506	0.559	0.553	0.676	0.568
	BVPA	RMSE	76.119	73.102	36.429	36.961	33.766	51.275
		R-squred	0.195	0.187	0.541	0.515	0.585	0.405
	LVPA	RMSE	14.110	13.435	13.692	13.953	14.978	14.034
		R-squred	0.446	0.479	0.467	0.450	0.409	0.450
	LU	Accuracy Rate	0.936	0.934	0.933	0.935	0.932	0.934

4.5 CONCLUSION

In this paper, we have demonstrated MICE as a flexible method to implement multiple imputations for the multivariate missing values in parcel data. We consider MICE via PMM, and two recursive portioning techniques (i.e., CART and Random Forests) as imputation engines for MICE. To assess their performance, we have conducted cross-validation and found that, under limited computing power, the CART-based MICE has performed best for imputing missing values in both continuous and discrete variables, especially when the distribution of the continuous variables to be imputed is skewed.

For future research, we plan to improve the performance of Random Forests-based MICE by increasing the number of trees, which is expected to produce more reliable imputers. In addition, although we have generated multiple imputed data sets, we have not taken full advantage of them. We plan to conduct standard statistical analyses with the resulting multiple imputed data sets and assess the uncertainty introduced via imputed data. Lastly, this study focuses primarily on recursive partitioning methods as an imputation model. There are many other machine-learning techniques, such as k-nearest neighbor, support vector machine, and Bayesian network that have been used in data imputation. It would be interesting to benchmark them with the methods in this study.

5.0 DISCREPANCY ANALYSIS OF ACTIVITY SEQUENCES

Over the past four decades, the goals of urban transportation planning and policies have shifted from meeting "long-term, supply-oriented, mobility" needs to facilitating "short-term, demand-oriented, accessibility" needs. This shift has created a new paradigm called activity-based travel behavior analysis (Pinjari and Bhat, 2011). The underlying theory of the activity-based analysis recognizes that travel is a derived demand to participate in activities that are separated in time and space. The theory implies that "travel can be best understood in the broader context of activity patterns" (Ettema, 1996). An individual's activity pattern is a complex phenomenon resulting from interactions of multiple dimensions, such as timing, duration, locations, activity types, travel mode, trip chaining, activity jointness, activity substitution, activity priority, activity planning horizon, and so on (Burnett and Hanson, 1982).

Given that it is very difficult to capture their full complexities with all the dimensions, two general approaches have been used to measure the complexity of activity-travel patterns (Burnett and Hanson, 1982). One is decomposing an individual's activity pattern into numerous dimensions and generating separate measures for each of the dimensions. The other is treating the pattern as a multidimensional "holistic" entity. Currently, the first approach is dominant in activity-pattern research, in part because most existing operational activity-based travel forecasting systems are implemented on a micro-simulation framework that consists of a series of calibrated econometric models to address the multiple dimensions either individually or jointly (Davidson et al., 2007). Discrete choice models, discrete-continuous choice models, and hazard-based duration models are widely used as the basis of the micro-simulation implementation framework (Pas, 1983). The second approach to measuring activity patterns, which relates to this paper, was popular in the early stage of the activity-based travel behavior analysis. Many transportation researchers recognized early on the complexity of activity patterns and emphasized the need to understand individual activity patterns as a whole (Koppelman and Pas, 1984; Pas, 1983; Recker et al., 1985). The holistic approach focuses more on interpreting people's daily or weekly activity patterns into homogeneous groups, and identifying determinants or constraints that influence the homogeneous patterns.

The early efforts on the holistic approach fall into two categories (Pas, 1983). First, each activity pattern is described by numerous measures, and then the measures are used for factor analysis or principal components analysis to identify its salient features. The latter information is often used to classify the whole set of activity patterns into a small number of similar groups. The second holistic method is comparing individuals' activity-travel patterns with each other, which is often represented on time-slice variables. The comparison produces a matrix of pairwise dissimilarities between the patterns, which is subsequently used for cluster analysis.

Given our limited knowledge about human activity decisions, both atomistic and holistic approaches to accounting for the complexity of activity patterns are equally important and complementary (Burnett and Hanson, 1982). The atomistic or decomposing approach serves as a

core part of activity-based models to predict the patterns, while the holistic approach provides theoretical and empirical foundations by identifying travel determinants. However, it is surprising that although attention to the former is prominent today, the latter has seen little progress since the development of the cosine similarity index by Koppelman and Pas (1984) and the feature extraction of Recker, McNally and Root (1985), even if those classifications may be not satisfactory in that they only explain a small amount of variability within the clusters (Schlich and Axhausen, 2003).

The holistic approach to understanding complex activity patterns has recently received new attention since the introduction of sequence alignment methods into time use and transportation research (Wilson, 1998). In existing literature, sequence alignment methods are almost always combined with cluster analysis. Although this cluster-based approach has been proven powerful for discovering a typology of activity patterns, it does not seem successful in identifying various contexts that would impact the patterns. This is because cluster analysis may cause too much information loss as a result of reducing a large set of observations into a limited number of clusters (Studer et al., 2011). Consequently, we need a direct analysis of the association between pairwise dissimilarity measures and explanatory variables without any prior clustering. Inspired from the work of Studer et al. (2011), this paper proposes a new combination of sequence alignment with ANOVA-like tools, not with cluster analysis. The proposed method was originally developed in ecology under the name of a non-parametric MANOVA (Anderson, 2001), and recently introduced into sociology with the name of discrepancy analysis (Studer et al., 2011) and ANODI (the Analysis of Dissimilarity) (Bonetti et al., 2013). In addition to the methodological combination, an induction tree is built to visualize how activity sequences may vary with the value of covariates.

5.1 SEQUENCE ALIGNMENT METHODS FOR ACTIVITY-PATTERN ANALYSIS

Sequence alignment methods, also known as optimal matching (OM), measure the dissimilarity between two sequences of characters by calculating the minimal cost of transforming one sequence into the other (Martin and Wiggins, 2011; Wilson, 1998). Two basic operations are used in the sequence transformation: substitutions and indels. The substitution operation replaces an element of one sequence with the different element located at the same position of the other sequence. The indel, an OM jargon standing for the insertion or deletion of an element, causes a one-position movement of all the elements to the right. The transformation costs for substitutions and indels are assigned by the researcher in advance either theoretically or empirically. A dynamic programming algorithm is usually used to repeat the sequence alignment for all pairs of sequences. The final output of sequence alignment is a pairwise distance matrix of activity sequences, which is almost always used as input for cluster analysis. The resulting cluster membership is often associated with other variables, either as a dependent variable (e.g., multinomial logit models) or as an independent variable (e.g., ANOVA).

Sequence alignment methods were initially developed in biology in the 1970s from the needs of analyzing DNA sequences of nucleic acids or protein sequences of amino acids, and introduced into social science in the 1980s (Abbott and Tsay, 2000). It was in the late 1990s that the

methods were first adopted for the analysis of people's activity patterns (Wilson, 1998). Recently, sequence alignment is applied in the goodness-of-fit testing for activity-based travel demand micro-simulation models, in which predicted activity-travel patterns are compared with the observed one at an agent level (Sammour et al., 2012). Table 5.1 summarizes some applications of sequence alignment methods to activity-pattern analysis.

Table 5.5 Applications of Sequence Alignment Methods to Activity-Pattern Analysis

Citation	Sequence Specification		Assigning Transformation Costs		Number of Clusters	Subsequent Data Analysis
	sequence element	time scale	substitutions	indels		
(Wilson, 2001)	activity only	30 min	varying by activity being compared	gap opening; extension	4	ANOVA
	activity; location; person present	30 min	complicated	gap opening; extension	4	ANOVA
(Shoval and Isaacson, 2007)	location only	1 sec	not used	gap opening; extension	3	contingency table analysis
(Saneinejad and Roorda, 2009)	activity; location	15 min	not used	gap opening; extension	9	descriptive statistics

Wilson (2001) examined the activity patterns of 248 Canadian women who were selected as a 5% random sample from the main time-use survey. The 248 activity diaries were converted into two sets of sequences: one was composed of only one dimension (i.e., activity type) and the other had three dimensions (i.e., activity type, location and the presence of other persons). The author defined 15 activity types regardless of in-home or out-of-home; five locations (i.e., home, workplace, other place, traveling or unknown); and five types of persons present with the survey respondent (i.e., alone, household members, friends, other persons or unknown). Both the activity-only sequences and the activity-setting sequences were specified in 30-minute time intervals. In the activity-only sequence alignments, the substitution costs varied with the type of activities being compared. The indel costs were also further refined into two types; a gap open penalty for the first indel position and a gap extension penalty for all the subsequent indels. For the activity-settings sequences, the transformation costs of substitutions were more complicated due to the many ways of combining levels of the activity settings, and the same indel costs were assigned as in the activity-only sequences. Both sets of sequences were visually clustered into four similar patterns. Subsequently, the author conducted ANOVA to examine discriminatory power of the cluster membership with regard to socioeconomic characteristics.

Shoval and Isaacson (2007) investigated the moving paths of visitors to the Old City of Akko in Israel. For this study, 40 visitors were given GPS devices to track their movements in space and time. The study area was divided into 26 polygons, with each polygon representing a single location, and each visitor's polygon locations were consecutively recorded every second during

the visit. Since visitors started their trip at different times of day and their visit durations were different from each other, the sequence lengths of visitors varied. Only indel operations were used to align the location sequences of different lengths. The authors identified three distinct moving paths. In addition, they conducted a contingency table analysis.

Saneinejad and Roorda (2009) measured similarities between weekly activity sequences of 282 individuals who participated in a special survey in which, among other information, respondents were directly asked to describe activities that they normally do every week. The authors defined 10 activity types (i.e., nine routine and one non-routine) and two activity locations (i.e., in-home and out-of-home). Two letters representing activity type and location were specified on every 15-minute time interval of five weekdays for each individual, and thus the length of all sequences is equal to 480. No substitutions were used, while two types of indels were defined: gap opening indels and gap extending indels. The authors identified nine different schedules of routine weekly activities. Then, they described socioeconomic characteristics of individuals within each cluster.

It has been demonstrated that sequence alignment methods outperform in classifying activity-travel patterns compared to other conventional dissimilarity measures, such as Euclidean distance measures and signal-processing theoretical measures (Joh et al., 2001). Only sequence alignment methods can capture sequential information imbedded in activity patterns. This unique ability of sequence alignment methods may yield better cluster solutions that are more likely to be sensitive to activity-travel constraints. On the other hand, there are many controversial issues for using sequence alignment in social science rather than in biology (Aisenbrey and Fasang, 2010). The issues include the meaning of indels and substitutions in the context of human behavior analysis; the arbitrary assignment of transformation costs for substitutions and indels; required symmetry of the pairwise distance matrix; the lack of proper support of multi-dimensional analysis; and time distortion by indels. Fortunately, a “second wave” of sequence alignment toward methodological improvements is currently observed in sociology (Aisenbrey and Fasang, 2010) as well as transportation (Joh et al., 2002; Wilson, 2008).

5.2 **DATA AND METHODS**

The Portland metropolitan portion of the 2011 Oregon Travel and Activity Survey (OTAS) is used for this study. The portion of the survey records all locations visited by approximately 15,000 persons in nearly 6,500 Portland-area households during a scheduled day. A random sample of 1,000 persons, which accounts for about 6.5% of the main survey, is selected to reduce the computational burden of performing a series of proposed methods. In addition, for simplicity reasons 24 types of self-reported original activities are aggregated into 12 major activity types, as shown in Table 2. It should be noted that all at-home activities are grouped into three different categories: home at the beginning of the day (HB); home returning temporarily in the middle of the day (HR); and home at the end of the day (HE). This categorization of at-home activities may help partially avoid time distortion that often occurs by indels. Time distortion by indels is a unique feature of sequence alignment in matching sequences of varying lengths. In the case of aligning sequences involving timing and duration of episodes, the indel operations need to be carefully used to prevent excessive time distortion (Lesnard, 2010; Wilson, 2001). A simpler

way of avoiding time distortion, as in this study, might be roughly disaggregating a sequence state (i.e., at-home activity) by time periods (i.e., at the beginning, middle and end of the day).

Table 5.6 Activity Aggregation and Average Duration (1,000 Persons)

12 Activity Types (Aggregated)	24 Activity Types (Originally Self-Reported)	Code	No. of Episodes	Average Duration (min.)
Home in the beginning	work at home; all other activities at home	HB	968	495
Home returning temporarily	work at home; all other activities at home	HR	410	121
Home in the end	work at home; all other activities at home	HE	848	560
Work	work; all other activities at work; work related	WK	625	312
School	attending class; all other activities at school	SC	218	355
Escort	drop off; pick up	EC	270	10
Eat out	eat meal outside of home	EO	196	44
Household maintenance	routine shopping; major shopping; household errands	HM	516	28
Personal business	service private vehicle; personal business; health care	PB	226	71
Social recreation	civic/religious activities; outdoor/indoor recreation; visit friends; loop trip	SR	375	134
Other	Other	OT	4	387
Trip for the activity	not categorized in the survey, but explicitly included for this study	TR	3656	19

The goal of this paper is to identify determinants that influence individuals' daily activity-travel patterns from the holistic perspective. To achieve this goal, four sequential steps are proposed: 1) representing an individual's activity diary as a sequence of characters; 2) performing sequence alignment to produce a pairwise distance matrix among all activity sequences; 3) conducting discrepancy analysis to examine the association between activity sequences characterized by the distance matrix and one or more categorical predictors; and 4) building an induction tree to help interpret how activity sequences change with the predictors. All of these steps are implemented in R with the TraMineR package (Gabadinho et al., 2011; Studer et al., 2011).

5.2.1 Sequence Representation

Like all other household activity-travel surveys, the OTAS data is released in a spell format, where each record represents an activity spell or episode of variable duration undertaken by a person, and each person may have more than one activity episode during a day. To conduct sequence alignment, an individual's activity diary first needs to be converted from the spell format to a sequence of letters representing activity states on a fixed time scale. In this study, individual activity diaries are reconstructed with 12 aggregated activity types on five-minute time intervals from 3 a.m. through 2:59 a.m. the next day, so that each activity sequence consists of 288 consecutive activity codes. Figure 1 shows a "sequence index plot" and a "state

distribution plot" for the transformed activity sequences. In the activity sequence index plot, the first 10 sequences of the subsample are individually rendered with stacked bars depicting the activity states over time. The sequence index plot is useful to visualize individual activity trajectories and the duration spent in each successive activity episode. The activity state distribution plot shows the distribution of activity states at each time interval for all sequences in the subsample (Gabadinho et al., 2011). Both plots will be used for building an induction tree as the displayed node content.

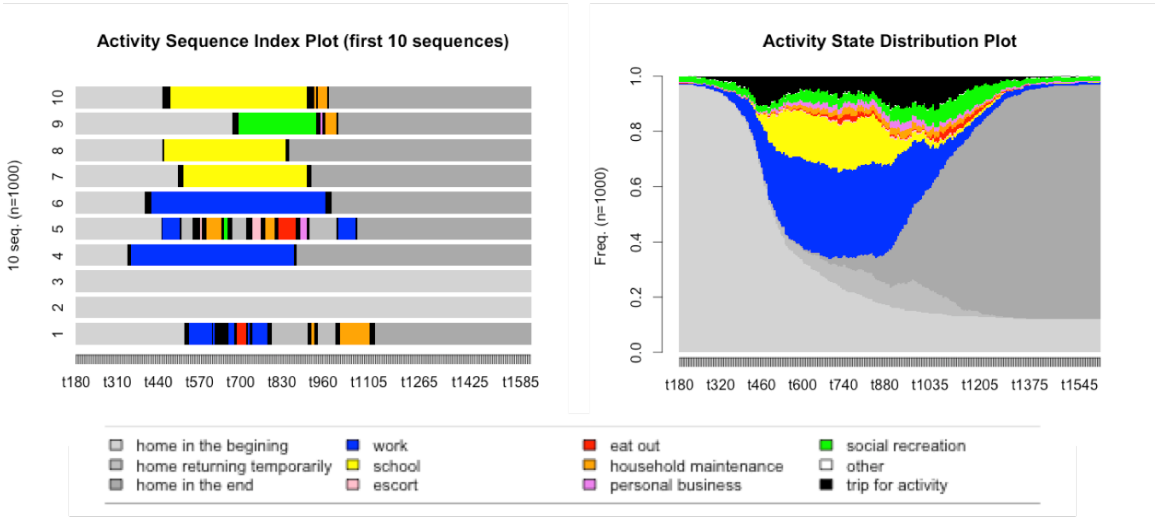


Figure 5.4 Activity Sequence Index Plot and Activity State Distribution Plot

5.2.2 Sequence Alignment

Once activity diaries are represented as state sequences with a single attribute of activity type on time intervals of a fixed length, the next step is to align the activity sequences for measuring the dissimilarities between each pair. One of the controversial issues of sequence alignment methods in human behavioral applications, as in this study, is how to set up transformation costs or penalties for the two basic operations – substitutions and indels. In fact, existing literature suggests various ways of assigning the transformation costs. For example, the indel cost can be separated into two types – gap opening penalty and gap extension penalty (Saneinejad and Roorda, 2009; Shoval and Isaacson, 2007; Wilson, 2001). A higher penalty can be given for gap openings in the first time position of an activity episode than for gap extensions in all the subsequent time positions of that episode. In addition, instead of using a single substitution cost, it is possible to develop a matrix that specifies the substitution costs between all pairs of sequence states (Wilson, 2001). The substitution cost matrix can also be derived from the probability of transition between sequence states, given that a higher transition rate between two states may indicate a less costly substitution of these states (Lesnard, 2010). However, this study follows the default settings, focusing more on evaluating the validity of sequence discrepancy analysis that is new in transportation research. In other words, the substitution and indel costs are set to 2 and 1, respectively, in this study.

5.2.3 Discrepancy Analysis of Activity Sequences

This paper introduces a new methodology for a direct analysis of the association between complex objects described by a distance matrix and one or more categorical variables; there is no need for any prior data reduction technique, such as cluster analysis. This new method is a generalization of MANOVA (Anderson, 2001). The standard MANOVA is concerned with measurable objects that are characterized by multiple continuous dependent variables. On the other hand, the generalized MANOVA can handle complex objects that are not directly measurable, but can be described by a pairwise dissimilarity matrix, such as ecosystems, life trajectories and activity diaries.

The purpose of ANOVA is to test for significant differences among group means by analyzing the variance. Recall that given a certain sample size, the sample variance is a function of the sum of squared deviation from the mean, or *SS* for short. The essence of ANOVA is partitioning the total variance (SS_T) into two different sources of variance: the within-group variance (SS_W) and the among-group variance (SS_A). Then, the two variance sources are compared to produce the test statistic of F-ratio. The larger the F-ratio value, the more likely it is to reject the null hypothesis that there is no difference among the group means.

For one-way univariate ANOVA, in which a single response variable is linked to one predictor, SS_W is the sum of (deviation) squares between individual cases and their group mean, while SS_A is the sum of (deviation) squares between group means and the overall sample mean. Next, consider one-way multivariate ANOVA, in which multiple responses are associated with one predictor. Traditionally, MANOVA compares the among-group variance/covariance matrix versus the within-group variance/covariance matrix, instead of the corresponding variances. The covariance here is included because the multiple response variables may be correlated and we need to consider these correlations for the significance test. In that case, the correlations between response variables do not really matter as in independent activity-sequence objects; however, we can simply add up the sums of squares across all response variables. Then, we can construct an F-ratio test statistic, as in the univariate ANOVA problem. Such an additive partitioning of the sums of squares in MANOVA can also be thought of geometrically, as shown in Figure 5.1 (Anderson, 2001).

The key to generalize the geometric approach of MANOVA to complex objects is based on the fact that “the sum of squared distances between points and their centroid is equal to the sum of squared interpoint distances divided by the number of points” (Anderson, 2001). This relationship has an important implication that an additive partitioning of sums of squares can be obtained without calculating the central locations of groups. For the Euclidean distance measure, the relationship between distances to the centroid and interpoint distances was well known early on. It was found that this key relationship holds for any non-Euclidean distance measures equivalently. The importance of this finding is substantial because, unlike the Euclidean distances, the calculation of a central location for non-Euclidean distances, such as a pairwise distance matrix resulting from sequence alignments, is often problematic. Further, Studer et al. (2011) demonstrate that if the distance measure is non-Euclidean, the non-Euclidean distances do not need to be squared before summing them. In short, the new method generalizes the notion of “sum of squares” in ANOVA to non-Euclidean measures of dissimilarity.

Once the test statistics of pseudo F-ratio with any non-Euclidean distance measure is obtained, we need to test the statistical significance. However, we cannot conduct the classical F-test as in the standard ANOVA because the distances between complex objects are not normally distributed, and thus the pseudo F-ratio statistic does not follow a Fisher distribution under the null hypothesis. Instead, we need to consider a permutation test in order to obtain a new distribution of the pseudo F-ratio under the null hypothesis. The permutation test works as follows: First, the complex objects are exchanged among the different groups of a categorical predictor through a random permutation. Second, a new pseudo F-ratio statistic, called $F_{permuted}$, is computed. Third, the first and second steps are repeated for all possible permutations, which give the entire distribution of the pseudo F-ratio statistic under the true null hypothesis. Fourth, from this distribution, the p-value of the observed pseudo F-ratio statistic ($F_{observed}$) is assessed by evaluating the proportion of $F_{permuted}$ that are higher than $F_{observed}$. Since the number of all possible permutations is often huge, it is usually practical to perform 1,000 permutations for tests with a 5% significance level (Anderson, 2001; Studer et al., 2011). Table 3 compares standard ANOVA versus generalized MANOVA in one-way design with respect to the calculation of a test statistic and the significance test.

Table 5.7 Standard ANOVA vs. Generalized MANOVA: One-Way Design

	Standard ANOVA	**Generalized MANOVA**
Test Statistic	$F - ratio = \frac{SS_A/(a-1)}{SS_W/(n-a)}$	$pseudo\ F - ratio = \frac{SS_A^*/(a-1)}{SS_W^*/(n^*-a)}$
	• $SS_T = SS_W + SS_A$	• $SS_T^* = SS_W^* + SS_A^*$
	• SS_T is the sum of squared Euclidean distances from individuals to the grand centroid	• SS_T^* is the sum of all pairwise distances divided by the number of objects
	• SS_W is the sum of squared Euclidean distances from individuals to their group centroid	• SS_W^* is the sum of all pairwise distances within groups divided by the number of objects
	• SS_A is the sum of squared Euclidean distances from group centroids and the grand centroid	• $SS_A^* = SS_T^* - SS_W^*$
p-value	F test	permutation test

Note: (1) *a* refers to the number of levels or groups of a covariate; (2) in the generalized MANOVA, $n^* = n\ (n - 1)\ /\ 2$ where *n* is the sample size.

In the above, the one-way design of discrepancy analysis was discussed; that is, a single factor is associated with a distance matrix of the complex sequence objects. The one-way design can be nicely extended to a multi-way design in which multiple factors are involved. For more information on formula of SS_T, SS_W, and SS_A to compute a pseudo F test statistic in the multi-way design, refer to McArdle and Anderson (2001). As in the one-way discrepancy analysis, since the F distribution is not suitable for evaluating the pseudo F-ratio statistic, we consider the permutation test again. This paper conducts the multi-way discrepancy analysis to

simultaneously find out multiple factors that explain discrepancies among individuals' activity patterns. Since those factors are often highly correlated, their unique effects after controlling for other effects are more appropriate than their marginal effects obtained from a series of the one-way discrepancy analysis.

5.2.4 Tree-Structured Analysis of Sequences

Indeed the sequence discrepancy analysis with either a single factor or multiple factors can explain which variables have significant effects on the discrepancy among activity sequences. However, it is hard to tell what the effects are, namely, how activity sequences may vary with the value of the predictors. To complement this limitation, an induction tree is built. In general, trees work as follows: First, all sequence objects are located in an initial node. Then, each node is recursively partitioned by the value of a predictor. The predictor and the split are determined so that the resulting child nodes are different from one another as much as possible. The procedure is repeated at every new node until certain stopping criteria are met. As building a tree with state sequences is very rare in existing literature, however, this study follows the instructions suggested by Studer et al. (2011). Their tree is slightly different than popular tree algorithms, such as CHi-squared Automatic Interaction Detection (CHAID), in several aspects. First, while CHAID can only handle a categorical variable, the proposed tree is built on the basis of sequence objects that are neither continuous nor categorical. Second, the proposed tree is binary in that each node is split into only two subsamples, unlike multi-branch trees of CHAID. Third, a pseudo R^2 derived from the one-way discrepancy analysis is used as a node splitting criterion. In other words, each node is split with the predictor and its value achieving the highest pseudo R^2 value. Fourth, the significance of the one-way pseudo F-ratio that is determined through permutation tests is used as a stopping criterion. At each node, the tree stops growing a branch once the selected split encounters a non-significant F value.

5.3 RESULTS AND DISCUSSIONS

5.3.1 Sequence Discrepancy Analysis with Multiple Factors

Table 4 shows the results of the multifactor discrepancy analysis of activity sequences characterized by a pairwise sequence alignment distance matrix. For illustrative purposes, only seven significant covariates were selected, including one interaction term. The set of covariates explained approximately 19.4% of the total discrepancy among the daily activity sequences of 1,000 persons since the global pseudo $R^2 = 0.194$. The overall model was statistically significant, indicated by the global pseudo F value of 34.064 and $p < .05$. The most significant factor was an indicator of whether or not the person was a K-12 student. If the K-12 indicator was removed, the global pseudo R^2 decreased by 0.046. This difference was significant since the pseudo F value for that indicator was 57.034, which was a value attained less than five times out of the thousand permutations. The worker indicator was also significant. Removing the indicator variable from the model reduced the global pseudo R^2 by 0.043, which was significant since the pseudo $F_{Worker} = 52.904$ was attained less than five times amongst the thousand permutations. As for the other indicator variables, results indicated that full-time college students, persons with a driver's license, and adults over age 65 made moderate but significant contributions to explain

the total discrepancy of activity sequences. There also existed statistically significant discrepancy in activity sequences among five groups of different household size (1, 2, 3, 4 and 5+). Finally, the sequence discrepancy of activity diaries were strongly influenced by an interaction of the Worker and Adult-over-age-65 covariates

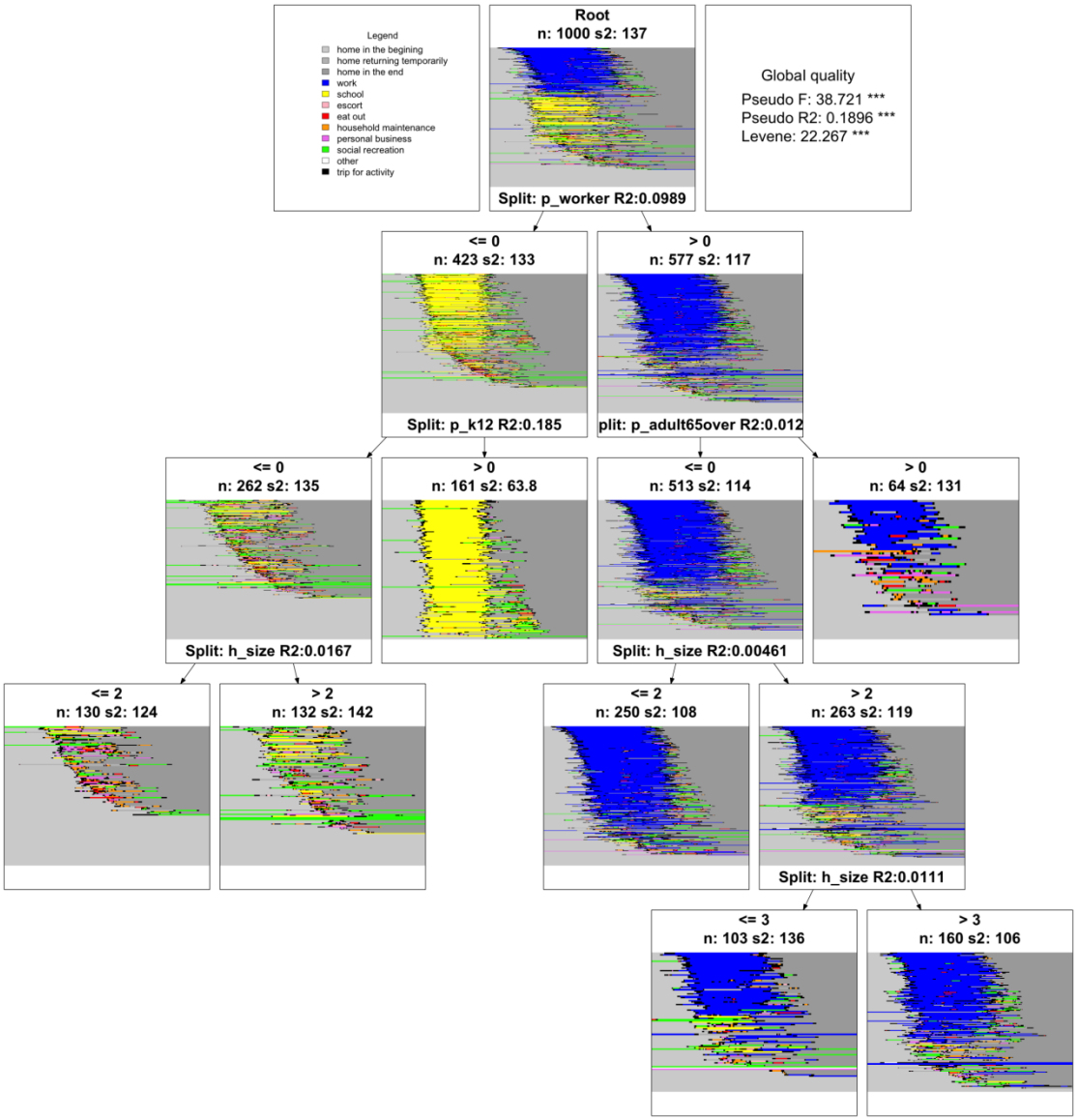
5.3.2 Tree Analysis of Activity Sequences

Using the same set of covariates as in the previous multifactor discrepancy analysis, but without the interaction term, an induction tree was built for the subsample of 1,000 activity trajectories, which is shown in Figure 2. To display more comprehensive information about the content at each node, the same tree was built with both activity sequence index plots and activity state distribution plots in the top and bottom panels of Figure 2, respectively. In addition to the node content, other important information is displayed on each node, including node size (*n*) and within-node discrepancy (*s2*). Moreover, a selected split covariate and its associated one-way pseudo R^2 are shown at the bottom of each parent node. The selected binary split of a covariate is indicated at the top of each child node.

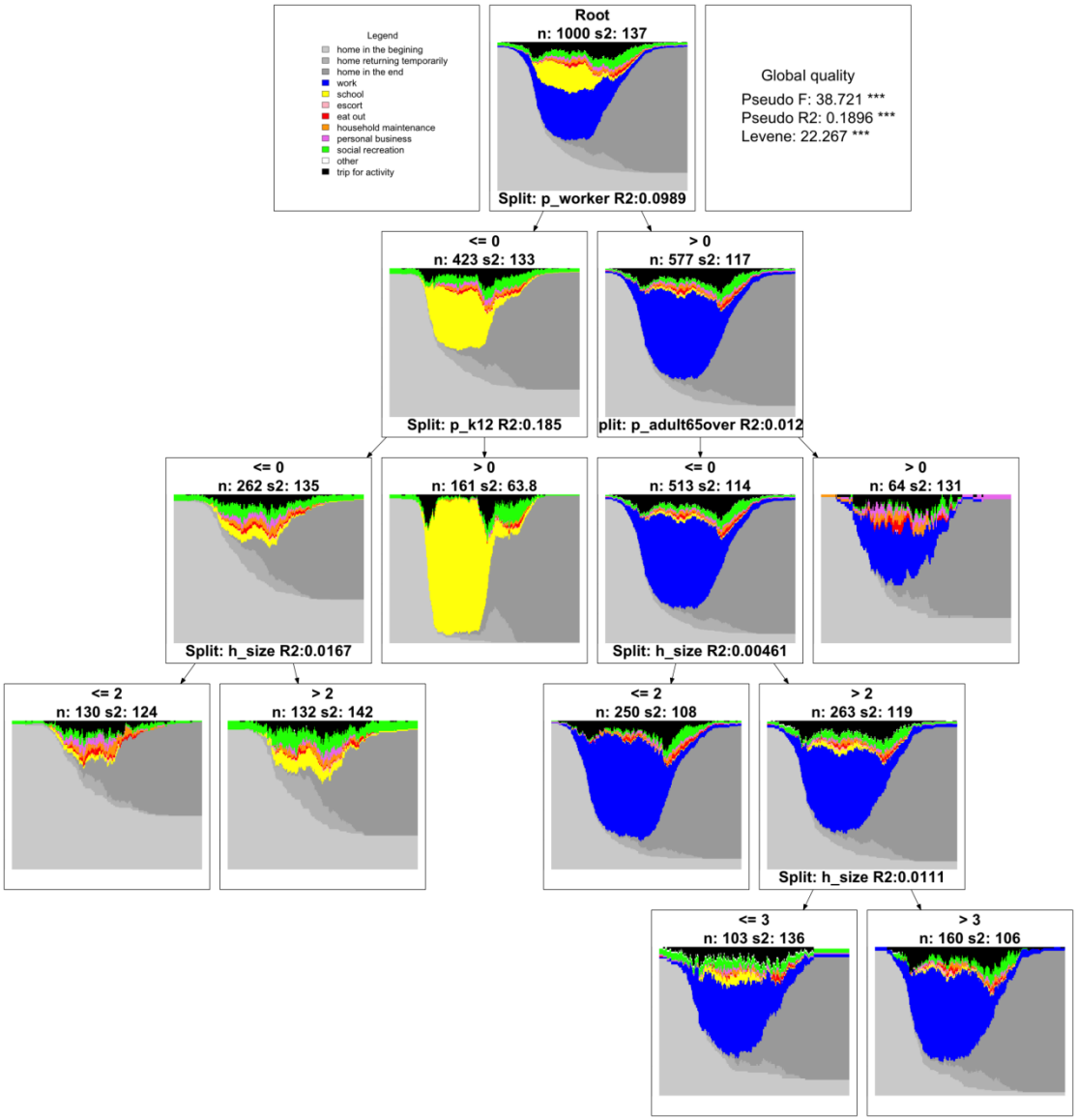
Table 5.8 Multifactor Discrepancy Analysis

Variable	**Variable Type**	**Pseudo F (for each variable)**	**∆Pseudo R^2 (for each variable)**	**p-value**
Worker	indicator	52.904	0.043	0.001
K-12 student	indicator	57.034	0.046	0.001
Full-time college student	indicator	2.829	0.002	0.008
Licensed	indicator	4.658	0.004	0.001
Adult over age 65	indicator	3.220	0.003	0.003
Household size	5 categories	2.361	0.002	0.017
Worker * Adult over age 65	interaction	2.337	0.002	0.027
Global		**Pseudo F (for total)**	**Pseudo R^2 (for total)**	**p-value**
		34.064	0.194	0.001

The global pseudo R^2 of the tree was 18.96%, which is slightly lower than that of the multifactor discrepancy analysis in Table 4. However, it should be noticed that while in the discrepancy analysis the seven covariates including one interaction term turned out to be significant, only four covariates of them were involved for developing the tree to produce the similar pseudo R^2 value. This might be because several interaction effects are automatically detected on the tree. For example, it was found that household size had more influences on the activity sequences of non-workers who do not go to K-12 school (e.g., homemaker) than those of K-12 students. As for workers, household size mostly influenced the daily activity patterns of younger workers (less than age 65), not older workers (over age 65).


Legend
home in the begining
home returning temporarily
home in the end
work
school
escort
eat out
household maintenance
personal business
social recreation
other
trip for activity
Root
n: 1000 s2: 137
Split: p_worker R2:0.0989
Global quality
Pseudo F: 38.721 ***
Pseudo R2: 0.1896 ***
Levene: 22.267 ***
<= 0
n: 423 s2: 133
Split: p_k12 R2:0.185
> 0
n: 577 s2: 117
plit: p_adult65over R2:0.012
<= 0
n: 262 s2: 135
Split: h_size R2:0.0167
> 0
n: 161 s2: 63.8
<= 0
n: 513 s2: 114
Split: h_size R2:0.00461
> 0
n: 64 s2: 131
<= 2
n: 130 s2: 124
> 2
n: 132 s2: 142
<= 2
n: 250 s2: 108
> 2
n: 263 s2: 119
Split: h_size R2:0.0111
<= 3
n: 103 s2: 136
> 3
n: 160 s2: 106

(a)



(b)

Figure 5.2 Induction Trees of Activity Sequences: (a) with activity sequence index plots and (b) with activity state distribution plots

In addition to the automatic detection of interaction effects and the provision of a comprehensive view of the sequence-covariate link, the induction tree yielded seven clusters at the terminal nodes. As shown in the tree with state distribution plots, it is possible to discover the differences among the seven clusters in the distribution of activity types at each five-minute time point. For example, see the terminal node indicating K-12 students. Most of them conducted the "school" activity (yellow) in the midday and their "trip" activity (black) was noticeably peaked at two time points. In addition, non-workers who are not a K-12 student were separated by household size. Those with small household size (h_size <= 2) spent more time at home (grey) than those with large household size (h_size > 2). Further, one cluster of older workers was discovered, which differs from the other three clusters of younger workers. Compared to younger workers, older workers spent less time on working (blue) and social recreation (green), and more time on personal business (violet) such as health care. Not surprisingly, the "trip" activity of older workers was spread over the midday without any clear peak time points.

5.4 **CONCLUSION**

Individuals' daily activity-travel patterns are complex due to interactions of numerous aspects embedded in them. An "old" question in activity-based travel behavior analysis is what explains the complexity of activity-travel patterns. To answer the question, this study proposed the discrepancy analysis of activity sequences. Viewing individual activity patterns as holistic and sequential objects, activity diaries were first converted into sequences of characters representing activity states. Then, the total discrepancy in the activity sequences was defined with a pairwise dissimilarity matrix between all sequences that was obtained from sequence alignment methods. Following the principle of ANOVA, the total sequence discrepancy was partitioned into explained among-groups discrepancy and residual within-groups discrepancy. This partition enabled to measure the strength of the association between activity sequences and covariates by calculating a pseudo R^2 and to assess the statistical significance of the association through the permutation tests of a pseudo F-ratio value. In addition to the sequence discrepancy analysis, this study developed an induction tree to help understand how individual activity sequences vary with the influential covariates.

Most of the existing applications of sequence alignment to activity-travel diary data have been restricted to calculating and classifying the dissimilarities of activity sequences, missing useful knowledge on activity sequences. It is expected that this research will allow us to explore the unknown area. In addition, this study would make a practical contribution to a micro-simulation framework for activity-based travel demand modeling. Some activity-based models assume a sequential scheduling process in which individuals decide what to do next at every time point. The approach is often criticized for the absence of any pre-planning process where activities are scheduled first on the basis of their priority and the schedule is implemented next. Gliebe and Kim (2010) respond to this criticism by assigning household roles to individuals before simulating them. This market segmentation can create a propensity for a certain type of behavior from which a wide range of activity-travel patterns may emerge. However, the market segmentation does not really capture the sequential decision process in advance. Discrepancy analysis with sequence alignment proposed in this paper may enable a more suitable segmentation for agents who are simulated through a sequential scheduling process.

This study certainly leaves room to be improved in several aspects. First of all, sequence alignment methods combined with discrepancy analysis in this paper can be further fine-tuned with different transformation cost settings for indels and substitutions. In addition, activity sequences can be represented in a different way by considering multiple activity attributes simultaneously, such as activity location, travel mode and time, the presence of persons, etc. Lastly, it is worthwhile to compare the cluster solutions discovered by building a tree with those obtained from a classical cluster analysis following sequence alignment methods. Thus, it may be possible to empirically verify whether or not the new methodological combination presented in this paper will truly overcome information loss caused by the previous popular cluster-based approach.

6.0 CONCLUSIONS

This paper presents the work of a data-integration project aimed to start addressing the challenges in data development for integrated land use and transportation modeling. As data sources increasingly update more frequently and even in real time, and policy decisions demand timely data and more rigorous analysis, data collection and updating increasingly become a continuous process. Traditional practice adopted by most planning agencies of updating datasets as a one-off effort at long intervals is very costly and increasingly challenged.

The project focuses on bringing interdisciplinary methods to make the best use of available data in the land use and transportation modeling domain. Utilizing statistics and machine-learning techniques, this project develops a continuous approach and reusable tools for data integration. We began by suggesting a data development process centered on a data warehouse (DW) for common data sets used in integrated modeling, in contrast to the commonly used ad-hoc approach. Accordingly, we propose a data development work flow that isolates generic components such as data-quality assurance and data imputation from region-specific data and system-specific requirements. This work flow involves ETL, data models, data-quality assurance, data imputation/synthesis, and finally data transformation to feed into applications.

These tools will be reusable for different time periods and for different regions. We tested and applied the approach and tools to our test metropolitan regions including Portland, OR, and the San Francisco Bay Area. The data structure, tools for data processing and quality monitoring, and documentation are available to the public on the project website https://cities.github.io/smartdata. These can be used by public agencies, researchers, or anyone else needing to develop a usable database for use in integrated planning and modeling.

With this project, we have nourished the goal of a data development process that can preserve historical as well as current data, and monitor data quality automatically and constantly as new data are being integrated on a continuous basis. Eventually we hope that both historical and current data are available on equal footing for modeling inputs, calibration and validation. Models using these data as inputs can seamlessly switch base year among years for which we

have data, essentially a process that helps alleviate some of the hardest challenges in integrated modeling.

We have started addressing some of the challenging problems, and some of solutions we came up with are very promising. However, probably not surprisingly, we were not able to solve all the problems within the scope of this project. A few related research topics we wish to explore in the future include data-synthesis approaches and their application in land use data. While population synthesizers have been widely used in both land use and transportation modeling, the application of data synthesis to land use data has not been fully explored. Although the MICE approach is versatile and produces reasonable results in our tests, there is vast number of alternative machine-learning approaches available and largely under-studied. An area of particular interest is to utilize the temporal dimension of the data for data imputation.

7.0 REFERENCES

Abayomi, K., Gelman, A., Levy, M., 2008. Diagnostics for multivariate imputations. Journal of the Royal Statistical Society: Series C (Applied Statistics) 57, 273–291. doi:10.1111/j.1467-9876.2007.00613.x

Abbott, A., Tsay, A., 2000. Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect. Sociological Methods & Research 29, 3–33.

Abraham, J., Andersen, K., Clay, M., Hunt, J., 2009. Calibrating a Synthetic Built Form Generator. Transportation Research Record: Journal of the Transportation Research Board 2133, 100–108. doi:10.3141/2133-11

Abraham, J., Weidner, T., Gliebe, J., Willison, C., Hunt, J., 2005. Three Methods for Synthesizing Base-Year Built Form for Integrated Land Use-Transport Models. Transportation Research Record: Journal of the Transportation Research Board 1902, 114–123. doi:10.3141/1902-14

Aisenbrey, S., Fasang, A.E., 2010. New Life for Old Ideas: The “Second Wave” of Sequence Analysis Bringing the “Course” Back Into the Life Course. Sociological Methods & Research 38, 420–462.

Anderson, M.J., 2001. A New Method for Non-Parametric Multivariate Analysis of Variance. Austral Ecology 26, 32–46.

Azur, M.J., Stuart, E.A., Frangakis, C., Leaf, P.J., 2011. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 20, 40–49. doi:10.1002/mpr.329

Bonetti, M., Piccarreta, R., Salford, G., 2013. Parametric and Nonparametric Analysis of Life Courses: An Application to Family Formation Patterns. Demography 50, 881–902.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees, 1 edition. ed. Chapman and Hall/CRC, New York, N.Y.

Burgette, L.F., Reiter, J.P., 2010. Multiple imputation for missing data via sequential regression trees. Am. J. Epidemiol. 172, 1070–1076. doi:10.1093/aje/kwq260

Burnett, K.P., Hanson, S., 1982. The Analysis of Travel as an Example of Complex Human Behavior in Spatially Constrained Situations: Definition and Measurement Issues. Transportation Research Part A 16, 87–102.
Buuren, S., Groothuis-Oudshoorn, K., 2011. MICE: Multivariate imputation by chained equations in R. Journal of statistical software 45.
Buuren, S. van, 2007. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res 16, 219–242. doi:10.1177/0962280206074463
Davidson, W., Donnelly, R., Vovsha, P., Freedman, J., Ruegg, S., Hicks, J., Castiglione, J., Picado, R., 2007. Synthesis of First Practices and Operational Research Approaches in Activity-Based Travel Demand Modeling. Transportation Research Part A 41, 464–488.
Devillers, R., Bédard, Y., Jeansoulin, R., Moulin, B., 2007. Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data. International Journal of Geographical Information Science 21, 261–282.
Donders, A.R.T., van der Heijden, G.J.M.G., Stijnen, T., Moons, K.G.M., 2006. Review: A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology 59, 1087–1091. doi:10.1016/j.jclinepi.2006.01.014
Doove, L.L., Van Buuren, S., Dusseldorp, E., 2014. Recursive partitioning for missing data imputation in the presence of interaction effects. Computational Statistics & Data Analysis 72, 92–104. doi:10.1016/j.csda.2013.10.025
Eskelson, B.N.I., Temesgen, H., Lemay, V., Barrett, T.M., Crookston, N.L., Hudak, A.T., 2009. The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. Scandinavian Journal of Forest Research 24, 235–246. doi:10.1080/02827580902870490
Ettema, D., 1996. Activity-Based Travel Demand Modeling. Technische Universiteit Eindhoven, Netherlands.
Farooq, B., Bierlaire, M., Hurtubia, R., Flötteröd, G., 2013. Simulation based Synthesis of Population (No. TRANSP-OR *130125*). Transport and Mobility Laboratory, EPFL.
Gabadinho, A., Ritschard, G., Muller, N.S., Studer, M., 2011. Analyzing and Visualizing State Sequences in R with TraMineR. Journal of Statistical Software 40, 1–37.
Gliebe, J., Kim, K., 2010. Time-Dependent Utility in Activity and Travel Choice Behavior. Transportation Research Record: Journal of the Transportation Research Board 2156, 9–16.
Gregor, B., 2012. GreenSTEP Model Overview. Oregon DOT Transportation Planning Analysis Unit, Salem, OR.
Hsu, D., 2014. Improving energy benchmarking with self-reported data. Building Research & Information 42, 641–656. doi:10.1080/09613218.2014.887612
Hu, Y., De, S., Chen, Y., Kambhampati, S., 2012. Bayesian Data Cleaning for Web Data. arXiv:1204.3677.
Joh, C., Arentze, T., Hofman, F., Timmermans, H., 2002. Activity Pattern Similarity: A Multidimensional Sequence Alignment Method. Transportation Research Part B 36, 385–403.
Joh, C., Arentze, T., Timmermans, H., 2001. Pattern Recognition in Complex Activity Travel Patterns: Comparison of Euclidean Distance, Signal-Processing Theoretical, and Multidimensional Sequence Alignment Methods. Transportation Research Record: Journal of the Transportation Research Board 1752, 16–22.

Knuth, D.E., 1984. Literate Programming. The Computer Journal 27, 97–111. doi:10.1093/comjnl/27.2.97
Koppelman, F.S., Pas, E.I., 1984. Travel-Activity Behavior in Time and Space: Methods for Representation and Analysis, in: P. Nijkamp, H. Leitner, Wrigley, N. (Eds.), Measuring the Unmeasurable. Kluwer Academic Pub.
Lakshminarayan, K., Harp, S.A., Samad, T., 1999. Imputation of Missing Data in Industrial Databases. Applied Intelligence 11, 259–275. doi:10.1023/A:1008334909089
Lesnard, L., 2010. Setting Cost in Optimal Matching to Uncover Contemporaneous Socio-Temporal Patterns. Sociological Methods & Research 38, 389–419.
Levinson, D., 2004. Metropolitan Travel Survey Archive: Final Report to Bureau of Transportation Statistics. Department of Civil Engineering, University of Minnesota.
Levinson, D., Deshpande, V., 2003. Metropolitan Travel Survey Archive. University of Minnesota.
Levinson, D., Deshpande, V., 2006. Metropolitan Travel Survey Archive Phase 2 - Final Report. University of Minnesota, Department of Civil Engineering, Minneapolis, MN.
Martin, P., Wiggins, R., 2011. Optimal Matching Analysis, in: Williams, M., Vogt, P. (Eds.), The SAGE Handbook of Innovation in Social Research Methods. SAGE Publications Ltd, pp. 385–408.
McArdle, B.H., Anderson, M.J., 2001. Fitting Multivariate Models to Community Data: A Comment on Distance-Based Redundancy Analysis. Ecology 82, 290–297.
Metro, 2009. Metroscope 3.0 Documentation. Metro, Portland, OR.
Müller, K., Axhausen, K.W., 2011. Population Synthesis for Microsimulation: State of the Art, in: Transportation Research Board 90th Annual Meeting.
Obe, R., Hsu, L., 2011. PostGIS in Action. Manning Publications.
Pas, E.I., 1983. A Flexible and Integrated Methodology for Analytical Classification of Daily Travel-Activity Behavior. Transportation Science 17, 405–429.
Pinjari, A.R., Bhat, C.R., 2011. Activity-Based Travel Demand Analysis, in: A. de Palma, R. Lindsey, E. Quinet, Vickerman, R. (Eds.), A Handbook of Transport Economics. Edward Elgar Pub, pp. 213–248.
Pipino, L.L., Lee, Y.W., Wang, R.Y., 2002. Data quality assessment. Commun. ACM 45, 211–218. doi:10.1145/505248.506010
Recker, W.W., McNally, M.G., Root, G.S., 1985. Travel/Activity Analysis: Pattern Recognition, Classification, and Interpretation. Transportation Research Part A 19, 279–296.
Rubin, D.B., 2004. Multiple Imputation for Nonresponse in Surveys. Wiley-Interscience.
Ruggles, S., ALEXANDER, T., Genadek, K., Goeken, R., Schroeder, M., Sobek, M., 2010. Integrated Public Use Microdata Series (IPUMS): Version 5.0 [Machine-readable database]. University of Minnesota, Minneapolis, MN.
Sammour, G., Bellemans, T., Vanhoof, K., Janssens, D., Kochan, B., Wets, G., 2012. The Usefulness of the Sequence Alignment Methods in Validating Rule-Based Activity-Based Forecasting Models. Transportation 39, 773–789.
Saneinejad, S., Roorda, M.J., 2009. Application of Sequence Alignment Methods in Clustering and Analysis of Routine Weekly Activity Schedules. Transportation Letters 1, 97–211.
Schafer, J., 1997. Analysis of incomplete multivariate data. Chapman & Hall, London; New York.
Schafer, J.L., Graham, J.W., 2002. Missing data: Our view of the state of the art. Psychological Methods 7, 147–177. doi:10.1037/1082-989X.7.2.147

Schlich, R., Axhausen, K.W., 2003. Habitual Travel Behaviour: Evidence from a Six-Week Travel Diary. Transportation 20, 13–36.
Shah, A.D., Bartlett, J.W., Carpenter, J., Nicholas, O., Hemingway, H., 2014. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. Am. J. Epidemiol. 179, 764–774. doi:10.1093/aje/kwt312
Shoval, N., Isaacson, M., 2007. Sequence Alignment as a Method for Human Activity Analysis in Space and Time. Annals of the Association of American Geographers 97, 282–297.
Singleton, P.A., 2013. Data Cleaning in Activity and Travel Surveys: Methodology Applied to Walk Trips. Presented at the Transportation Research Board 92nd Annual Meeting.
Studer, M., Ritschard, G., Gabadinho, A., Muller, N.S., 2011. Discrepancy Analysis of State Sequences. Sociological Methods & Research 40, 471–510.
ULTRANS ITS UC Davis, HBA Specto Incorporated, 2011. California Statewide Travel Demand Model. Institute of Transport Studies, University of California. Davis, Davis, CA.
Waddell, P., 2009. Parcel-Level Microsimulation of Land Use and Transportation: The Walking Scale of Urban Sustainability, in: Proceedings of the 2009 IATBR Workshop on Computational Algorithms and Procedures for Integrated Microsimulation Models. Tempe, AZ.
Waddell, P., Caballero, P., Duisberg, R., Kim, H., Peak, C., Socha, D., Wang, L., 2005. UrbanSim:Model Estimation for the Puget Sound Region (No. CUSPA-04-02). Center for Urban Simulation and Policy Analysis.
Waddell, P., Wang, L., Charlton, B., Olsen, A., 2010. Microsimulating parcel-level land use and activity-based travel: Development of a prototype application in San Francisco. Journal of Transport and Land Use 3, 65–84.
Wang, L., Waddell, P., Outwater, M.L., 2011. Incremental Integration of Land Use and Activity-Based Travel Modeling. Transportation Research Record: Journal of the Transportation Research Board 2255, 1–10.
Wegener, M., 2011. From Macro to Micro—How Much Micro is too Much? Transport Reviews 31, 161–177.
Weidner, T., Abraham, J.E., Hunt, J.D., Gregor, B.J., 2007. Floorspace supply/demand calibration in two land-use transport models, in: 11th World Conference on Transport Research.
Weidner, T., Knudson, B., Picado, R., Hunt, J., 2009. Sensitivity Testing with the Oregon Statewide Integrated Model. Transportation Research Record: Journal of the Transportation Research Board 2133, 109–122. doi:10.3141/2133-12
White, I.R., Royston, P., Wood, A.M., 2011. Multiple imputation using chained equations: Issues and guidance for practice. Statist. Med. 30, 377–399. doi:10.1002/sim.4067
Wilson, C., 1998. Analysis of Travel Behavior Using Sequence Alignment Methods. Transportation Research Record: Journal of the Transportation Research Board 1645, 52–59.
Wilson, C., 2001. Activity Patterns of Canadian Women: Application of ClustalG Sequence Alignment Software. Transportation Research Record: Journal of the Transportation Research Board 1777, 55–67.
Wilson, C., 2008. Activity Patterns in Space and Time: Calculating Representative Hagerstrand Trajectories. Transportation 35, 485–499.

Yang, X., Blower, J.D., Bastin, L., Lush, V., Zabala, A., Masó, J., Cornford, D., Díaz, P., Lumsden, J., 2013. An integrated view of data quality in Earth observation. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371.

8.0 APPENDICES

APPENDIX A-1 R AND SHELL SCRIPTS FOR ETL

Extract

```
# extracts data from dbf files
require('foreign')
data.df  <- read.dbf('/workspace/sanfrancisco/shapefiles/zone.dbf')

# extracts data from csv files
data.df <- read.csv("/workspace/sanfrancisco/parcels.txt", sep=",", header=T)

#extract data from SAS export file
require('foreign')
data.df <- read.xport('/workspace/sanfrancisco/pub_comp_house1f_with_xy.xpt')

#extract data from DBMS (e.g. MySQL)
require(RMySQL)
conn <- dbConnect(MySQL(), group='db_server', dbname='sanfrancisco_baseyear')
data.df  <- dbReadTable(conn, 'zones', row.names=NULL)

#Extract fixed width format (Bay Area Travel Survey data, available at
ftp://ftp.abag.ca.gov/pub/mtc/planning/BATS/BATS2000)
fname = "/workspace/MTC/BATS2000_Public_Release_v3.0/vehicle.dat"
data.df<-read.fwf(fname, widths=c(6,2,3,5,3,5,7,7,7,2,2,6,2,2,2,7,2,5,3,5,3),
col.names=c("HHID","HHVEHNUM","MAKMODEL","VEHYEAR","MONPOSS","YEARPO
SS","MILEPOSS","MILEDAY1","MILEDAY2","MAIL","PROXY","HOMEZIP","SAMPTYP
","STATUS","HHVEHICLES","C_TRACT","C_BLKGRP","C_PUMA","R_SD","R_TAZ1454
","R_COUNTY_FIPS"), comment.char="")

#extract data from Excel spreadsheets
require(gdata)
data.df  <- read.xls("Puma2000.xls", sheet=1)

#read data directly from URL
require(RCurl)
```

```
urlfile = getURL(“/url/to/csv/file”)
data.df <- read.csv(textConnection(urlfile))

#Load GIS data into PostgreSQL/PostGIS in a linux shell with PostGIS installed
# Example - Metro UGB shapefile; 2913 is the Spatial Reference Identifier for NAD83(HARN)
shp2pgsql -d -I -s 2913 /workspace/RLIS/ugb.shp | psql -h postgres_server -U
postgres_username -d portland

# load all shapefile in the current directory
for f in *.shp; do shp2pgsql -d -I -s 2913 $f | psql -h postgres_server -U postgres_username -d
portland; done
```

Transform

```
# merge two data tables
data.df.extra <- merge(data.df, df.extra, by.x='TAZ', by.y='zone_id')

#append rows
```

Load (to a PostgreSQL database)

```
# Load R data.frame in data.df to table “table.name” in PostgreSQL database bayarea
require(RPostgreSQL)
conn <- dbConnect(PostgreSQL(), user= "postgres", password="***", dbname="bayarea")
dbListTables(conn)
if (dbExistsTable(conn, table.name)) dbRemoveTable(conn, table.name)
dbWriteTable(conn, table.name, data.df, row.names = F)
```

APPENDIX A-2 DATA-QUALITY INDICATORS

```
# Example for single and multiple family housing
# comparing ACS and RLIS

setwd("/workspace/DQI")

library(ggplot2)
library(RPostgreSQL)
library(plyr)
library(data.table)
library(reshape2)

############################
# Data Quality Indicators #
############################

# functions to calculate several data quality indicators
count.na <- function(x) sum(is.na(x))
all.na <- function(x, y) max(count.na(x), count.na(y))==length(y)
rmse <- function(x=x, y=y) sqrt(mean((x-y)^2))
r2 <- function(x=x, y=y, ...) {if(!all.na(x, y)) summary(lm(y~x,...))$r.squared else 0.0}
intercept <- function(x=x, y=y, ...) {if(!all.na(x,y)) coef(lm(y~x,...))["(Intercept)"] else 0.0}
slope <- function(x=x, y=y, ...) {if(!all.na(x,y)) coef(lm(y~x,...))["x"] else 0.0}

# create a table for data quality indicators
all.dt <- data.table(all, key = "year")
dqi <- ddply(all.dt, .(year,unit_type), summarise,
                    rmse=round(rmse(x=num_units.rlis, y=num_units.acs),2),
                    r2=round(r2(x=num_units.rlis, y=num_units.acs),2),
                    intercept=round(intercept(x=num_units.rlis, y=num_units.acs),2),
                    slope=round(slope(x=num_units.rlis, y=num_units.acs),2))
dqi <- dqi[order(rev(dqi$year),dqi$unit_type),]
dqi[dqi==0] <- NA
dqi$year <- as.numeric(dqi$year)

# create graphs
rmse <- ggplot(data=dqi,aes(x=year,y=rmse))+
                    geom_point(aes(shape=unit_type),size=5)+
                    theme(legend.position="none")+
                    xlim(2007,2011)
r2 <- ggplot(data=dqi,aes(x=year,y=r2))+
                    geom_point(aes(shape=unit_type),size=5)+
                    theme(legend.position="none")+
                    xlim(2007,2011)
intercept <- ggplot(data=dqi,aes(x=year,y=intercept))+
```

```
                    geom_point(aes(shape=unit_type),size=5)+
                    theme(legend.position="bottom")+
                    xlim(2007,2011)
slope <- ggplot(data=dqi,aes(x=year,y=slope))+
                    geom_point(aes(shape=unit_type),size=5)+
                    theme(legend.position="none")+
                    xlim(2007,2011)

png(file="dqi_housing.png")
multiplot(rmse,slope,r2,intercept,cols=2) # the "multiplot" function must be generated first
dev.off()

# Multiple plot function
#
# ggplot objects can be passed in ..., or to plotlist (as a list of ggplot objects)
# - cols:   Number of columns in layout
# - layout: A matrix specifying the layout. If present, 'cols' is ignored.
#
# If the layout is something like matrix(c(1,2,3,3), nrow=2, byrow=TRUE),
# then plot 1 will go in the upper left, 2 will go in the upper right, and
# 3 will go all the way across the bottom.
#
multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {
  require(grid)

  # Make a list from the ... arguments and plotlist
  plots <- c(list(...), plotlist)

  numPlots = length(plots)

  # If layout is NULL, then use 'cols' to determine layout
  if (is.null(layout)) {
    # Make the panel
    # ncol: Number of columns of plots
    # nrow: Number of rows needed, calculated from # of cols
    layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                     ncol = cols, nrow = ceiling(numPlots/cols))
  }

 if (numPlots==1) {
    print(plots[[1]])

  } else {
    # Set up the page
    grid.newpage()
    pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))
```

```
    # Make each plot, in the correct location
    for (i in 1:numPlots) {
      # Get the i,j matrix positions of the regions that contain this subplot
      matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))

      print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
                                      layout.pos.col = matchidx$col))
    }
  }
}
```

APPENDIX A-3 DATA IMPUTATION WITH MICE

```
setwd("/workspace/data_imputation")

library(RPostgreSQL)# to connect the PostgreSQL database
library(mice)
rm(list=ls())

### connect to the PostgreSQL database
conn <- dbConnect(PostgreSQL(), host="postgres_server", user="postgres_user",
password="postgres_passwd", dbname="portland")

### read taxlots20XX

taxlots2011 <- dbReadTable(conn, c("rlis","taxlots2011"))

### create a data set for imputation

ds <- taxlots2011[c("area","bldgsqft","landval","bldgval","landuse","juris_city")]

SFR <- subset(ds, landuse=='SFR')                    # select tax lots with SFR

SFR$bldgsqft[which(SFR$bldgsqft==0)] <- NA           # recode bldgsqft with zero to NA

SFR$landval[which(SFR$landval==0)] <- NA             # recode landval with zero to NA

SFR$bldgval[which(SFR$bldgval==0)] <- NA             # recode bldgval with zero to NA

SFR$juris_city <- as.factor(SFR$juris_city)          # convert data type of juris_city from
character to factor

SFR$far <- SFR$bldgsqft / SFR$area                   # floor area ratio

SFR$lvpa <- SFR$landval / SFR$area                   # land value per area

SFR$bvpa <- SFR$bldgval / SFR$area                   # building value per area


SFR <- SFR[c("area","juris_city","far","lvpa","bvpa")]

### validation

# create a validation data set

# in which we draw a 5% random sample for each variable and replace them with NA

SFR.val <- subset(SFR, !is.na(far) & !is.na(lvpa) & !is.na(bvpa))
```

```
set.seed(1234)
SFR.val$far[sample(nrow(SFR.val),0.05*nrow(SFR.val))] <- NA
SFR.val$lvpa[sample(nrow(SFR.val),0.05*nrow(SFR.val))] <- NA
SFR.val$bvpa[sample(nrow(SFR.val),0.05*nrow(SFR.val))] <- NA
# inspect missing values across the variables from the validation data set
md.pattern(SFR.val)

# conduct MICE via CART using the validation data set
system.time(imp.cart.SFR.val <- mice(SFR.val, seed=1234, method="cart", minbucket=5))
print(imp.cart.SFR.val)

SFR$rowname <- as.numeric(row.names(SFR))

imp.far <- imp.cart.SFR1$imp$far
imp.far$rowname <- as.numeric(row.names(imp.far))
imp.far <- merge(imp.far, SFR, by="rowname")

imp.lvpa <- imp.cart.SFR1$imp$lvpa
imp.lvpa$rowname <- as.numeric(row.names(imp.lvpa))
imp.lvpa <- merge(imp.lvpa, SFR, by="rowname")

imp.bvpa <- imp.cart.SFR1$imp$bvpa
imp.bvpa$rowname <- as.numeric(row.names(imp.bvpa))
imp.bvpa <- merge(imp.bvpa, SFR, by="rowname")
```

```
rmse.1 <- sqrt(mean((imp.far$'1'-imp.far$far)^2))

rmse.2 <- sqrt(mean((imp.far$'2'-imp.far$far)^2))

rmse.3 <- sqrt(mean((imp.far$'3'-imp.far$far)^2))

rmse.4 <- sqrt(mean((imp.far$'4'-imp.far$far)^2))

rmse.5 <- sqrt(mean((imp.far$'5'-imp.far$far)^2))

r2.1 <- summary(lm(imp.far$'1' ~ imp.far$far))$r.squared

r2.2 <- summary(lm(imp.far$'2' ~ imp.far$far))$r.squared

r2.3 <- summary(lm(imp.far$'3' ~ imp.far$far))$r.squared

r2.4 <- summary(lm(imp.far$'4' ~ imp.far$far))$r.squared

r2.5 <- summary(lm(imp.far$'5' ~ imp.far$far))$r.squared


table.far <- matrix(c(rmse.1, rmse.2, rmse.3, rmse.4, rmse.5, r2.1, r2.2, r2.3, r2.4, r2.5), 2, 5,
byrow=TRUE)

dimnames(table.far) = list(c("rmse", "r2"), c("mi_1", "mi_2", "mi_3", "mi_4", "mi_5"))

table.far


rmse.1 <- sqrt(mean((imp.lvpa$'1'-imp.lvpa$lvpa)^2))

rmse.2 <- sqrt(mean((imp.lvpa$'2'-imp.lvpa$lvpa)^2))

rmse.3 <- sqrt(mean((imp.lvpa$'3'-imp.lvpa$lvpa)^2))

rmse.4 <- sqrt(mean((imp.lvpa$'4'-imp.lvpa$lvpa)^2))

rmse.5 <- sqrt(mean((imp.lvpa$'5'-imp.lvpa$lvpa)^2))

r2.1 <- summary(lm(imp.lvpa$'1' ~ imp.lvpa$lvpa))$r.squared

r2.2 <- summary(lm(imp.lvpa$'2' ~ imp.lvpa$lvpa))$r.squared

r2.3 <- summary(lm(imp.lvpa$'3' ~ imp.lvpa$lvpa))$r.squared

r2.4 <- summary(lm(imp.lvpa$'4' ~ imp.lvpa$lvpa))$r.squared

r2.5 <- summary(lm(imp.lvpa$'5' ~ imp.lvpa$lvpa))$r.squared
```

```
table.lvpa <- matrix(c(rmse.1, rmse.2, rmse.3, rmse.4, rmse.5, r2.1, r2.2, r2.3, r2.4, r2.5), 2, 5, byrow=TRUE)

dimnames(table.lvpa) = list(c("rmse", "r2"), c("mi_1", "mi_2", "mi_3", "mi_4", "mi_5"))

table.lvpa


rmse.1 <- sqrt(mean((imp.bvpa$'1'-imp.bvpa$bvpa)^2))

rmse.2 <- sqrt(mean((imp.bvpa$'2'-imp.bvpa$bvpa)^2))

rmse.3 <- sqrt(mean((imp.bvpa$'3'-imp.bvpa$bvpa)^2))

rmse.4 <- sqrt(mean((imp.bvpa$'4'-imp.bvpa$bvpa)^2))

rmse.5 <- sqrt(mean((imp.bvpa$'5'-imp.bvpa$bvpa)^2))

r2.1 <- summary(lm(imp.bvpa$'1' ~ imp.bvpa$bvpa))$r.squared

r2.2 <- summary(lm(imp.bvpa$'2' ~ imp.bvpa$bvpa))$r.squared

r2.3 <- summary(lm(imp.bvpa$'3' ~ imp.bvpa$bvpa))$r.squared

r2.4 <- summary(lm(imp.bvpa$'4' ~ imp.bvpa$bvpa))$r.squared

r2.5 <- summary(lm(imp.bvpa$'5' ~ imp.bvpa$bvpa))$r.squared


table.bvpa <- matrix(c(rmse.1, rmse.2, rmse.3, rmse.4, rmse.5, r2.1, r2.2, r2.3, r2.4, r2.5), 2, 5, byrow=TRUE)

dimnames(table.bvpa) = list(c("rmse", "r2"), c("mi_1", "mi_2", "mi_3", "mi_4", "mi_5"))

table.bvpa


prmse1 <- rmse1 / nrow(imp.far)

prmse2 <- rmse2 / nrow(imp.far)

prmse3 <- rmse3 / nrow(imp.far)

prmse4 <- rmse4 / nrow(imp.far)
```

```
prmse5 <- rmse5 / nrow(imp.far)

imp.lvpa <- imp.cart.SFR1$imp$lvpa

imp.lvpa$rowname <- row.names(imp.far)

imp.bvpa <- imp.cart.SFR1$imp$bvpa

imp.far$rowname <- row.names(imp.far)

system.time(imp.pmm <- mice(SFR, seed=1234))

# stripplot

png("stripplot.pmm.png"); stripplot(imp.pmm, pch=20, cex=1.2); dev.off()

png("stripplot.cart.png"); stripplot(imp.cart, pch=20, cex=1.2); dev.off()

# xyplot

png("xyplot.pmm1.png"); xyplot(imp.pmm, bldgsqft~landval | .imp, pch=20, cex=1.4); dev.off()

png("xyplot.pmm2.png"); xyplot(imp.pmm, bldgsqft~bldgval | .imp, pch=20, cex=1.4); dev.off()

png("xyplot.pmm3.png"); xyplot(imp.pmm, bldgsqft~juris_city | .imp, pch=20, cex=1.4);
dev.off()

png("xyplot.cart1.png"); xyplot(imp.cart, bldgsqft~landval | .imp, pch=20, cex=1.4); dev.off()

png("xyplot.cart2.png"); xyplot(imp.cart, bldgsqft~bldgval | .imp, pch=20, cex=1.4); dev.off()

png("xyplot.cart3.png"); xyplot(imp.cart, bldgsqft~juris_city | .imp, pch=20, cex=1.4); dev.off()

png("xyplot.cart1.png"); xyplot(imp.cart, far~lvpa | .imp, pch=20, cex=1.4); dev.off()

png("xyplot.cart2.png"); xyplot(imp.cart, far~bvpa | .imp, pch=20, cex=1.4); dev.off()

png("xyplot.cart3.png"); xyplot(imp.cart, far~juris_city | .imp, pch=20, cex=1.4); dev.off()
```

```
# densityplot
png("densityplot.pmm.bldgsqft.png"); densityplot(imp.pmm, ~bldgsqft | .imp); dev.off()
png("densityplot.pmm.landval.png"); densityplot(imp.pmm, ~landval | .imp); dev.off()
png("densityplot.pmm.bldgval.png"); densityplot(imp.pmm, ~bldgval | .imp); dev.off()

png("densityplot.cart.bldgsqft.png"); densityplot(imp.cart, ~bldgsqft | .imp); dev.off()
png("densityplot.cart.landval.png"); densityplot(imp.cart, ~landval | .imp); dev.off()
png("densityplot.cart.bldgval.png"); densityplot(imp.cart, ~bldgval | .imp); dev.off()

png("densityplot.cart.far.png"); densityplot(imp.cart, ~far | .imp); dev.off()
png("densityplot.cart.lvpa.png"); densityplot(imp.cart, ~lvpa | .imp); dev.off()
png("densityplot.cart.bvpa.png"); densityplot(imp.cart, ~bvpa | .imp); dev.off()

# after MICE
fit <- with(imp.cart, lm(far~lvpa+bvpa))
summary(fit)
est <- pool(fit)
est
```