DATA FUSION: TECHNIQUES AND APPLICATIONS

2017-11-22

Transportation Tomorrow Survey 2.0

Daming Gao, Siva Srikukenthiran

Khandker Nurul Habib & Eric J. Miller

UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE & ENGINEERING
Transportation Research Institute

EXECUTIVE SUMMARY

Travel survey data have been a primary source for planning and simulation purposes; however, in recent years, researchers have begun considering alternative data sources to complement these survey data. For example, passively collected data such as GPS, smartphone, and transit smart-cards have been an increasing focus of travel location research. They offer a continuous stream of travel information of a large population at a relatively inexpensive cost – these advantages can effectively address many of the shortcomings of the travel survey data and provide possibilities of fusion between the two. Goulias et al. (2013) proposed a sustainable framework for integration of multiple surveys and datasets. This framework is called core-satellite - the core is the current travel survey, while the satellites can be a series of specific surveys to fill in the gaps and add details to the core through common variables. Additionally, complementary datasets such as passive data could be used to enrich and augment the core-satellite structure. The method of linking various datasets is the motivation behind the need to research data fusion techniques.

The objective of this report is to provide an overview of common data fusion techniques and real-world applications. It aims to identify gaps between theory and practice, and provide a sense of direction for future research. This report first reviews the context and definition of data fusion. Out of two branches of research - sensor fusion and statistical matching of survey records, this report focusses on the latter. Next, data harmonization issues such as spatial, temporal and semantic incompatibilities are thoroughly discussed, along with a list of possible treatments. After resolving these issues, mathematical procedures on choosing common variables are outlined. Moreover, common fusion techniques in the statistics community, such as the nearest-neighbor, record linkage, parameter estimation, and imputation, are expanded to the key procedural level. This report also briefly discusses data integration techniques in transportation research: population synthesis, RP/SP survey integration, and pseudo-panels. Also discussed are procedure that can be used to test the validity of the fusion process through the use of indicators, such as the marginal distribution of variables and correlation structures. The remainder of the report presents case studies showing some attempts that have been made to fuse data in the context of travel surveys, and identifies datasets that can be fused in preparation for the next TTS in 2021.

Based on this review of techniques, while data fusion has the potential to enrich travel survey data, the actual implementation can be very challenging, given the rarity of real-life applications in the literature. The largest barrier is the incompatibility of datasets, as most datasets are not designed for integration. Methods to resolve data incompatibility are difficult to generalize as each dataset is unique and should be treated on a case-by-case basis. Another barrier is a lack of validity measures – current methods offer insights on variable distributions and correlations, but finer levels of measurements do not yet exist due to mathematical constraints. As for the fusion techniques, many are rigorously discussed at the theoretical level, but few are put into actual use. Partially caused by the complexity and incompatibility of real-life datasets, there is no specific statement on which method should be recommended under certain situations. As a result, there is a need to conduct further research on evaluating the feasibility and applicability of each fusion method.

Due to the above reasons, real-life examples reviewed in this report are not strictly data fusion. However, they are still valuable as they offer interesting results and possibilities on how multiple datasets can be utilized together. Some studies, as illustrated by Spurr et al. (2015) on directly matching smart-card records with travel survey smart card holders, and Grapperon et al. (2016) on enriching smart-card information with census, are the ones that best approximate data fusion processes and deserve careful attention.

Data Fusion: Techniques and Applications

TRANSPORTATION TOMORROW SURVEY 2.0

Table of Contents

Executive Summary 1

1 Introduction 4

2 Data Fusion Perspectives and Definition 6

3 Fusion Prerequisites 7
- **3.1 Data Harmonization** 7
- **3.2 Choice of Matching Variables** 8

4 Common Fusion Techniques 10
- **4.1 Classifications** 10
- **4.2 Non-parametric Approach** 11
  - 4.2.1 Nearest-neighbor Approach 11
  - 4.2.2 Exact Matching 14
- **4.3 Parametric Approach** 15
- **4.4 Mixed Approach** 16
- **4.5 Imputation** 16

5 Techniques Relevant TO Data Fusion 18
- **5.1 Population Synthesis** 18
- **5.2 RP/SP Survey Fusion** 18
- **5.3 Pseudo-Panel** 19

6 Data Fusion Validity 20

7 Case Studies 21
- **7.1 Survey Integration** 21
- **7.2 Data Enrichment** 22
- **7.3 Data Matching** 22
- **7.4 Pattern Identification** 24

8 Hypothetical Cases for TTS 2.0 25
- **8.1 Household Income Imputation** 25
- **8.2 PRESTO Card Data Analysis** 25
- **8.3 Time-use Survey** 26

8.4 Cellint Cellular Data 26
8.5 StudentMoveTO Data 26
9 Concluding Remarks 27
10 Bibliography 28

1 INTRODUCTION

Travel survey data have been a primary source for planning and simulation purposes; however, in recent years, with advances in technology, researchers have begun considering other data sources to complement this survey data. Passively collected datasets such as those produced by cellphone towers, GPS, smartphone devices, and from the use of transit smart-cards have become an increasing focus of travel location research. Such data are continuously collected, can be easily made available, are relatively inexpensive to obtain, and provide continuous detail on mobility behaviors for a large population. However, the accuracy of passive data is not always reliable and socio-economic information is generally absent. As a result, there is a need to look to census data, points of interest (POI), and land use information to enrich these datasets to make them viable for the replacement of or fusion with traditional survey data.

This need to look for alternative data sources has been driven by several limitations in travel survey data. First, travel survey data are generally cross-sectional in nature. Respondents are usually asked to recall their trips on a given day or a few consecutive days, their answers cannot adequately capture the long-term temporal variations in travel demand patterns. Second, survey data are not collected on a frequent basis due to their high cost. For example, the Transportation Tomorrow Survey (TTS) is conducted only once in every 5 years in the Greater Golden Horseshoe (GGH) area (Ministry of Transportation Ontario, 2016). Even though this is on the high-end of the frequency of large-scale surveys, collected data can become obsolete as years pass before the next survey; this is especially true if there is rapid development during the period which changes overall travel behavior. Finally, the collected data have been gradually losing their representativeness of the real population, because of dropping response rates of certain demographics and varying demographics of survey modes. For example, younger populations have become increasingly underreported in traditional landline surveys, while having a disproportionate representation in smartphone-based surveys.

With many potential datasets that could be used to address these limitations, it is crucial to organize them in a way that is sustainable, generalizable, and expandable. Goulias et al. (2013) provide a core-satellite paradigm to achieve these objectives. The core can be a large-sample travel survey including socio-economic variables and travel behaviors (i.e. the current traditional travel survey). On the other hand, satellites can be a series of smaller, but more specific surveys to fill in gaps and add details to the core. All data produced by the satellite should be linkable to the core through common variables, so more datasets can be later added. Complementary datasets such as passive data could be used to enrich and augment the core-satellite structure (Miller, et al., 2014). The TTS 2.0 project is considering such a structure for the next TTS in 2021.

In considering the core-satellite structure, the method of linking together the resulting varying datasets lead to the need to research data fusion techniques. In the statistics community, data fusion techniques have been developed and discussed since the 1960s (Rässler, 2002), and many techniques exist that are used by market research firms and statistics agencies. However, in the transportation survey field, there is limited research, focusing mainly on conceptual methods, with actual implementations rare. This could be attributed to the complexity of travel data, and the lack of context or exact purposes for data fusion. Moreover, the statistical properties of the fused dataset remain difficult to quantify (Bayard, et al., 2009). Furthermore, it is only in recent years that various data sources have become available to the transportation community. The combination of these factors not only presents a challenge but also provides an opportunity for future research to investigate ways to combine datasets from varying sources.

This report presents an overview of common data fusion techniques and prior real-world applications. It aims to identify gaps between theory and practice, and provide a sense of direction for future research. **Section 2** contains a detailed definition of data fusion from multiple perspectives. **Section 3** describes the preparation steps required before attempting fusion. **Section 4** classifies traditional fusion techniques and provides overviews and examples for each. **Section 5** identifies techniques related to fusion. **Section 6** introduces measures to validate output produced by data fusion. **Section 7** categorizes and compares some of the transportation case studies.

Section 8 consists of potential scenarios of data fusion for the TTS. Finally, **Section 9** concludes the report and provides research recommendations.

2 DATA FUSION PERSPECTIVES AND DEFINITION

The definition of data fusion depends on its context. In marketing research, it is generally associated with "statistical matching", where records from different consumer surveys are matched based on common variables amongst the data files. This process is also often referred to as data integration. The techniques were developed in Europe during the 1960s when media planners were interested in the relationship between people's television viewing and purchasing behaviors to improve media targeting. Since collecting all the information in a single survey could be prohibitively expensive and difficult, and respondents would also find the questionnaire too long to complete, researchers developed techniques to match individual records from existing surveys. Later in the 1970s, federal agencies in the USA and Canada employed similar techniques to match data from different surveys to obtain more comprehensive household income information. (Rässler, 2002)

Another frequently mentioned concept is multi-sensor data fusion, where observations from different sensors are combined to better describe the state of the environment of interest. It is commonly used in intelligent transportation system (ITS) research. Fusion in this context requires definitions of sensor configurations and fusion system architectures before applying specific mathematical procedures (Bachmann, 2011). For example, Bachmann et al. (2013) showed that fusing Bluetooth and loop detector data could yield a higher accuracy of traffic speed estimates than using loop detectors alone. Since travel surveys produce very different data types compared to sensor data, this report focuses predominantly on the "matching" context of data fusion.

As a result, fusion in the transportation survey data context involves the integration of files from different sources (Saporta, 2002), and typically follows a donor-receptor framework. As illustrated in **Figure 1**, a **donor file** usually has a larger size than a **receptor file**, it contains information that receptor file does not have but preferably should have. The receptor file can also have variables that are not present in the donor file (Figure 1, right). Variables unique to either file are called **specific variables**, while variables exist in both files are called **common variables**. In general, common variables best describe the "identity" of each data set. They can be age, gender, or unique ID.

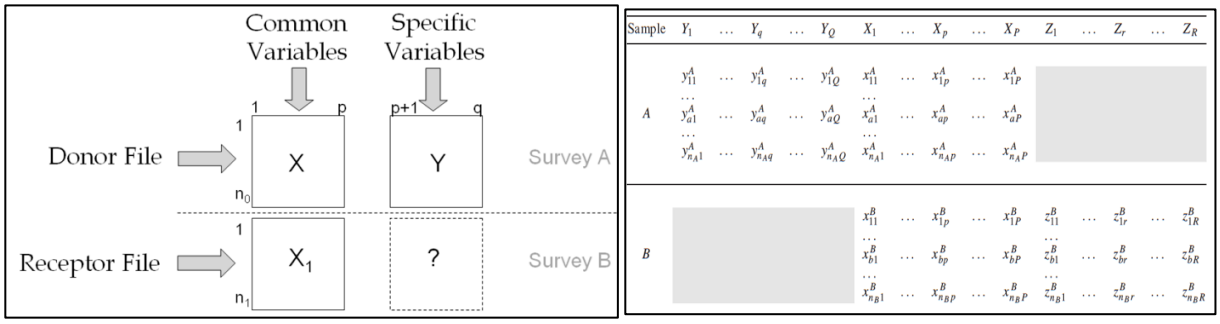


FIGURE 1: DATA FUSION FRAMEWORK. LEFT: (D'AMBROSIO, ET AL., 2007), RIGHT: (D'ORAZIO, ET AL., 2006)

There are two ways that matching can proceed. If common variables can be matched with certainty (e.g. identical ID, zone, or postal code), then **exact matching** or **record linkage** occurs (Section 4.2.2). If common variables are slightly different, but samples from the data files are still drawn from the same population, statistical techniques, usually a distance measure, are required to determine the closest possible match. This process is called **statistical matching** (Section 4.2.1). After pairing of records, imputation may be required for specific variables that the donor file has but the receptor file does not; various imputation techniques exist (Section 4.5).

3 FUSION PREREQUISITES

3.1 Data Harmonization

Data harmonization is a process of minimizing inconsistency and incompatibilities between datasets. Datasets from distinct sources are generally not designed for integration purposes due to different sample populations and variable definitions. Without resolving such issues, the precision of the data fusion outcome can be sacrificed (Bayard, et al., 2009). Table 1 presents a summary of data incompatibilities and possible solutions. For data harmonization process, Van der Lann (2000) proposed 8 types of harmonization issues to consider:

1) Harmonization of the definition of units - *Are units uniformly defined?*
2) Harmonization of reference periods - *Are survey data collected at roughly the same time?*
3) Completion of populations - *Do all sources cover the same target population?*
4) Harmonization of variables - *Are corresponding variables defined in the same way?*
5) Harmonization of classifications – *Are variables classified in the same way?*
6) Adjustment for measurement errors (accuracy) – *After harmonizing definitions, do the corresponding variables have the same value?*
7) Adjustment for missing data – *Do all the variables possess a value?* (**Note:** TTS 2.0 only takes complete data records, unless the question is optional, like income information of 2016 TTS survey. However, external datasets used to enrich the core survey may not possess certain variables).
8) Derivation of variables - *Are all variables derived using the combined information from different* sources?

Scanu et al. (2008)[1] condensed the above issues into three main categories: the harmonization of the **unit** (1-3), harmonization of **variables** (4, 5, 8), and other aspects (6, 7). We would discuss each of these issues separately:

For the harmonization of units, datasets need to refer to the same **population**, or subsets of the same population. This way datasets are more comparable because variables could have similar patterns or distributions (Roszka, 2015). For example, while fusing TTS survey data and household travel survey from Montreal would not generate a meaningful result, fusing TTS core and satellite surveys could enrich both datasets since they are both drawn from the same GTHA population. Another possibility is that two datasets are temporally lagged (Scanu & Tuoto, 2008). Temporally lagged datasets can be those collected by a single survey conducted in multiple years (e.g. previous TTS surveys), or different surveys of the same population conducted at different times.

For the harmonization of variables, it is important to identify the **common variables**, and determine whether they can be used for matching. If common variables have different definitions or incompatible classifications, there is no possibility for harmonization. Some common variables possess different categorizations of the same definition, and a new category can be derived by aggregating information (Refer to Table 1, semantic incompatibility). Variables such as age class, zonal area, or any continuous variable divided into classes can be re-categorized, though there may be a loss of resolution after such a process. Moreover, new common variables can be derived by transforming existing variables, if they satisfy certain criteria such as similar definitions and distributions. (Roszka, 2015) (D'Orazio, et al., 2006).

For other aspects such as measurement **errors** and **missing data**, there are no general solutions to deal with them, as the property of "error" and "missingness" is unique for each dataset. Metrics can be designed to estimate the errors and bias of satellite surveys before linking to the core, but there is no systematic solution on how to account for the errors in the literature. For missing data, imputation techniques exist, as described in Section 4.5.

[1] ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data, Report of WP2. Recommendations on the use of methodologies for the integration of surveys and administrative data. Link: http://cenex-isad.istat.it/archivio/Reports_of_the_project_workpackages/Report_of_WP2.pdf. Written by multiple authors.

TABLE 1: INCOMPATIBILITIES BETWEEN DATA FILES AND SOLUTIONS (BAYARD, ET AL., 2009) (MILLER, ET AL., 2014)

Types of incompatibility	Examples	Solutions
Spatial	The overlap between zones. Specifically, a traffic zone may intersect with several census tracts.	1. Aggregate both zone systems into a super-zone, where each of the original zones is fully contained inside the super-zone (higher level of aggregation). 2. Allocate units of observation from one zone to the other zone at which the first zone intersects the second. 3. Disaggregate both zone systems into a common, finer zone system.
Temporal	Temporal incompatibility between a continuous data collection system (e.g. smart card) and a transportation survey conducted at a certain period.	Solutions are a case-by-case basis. For non-consistent reference periods for the population, demographic projections need to be performed (Roszka, 2015).
	The same survey conducted at different times (e.g. previous TTS surveys).	Create a (pseudo)-panel to study the evolution of variables over time (Scanu & Tuoto, 2008)
Semantic	Two data files have a common variable but use different aggregation categories (e.g. 1-year period versus 5-year period).	1. If two categorizations map consistently into one another (5-year period versus 10-year period) then aggregate two 5-year categories into one 10-year category. 2. Assume uniform distribution, split one categorization and allocate portions to the other categorization. 3. Assume uniform distribution, use a common finer categorization, this means both aggregation categories need to be disaggregated statistically.

3.2 Choice of Matching Variables

Matching variables should be selected from common variables; mathematically, this can be stated as $Z_{matching} \subset Z_{common}$. Theoretically, all common variables could be used in matching procedures, but the computational effort can become substantial as the number of matching variables increases (D'Orazio, et al., 2006). Therefore, it is wise to select as fewer matching variables as possible, specifically those that have high explanatory power in their respective datasets.

Based on the literature, there are, in general, 4 ways to determine matching variables:

1) Compute the weighted frequency distributions and differences for each variable (Leulescu & Agafitei, 2013). If the differences are higher than a prescribed threshold, then variables are *incoherent*. This method is empirical but there is no theory to determine how large the threshold should be. Scanu et al. (2008) also recommended a similar approach by plotting histograms.
2) Use a distance measure to quantify the differences in the distribution of common variables. Leulescu & Agafitei (2013) used the Hellinger distance which ranges from 0 to 1, where 0 indicates perfect similarity of probability distributions. An empirical value of less than 0.05 indicates acceptable similarity of the distributions. However, one major limitation is that the Hellinger distance does not take sampling variability into account.

3) Use statistical measures of association or correlation (such as χ^2) amongst all common variables. Note that this can be different depending on the type of variables (e.g. nominal, ordinal, or interval). More detail is provided by D'Orazio et al. (2006).
4) Consider dependence relationships between common and specific variables. Linear regression models and **stepwise regression** procedures can be used for selecting matching variables. A more advanced generalized linear model can also be applied (McCullagh & Nelder, 1989). If the relationship is non-linear, **Classification and Regression Trees** (CART; Breiman et al., 1984) and **Random Forest** technique are available (Breiman, 2001). Graphical structure, such as **Bayesian Networks** (Cowell et al., 1999), offers a clear solution for understanding the probabilistic relationships between variables. (Scanu & Tuoto, 2008)

4 COMMON FUSION TECHNIQUES

4.1 Classifications

Three primary classifications exist for data fusion techniques in literature: matching with certainty/uncertainty, implicit/explicit methods, and parametric/non-parametric approaches. All classifications consider the **objectives** of the data fusion to define proper classes. If the objective is to generate a synthetic dataset that combines all the data or concatenate two datasets together, then the process is classified as **micro** (Refer to Figure 2). If the objective is to study the relationship between variables from different datasets, such as estimation of correlation coefficients, then the process is classified as **macro**. Macro processes do not integrate datasets at micro level (e.g. record by record), which means datasets are not "physically" fused (D'Orazio, 2013) (Bayard, et al., 2009). In transportation research, the **micro** approach has greater relevance for modelling, but understanding the joint relationship between variables from a **macro** approach is also necessary (Bayard, et al., 2009). Table 2 illustrates how classifications of methods and data fusion objectives relate to each other.

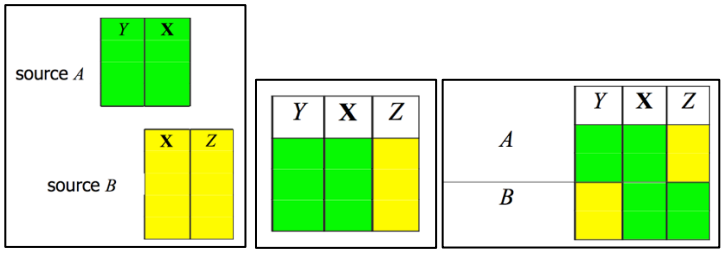


FIGURE 2: LEFT: DATA FUSION FRAMEWORK, COMMON VARIABLE X AND SPECIFIC VARIABLES Y AND Z. MIDDLE: MICRO APPROACH USING STATISTICAL MATCHING; RIGHT: MICRO APPROACH USING FILE CONCATENATION. (D'ORAZIO, 2013)[2]

TABLE 2: THE RELATIONSHIP BETWEEN DATA FUSION OBJECTIVES AND METHODS.

Objectives	Methods & Use Cases			Further Categorization	
Macro	Correlation coefficient, or regression coefficient between specific variables Y and Z.			Parametric	
Micro	Matching with certainty	Exact matching/record linkage		Non-parametric	
	Matching with uncertainty	Explicit	Use regression models to impute variables in receptor file	Mostly parametric	Mixed
		Implicit	Nearest neighbor statistical matching	Non-parametric	

Miller et al. (2014) classified data fusion into two categories based on whether the common variables from the two datasets can be matched with **certainty** or not. If yes, then no statistical procedure is needed; this method is called **exact matching**, or record linkage. If matching can only occur with **uncertainty**, explicit and implicit methods are available (D'Ambrosio, et al., 2007). **Explicit** methods construct a **regression** model Y = f(X) using the donor file to predict a specific variable as a function of common variables. The model is then used to **impute** the same specific

[2] Link: http://en.eustat.eus/productosServicios/datos/Seminario_55_Slides.pdf. A succinct summary of D'Orazio's textbook and functions of his statistical matching package R-StatMatch.

variable that is missing in the receptor file (Miller, et al., 2014). **Implicit** methods find the closest match between donor and receptor records based on a nearest-neighbor principle (e.g. distance between common/matching variables).

Data fusion can also be classified based on the parametric features of the model, as illustrated in Figure 3. If the joint distribution of variables is assumed to be one of the known probability distributions, then the fusion problem is **parametric;** this mainly involves parameter estimations using maximum likelihood principles. If no probability distribution is specified, then the process is **non-parametric** (Leulescu & Agafitei, 2013). **Mixed** approaches usually contain parametric step to impute the missing value, followed by statistical matching to find the closest match. The following sub-sections will provide an overview of common fusion techniques.

	Approaches		
Objectives SM	Parametric	Nonparametric	Mixed
Macro	Yes	Yes	No
Micro	Yes	Yes	Yes

FIGURE 3: COMBINATIONS OF OBJECTIVE AND APPROACHES (D'ORAZIO, 2013). D'ORAZIO ET AL. (2006) DESCRIBED EACH POSSIBLE COMBINATION IN DETAIL.

4.2 Non-parametric Approach

4.2.1 Nearest-neighbor Approach

Based on the classification method described in Table 2, the nearest-neighbor approach is a *micro*, *non-parametric*, and *implicit* technique. It is used when records from two surveys cannot be matched with certainty – if two records share very similar characteristics but there is no way to confirm that they belong to the same respondent. When common and matching variables are properly selected, a distance function d_{ij} is defined to measure the closeness between matching variables. The objective is to find the closest pairs of donor and receptor records, provided that the distance function is minimized, as referred in Equation 1 (Rässler, 2002).

$$\sum_{j=1}^{n_B}\sum_{i=1}^{n_A} \boldsymbol{d_{ij}w_{ij}} \qquad \boldsymbol{w_{ij} \geq 0,\ i = 1, \dots\dots, n_A,\ j = 1, \dots\dots, n_B} \qquad \textit{(1)}$$

where,

- i: Donor dataset A. Index ranges from 1 to n_A;
- j: Receptor dataset B. Index ranges from 1 to n_B;
- d_{ij}: Distance between matching variables from donor and receptor datasets;
- w_{ij}: Weight attached to the distance function. $w_{ij} \geq 0$. Previous literatures did not specify if it must be an integer but worked examples from Rodgers (1984) and Rubin (1986) used positive integer values.

A few variations or refinements of the algorithm are listed below (Rässler, 2002):

- ***Normalization of matching variables***: Matching variables may be of different scales in different data files; therefore, it would be convenient to standardize them as mean of 0 and standard deviation of 1.

- ***Weighted distance functions*** $d_{ij}w_{ij}$: Each matching variable exhibits different importance for the matching process. A weight w_{ij} could be multiplied to the distance between each pair of matching variables.
- ***Definition of critical variables***[3]: Matching variables such as gender, or region, are critical for the matching process. Matching occurs amongst records of the *same* critical variables (i.e. **matching class**). Establishing matching classes can reduce the computation power required, and offers a "sanity" check for the result. For example, a male in the donor file should never match with a female in the receptor file.
- ***Limited usage for each donor record***: If a donor record is used multiple times, the estimation of true variance would be underestimated. A penalty weight, which would be added to the distance function, can be assigned to already used donors.

Rodgers (1984) presented 2 nearest-neighbor based techniques: **unconstrained** and **constrained matching**. For **unconstrained matching**, there are no restrictions on how many times each donor record is taken for matching or whether each donor record must be used for matching or not. It has an advantage of allowing for the closest possible match between records. A worked example is provided by Rodgers (1984), and subsequently adapted by Rässler (2002):

File A

Unit i	Weight w_i^A	Z_1^A	Z_2^A	X
$A1$	3	1	42	x_1^A
$A2$	3	1	35	x_2^A
$A3$	3	0	63	x_3^A
$A4$	3	1	55	x_4^A
$A5$	3	0	28	x_5^A
$A6$	3	0	53	x_6^A
$A7$	3	0	22	x_7^A
$A8$	3	1	25	x_8^A

File B

Unit j	Weight w_j^B	Z_1^B	Z_2^B	Y
$B1$	4	0	33	y_1^B
$B2$	4	1	52	y_2^B
$B3$	4	1	28	y_3^B
$B4$	4	0	59	y_4^B
$B5$	4	1	41	y_5^B
$B6$	4	0	45	y_6^B

Statistically matched file, recipient file A

Matched unit ij	Weight w_{ij}	Z_1^A	Z_2^A	Z_2^B	Distance d_{ij}	X	Y
$A1, B5$	3	1	42	41	1	x_1^A	y_5^B
$A2, B5$	3	1	35	41	6	x_2^A	y_5^B
$A3, B4$	3	0	63	59	4	x_3^A	y_4^B
$A4, B2$	3	1	55	52	3	x_4^A	y_2^B
$A5, B1$	3	0	28	33	5	x_5^A	y_1^B
$A6, B4$	3	0	53	59	6	x_6^A	y_4^B
$A7, B1$	3	0	22	33	11	x_7^A	y_1^B
$A8, B3$	3	1	25	28	3	x_8^A	y_3^B

FIGURE 4: AN EXAMPLE OF UNCONSTRAINED MATCHING (RODGERS, 1984) (RÄSSLER, 2002). LEFT: DONOR AND RECEPTOR FILES BEFORE MATCHING. RIGHT: MATCHED FILE. NOTE THAT RECORD B6 IS NOT USED AT ALL.

As shown in Figure 4, records from donor file A are paired with records in receptor file B. Common variable Z_1 in both files is regarded as a critical variable; therefore, records must satisfy $Z_1^A = Z_1^B$ to be considered as matching candidates. Moreover, as the goal is to minimize the distance between Z_2^A and Z_2^B, record B6 is not present in the matched file because it generally has a very large distance to the matching candidates. However, while the distance is minimized, it is not guaranteed that the original properties, the mean and standard deviation of file A and B are kept consistent. For example, mean of Y in original file B and matched file is:

[3] Also known as "donation classes", in D'Orazio, et al. (2006)

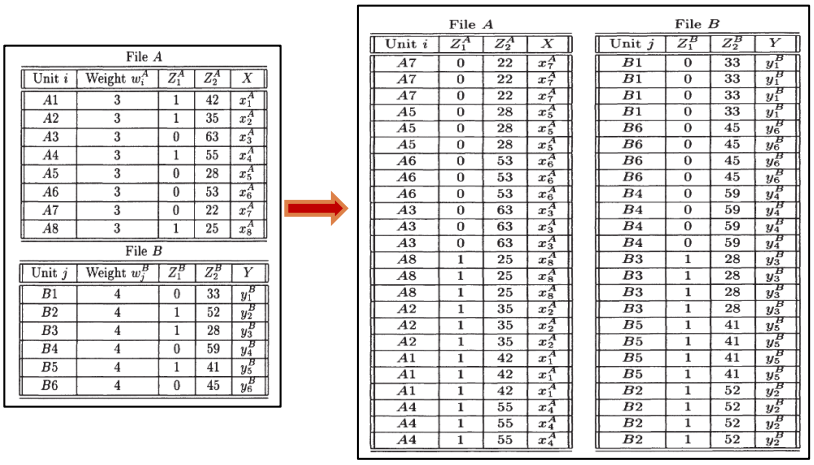
$$\frac{4}{24}\sum_{i=1}^{6} y_i^B \neq \frac{3}{24}(2y_1^B + y_2^B + y_3^B + 2y_4^B + 2y_5^B) \qquad (2)$$

The mean is different, it turns out the standard deviation is also not identical. It can be concluded that the matching file does not preserve the marginal distribution in the original files, hence the result is biased. To avoid this problem, another approach, **constrained matching**, is used which requires a preservation of the original weights w_i and w_j, as shown in Equation 3-4. This process involves two steps (Rubin, 1986):

$$\sum_{j=1}^{n_B} w_{ij} = w_i \qquad i = 1, \ldots\ldots, n_A, \quad and \qquad (3)$$

$$\sum_{i=1}^{n_A} w_{ij} = w_j \qquad j = 1, \ldots\ldots, \mathbf{n_B} \qquad (4)$$

Step 1: "Explode" the original files. If a record in file A has weight w_{Ai}, then the record is duplicated w_{Ai} times. Reorder the exploded files based on common variables Z_1 and Z_2, where Z_1 (critical variable) has higher precedence. As shown in Figure 5 below, for each Z_1 value, Z_2 is ranked in order of increasing value.



File A

Unit i	Weight w_i^A	Z_1^A	Z_2^A	X
A1	3	1	42	x_1^A
A2	3	1	35	x_2^A
A3	3	0	63	x_3^A
A4	3	1	55	x_4^A
A5	3	0	28	x_5^A
A6	3	0	53	x_6^A
A7	3	0	22	x_7^A
A8	3	1	25	x_8^A

File B

Unit j	Weight w_j^B	Z_1^B	Z_2^B	Y
B1	4	0	33	y_1^B
B2	4	1	52	y_2^B
B3	4	1	28	y_3^B
B4	4	0	59	y_4^B
B5	4	1	41	y_5^B
B6	4	0	45	y_6^B

File A

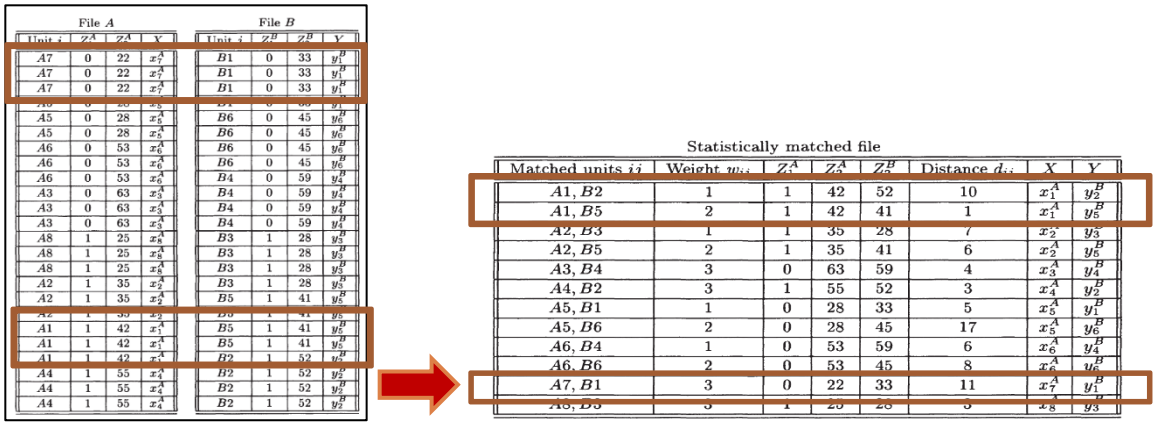
Unit i	Z_1^A	Z_2^A	X
A7	0	22	x_7^A
A7	0	22	x_7^A
A7	0	22	x_7^A
A5	0	28	x_5^A
A5	0	28	x_5^A
A5	0	28	x_5^A
A6	0	53	x_6^A
A6	0	53	x_6^A
A6	0	53	x_6^A
A3	0	63	x_3^A
A3	0	63	x_3^A
A3	0	63	x_3^A
A8	1	25	x_8^A
A8	1	25	x_8^A
A8	1	25	x_8^A
A2	1	35	x_2^A
A2	1	35	x_2^A
A2	1	35	x_2^A
A1	1	42	x_1^A
A1	1	42	x_1^A
A1	1	42	x_1^A
A4	1	55	x_4^A
A4	1	55	x_4^A
A4	1	55	x_4^A

File B

Unit j	Z_1^B	Z_2^B	Y
B1	0	33	y_1^B
B1	0	33	y_1^B
B1	0	33	y_1^B
B1	0	33	y_1^B
B6	0	45	y_6^B
B6	0	45	y_6^B
B6	0	45	y_6^B
B4	0	59	y_4^B
B4	0	59	y_4^B
B4	0	59	y_4^B
B4	0	59	y_4^B
B3	1	28	y_3^B
B3	1	28	y_3^B
B3	1	28	y_3^B
B3	1	28	y_3^B
B5	1	41	y_5^B
B5	1	41	y_5^B
B5	1	41	y_5^B
B5	1	41	y_5^B
B2	1	52	y_2^B
B2	1	52	y_2^B
B2	1	52	y_2^B
B2	1	52	y_2^B

FIGURE 5: FILE "EXPLOSION" PROCESS. EACH RECORD IS REPLICATED BASED ON ITS WEIGHT.

Step 2: Match the exploded files A and B in parallel. For example, in Figure 5, A7 has a weight of 3; therefore, it is repeated 3 times in the exploded file A. On the other hand, as B1 has a weight of 4, it is repeated 4 times in file B. Since file A is a donor, A7 is matched with 3 B1, the remaining B1 is matched with A5, this leaves 2 A5 remaining, which are to be matched with B6. Repeat this process until all records in both files are used in the matching process, resulting in the final matched file as shown in the right image in Figure 6.



File A

Unit i	Z_1^A	Z_2^A	X
$A7$	0	22	x_7^A
$A7$	0	22	x_7^A
$A7$	0	22	x_7^A
$A5$	0	28	x_5^A
$A5$	0	28	x_5^A
$A6$	0	53	x_6^A
$A6$	0	53	x_6^A
$A6$	0	53	x_6^A
$A3$	0	63	x_3^A
$A3$	0	63	x_3^A
$A3$	0	63	x_3^A
$A8$	1	25	x_8^A
$A8$	1	25	x_8^A
$A8$	1	25	x_8^A
$A2$	1	35	x_2^A
$A2$	1	35	x_2^A
$A1$	1	42	x_1^A
$A1$	1	42	x_1^A
$A1$	1	42	x_1^A
$A4$	1	55	x_4^A
$A4$	1	55	x_4^A
$A4$	1	55	x_4^A

File B

Unit i	Z_1^B	Z_2^B	Y
$B1$	0	33	y_1^B
$B1$	0	33	y_1^B
$B1$	0	33	y_1^B
$B6$	0	45	y_6^B
$B6$	0	45	y_6^B
$B6$	0	45	y_6^B
$B4$	0	59	y_4^B
$B4$	0	59	y_4^B
$B4$	0	59	y_4^B
$B3$	1	28	y_3^B
$B3$	1	28	y_3^B
$B3$	1	28	y_3^B
$B5$	1	41	y_5^B
$B5$	1	41	y_5^B
$B5$	1	41	y_5^B
$B2$	1	52	y_2^B
$B2$	1	52	y_2^B
$B2$	1	52	y_2^B

Statistically matched file

Matched units ij	Weight w_{ij}	Z_1^A	Z_2^A	Z_2^B	Distance d_{ij}	X	Y
$A1, B2$	1	1	42	52	10	x_1^A	y_2^B
$A1, B5$	2	1	42	41	1	x_1^A	y_5^B
$A2, B3$	1	1	35	28	7	x_2^A	y_3^B
$A2, B5$	2	1	35	41	6	x_2^A	y_5^B
$A3, B4$	3	0	63	59	4	x_3^A	y_4^B
$A4, B2$	3	1	55	52	3	x_4^A	y_2^B
$A5, B1$	1	0	28	33	5	x_5^A	y_1^B
$A5, B6$	2	0	28	45	17	x_5^A	y_6^B
$A6, B4$	1	0	53	59	6	x_6^A	y_4^B
$A6, B6$	2	0	53	45	8	x_6^A	y_6^B
$A7, B1$	3	0	22	33	11	x_7^A	y_1^B
$A8, B3$	3	1	25	28	3	x_8^A	y_3^B

FIGURE 6: LEFT: EXPLODED FILE A AND B. MATCH BOTH FILES "IN PARALLEL". RIGHT: THE RESULT OF MATCHING PROCESS.

Constrained matching also has a few limitations. Although the original weights are preserved, leading to less biased fused record compared to unconstrained matching, this method cannot guarantee the closest match between donor and receptor files. Moreover, the exploded files essentially contain repeated records and, therefore, do not represent the population values (Rubin, 1986). Finally, D'Orazio, et al. (2006) suggests the computational cost for this algorithm is heavy.

Constrained and constrained matching are also subject to **Conditional Independence Assumption** (CIA) – the independence of specific variables Y and Z, given the common variable X, as shown in Equation 5 below.

$$f(x, y, z) = f(y|x)f(z|x)f(x) \qquad (5)$$

In other words, the relationship between Y and Z can be completely explained by X. Care must be taken as CIA is untestable by examining the fused dataset. As a result, if CIA is wrongly assumed, the matching result will be misleading. Alternatively, approaches, such as using the auxiliary sources, are available to overcome the CIA (D'Orazio, et al., 2006).

4.2.2 Exact Matching

Exact matching, also called **record linkage**, is a special case of the *non-parametric* approach. It does not require statistical measures since records from donor and receptor files belong to the same individuals. However, in real-life scenarios, the unique identifiers across datasets do not always use the same naming system. As a result, the use of an individual's characteristics, such as address, is necessary. Complicating matters, typing errors, and harmonization issues, such as different variable definitions, categorization, and temporal incompatibility, need to first be resolved (See Section 3). The goal for exact matching is to use a set of matching criteria and decision rules to classify all record pairs into matches, non-matches, and possible matches (Denk & Hackl, 2004).

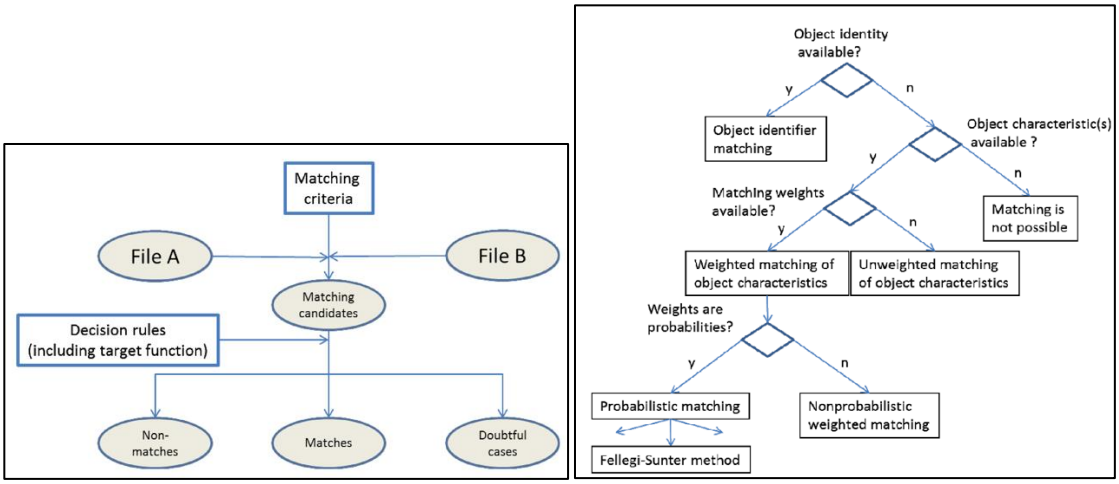


FIGURE 7: EXACT MATCHING PROCEDURES. LEFT: MAIN INGREDIENTS; RIGHT: A DECISION TREE OF MATCHING METHODS. [WILLENBORG & VAN DER LAAR (2014) ADAPTED FROM WILLENBORG & HEERSCHAP (2012)]

Willenborg & Heerschap (2012)[4] elaborated key components and methods in exact matching:

Key Components (Figure 7, left):

- ***Matching criteria***: Adjustable metrics to determine matching candidates. They can be a measure of similarity between records (e.g. Hamming distance), with cut-off values imposed to limit the number of matching candidates.
- ***Decision rules***: Criteria to identify matches, non-matches, and doubtful cases.
- ***Blocking procedures***: Partitioning of matching files based on selected common variables, such as gender, to partition matching files. This is similar to the matching class described in Section 4.2.1.

Methods (Figure 7, right):

- ***Object identifier/characteristics matching***: Matching that occurs when the unique identifiers or primary matching keys (such as survey ID) are of sufficient quality and common in both datasets. Otherwise, use object characteristics (such as address), or secondary matching keys are used to determine the matching candidates.
- ***Weighted/unweighted matching***: Weights used to indicate the importance of matching variables.
- ***Probabilistic matching***: Matching based on Fellegi-Sunter model (1969) to handle differences between matching variables due to processing errors, temporal and semantic compatibilities.

4.3 Parametric Approach

Parametric approach is a process of estimating covariance or the regression coefficients of the fused dataset. As illustrated in Figure 3, the parametric approach has micro and macro objectives, which employ different mathematical techniques. The macro objective is concerned with the estimation of covariance of the specific variables. Typically, maximum likelihood estimation is used for continuous variables, whereas the log-linear model is used for categorical variables. On the other hand, the micro objective focuses on the estimation of regression coefficients. Two steps are involved: first, **file concatenation** is performed (Rubin, 1986, as shown in Figure 8), followed by the use of conditional mean matching (e.g. with regression imputation) or draws from the predicted

[4] *Matching Series (Theme: Matching):* produced by Statistics Netherlands, focuses on the theoretical aspects of exact matching methods in practice. Link: https://www.cbs.nl/NR/rdonlyres/0EDC70A4-C776-43F6-94AD-A173EFE58915/0/2012Matchingart.pdf
Similar document produced by Statistics New Zealand, *Data Integration Manual: 2nd edition* (2013), focuses on the legal and operational aspects. Link: http://m.stats.govt.nz/methods/data-integration/data-integration-manual-2edn.aspx

distribution (e.g. with stochastic regression imputation) to impute the missing variables (D'Orazio et al., 2006)[5]. Overall, the parametric approach may be unreliable if the regression model is incorrectly specified (D'Orazio et al., 2006).

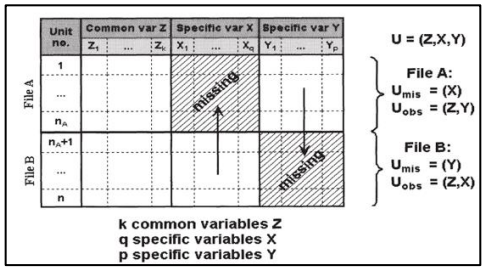


FIGURE 8: FILE CONCATENATION (RÄSSLER, 2002). FILE B DIRECTLY APPENDS TO FILE A. NOTE THAT THE COMMON VARIABLES ARE Z, NOT X.

4.4 Mixed Approach

The mixed approach combines parametric and non-parametric processes and takes the advantages of both. Specifically, the non-parametric process offers "protection" against the model misspecification in the parametric process (D'Orazio, 2013). For the parametric step, it involves the estimation of parameters of a model that describes the relationship between specific variables Y, Z, and the common variable X. For continuous variables, the model is usually linear regression with a residual term, whereas categorical variables employ log-linear models. Once the parameters are confirmed, the models are used to impute missing variables such as non-response from the survey questions or specific variables only contained in one dataset. Next, a nearest-neighbor approach (described in Section 4.2.1) is used to match records and produce a fused synthetic dataset. Overall, the matching step is similar for any types of variable, but several variations exist for the model estimation as elaborated formally in D'Orazio et al. (2006)[6]; a more succinct description is found in and D'Orazio (2013). Rubin (1986)[7] and Rässler (2002)[8] also provided relevant worked examples.

4.5 Imputation[9]

Imputation is a process that utilizes existing information to fill in missing values caused by survey non-response or file concatenation. Imputation can occur before or after the nearest-neighbor matching step, or it could be a standalone process if the objective of data fusion is macro only. Missing values can be imputed based on parametric models, such as conditional mean imputation or regression (Rässler, 2002). Imputed values are generated by models and hence are not "live" or observed values. Alternatively, missing values can be imputed by transferring existing values from the donor file to the receptor file. Methods such as random hot-deck, rank hot-deck, and distance hot-deck are the most commonly used (D'Orazio et al., 2006)[10]. These techniques are categorized as single imputation, which only creates one complete dataset but ignores the uncertainty of the process of predicting missing values.

To address this uncertainty, Rubin (1987) proposed multiple imputations. In this method, several values are first imputed per missing datum to create the same number of complete dataset (i.e. 3 imputations correspond to 3 complete datasets, as illustrated in Figure 9). Next, statistical procedures are performed to analyze each dataset. Finally, the results are pooled to produce the final estimates of the missing values. The goal of

[5] Statistical matching: theory and practice, Chapter 2.1-2.2.
[6] Statistical matching: theory and practice, Chapter 2.5 and 3.6.
[7] Statistical matching using file concatenation with adjusted weights and multiple imputations, pp. 91-93.
[8] Statistical matching: a frequentist theory, practical applications, and alternative Bayesian approaches, Chapter 2.4.
[9] Rudra (2014) provided a comprehensive review of imputation techniques to date, tested and compared them in the context of missing response in *GTHA Movement of Commercial Goods and Service Survey*. Link: https://tspace.library.utoronto.ca/bitstream/1807/68004/1/Rudra_Malvika_201411_MAS_thesis.pdf
[10] Statistical matching: theory and practice, Chapter 2.4.1 – 2.4.3. Distance hot-deck is identical to the nearest neighbor approach.

multiple imputations is not to produce the "correct" missing values, but instead to generate datasets that provide statistically valid inferences of parameters from the incomplete data (Rudra, 2014). As a summary, Table 3 compares the methodology and properties of single and multiple imputations.

TABLE 3: AN OVERVIEW OF IMPUTATION TECHNIQUES (RÄSSLER, 2002) (RUBIN, 1987)

	Single Imputation	**Multiple Imputation**
Method	"Fill in" plausible values based on mean, regression, and other techniques.	"Fill in" plausible values based on some probability distributions/assumptions. Assumptions are based on missing data pattern.
Property	Every single missing value is imputed once only. Imputed value does not reflect sample variability (i.e. uncertainty of missing data).	Every single missing value is imputed m times (these m values have certain distributions), that means each imputation creates one unique data set.

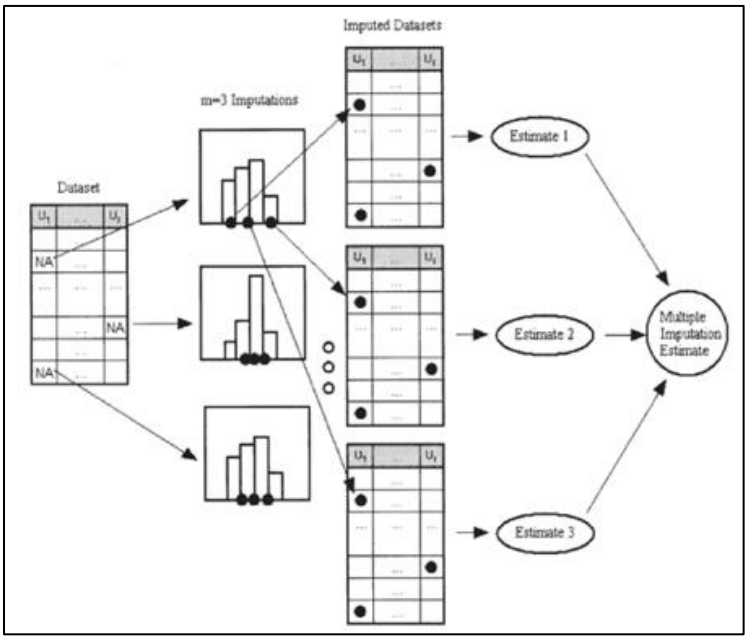


FIGURE 9: MULTIPLE IMPUTATION CONCEPT, GRAPHIC CREATED BY RÄSSLER (2002).

5 TECHNIQUES RELEVANT TO DATA FUSION

This section provides an overview of common techniques that may not be fusion *by definition*, but they do combine datasets together as an input of simulation models. Frequently mentioned techniques in transportation research include population synthesis, RP/SP, and the pseudo-panel.

5.1 Population Synthesis

The demand for accurate travel simulation, such as agent-based simulation model requires a population input at the individual or household level via population synthesis. Traditional travel surveys provide individual records of travel information, but they only cover a small percentage of the overall population. On the other hand, the Census, while collecting sociodemographic variables of the entire population, is aggregated to the zonal level due to privacy reasons. Techniques of combining both datasets have been developed and refined since the 1940s (Deming & Stephan, 1940). Standard procedures in transportation context were developed by Beckman et al. (1996). They used the iterative proportional fitting (IPF) technique that iteratively updates a disaggregated dataset such that it matches the marginal total of the of the aggregated dataset, as illustrated in Figure 10. Prichard et al. (2011) provided a comprehensive overview of IPF techniques, and propose new computational methods for synthesizing more and finer categories of sociodemographic attributes. These methods are classified as "sample-based", as census data serves as a reference sample for the synthetic population. Namazi-Rad et al. (2017) also introduced "sample-free" method, a heuristic approach that does not solely rely on the census.

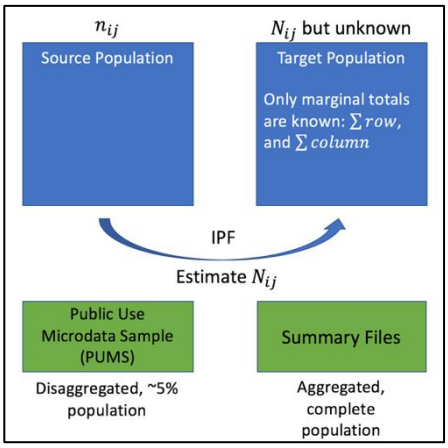


FIGURE 10: POPULATION SYNTHESIS THROUGH IPF PROCESS.

5.2 RP/SP Survey Fusion

Traditional travel survey questions are mostly retrospective, meaning people recall what they did and state their actual choices (e.g. choosing to take a bus). This type of survey is called revealed preference (RP). RP surveys are generally reliable, but they cannot extrapolate peoples' choices in hypothetical scenarios, such as what would happen if the bus ticket price increases by $2. However, stated preference (SP) surveys allow respondents to choose their actions during hypothetical scenarios; this type of survey can offer valuable insights on transport planning. The main disadvantage of SP is that people's choice may be biased because they may not necessarily behave the same way if the hypothetical situation becomes real. By combining RP/SP, their limitations can be mitigated, while retaining their advantages (Train, 2002).

In the transportation context, joint RP/SP data is often the input when estimating discrete choice models. The process of joining RP/SP is exact matching by a unique ID, as survey respondents answer both types of questions; hence, data fusion is mainly the estimation process of the scale parameters that correct the systematic bias of the SP data. Research has been done on using RP/SP to examine mode choice behavior of

cross-regional commuters by comparing three econometric models, and the authors found the joint RP/SP data is able to quantify the probabilistic response in terms of level-of-service attributes caused by the change of policies (Mahmoud, et al., 2015). Moreover, research by Lin, et al. (2017) used the StudentMoveTO dataset, a student-focused travel survey containing RP and SP questions, to evaluate the effect of latent attitudes that affect the level of multimodality among university students (Lin, et al., 2017).

5.3 Pseudo-Panel

The pseudo-panel is a combination of datasets from repeated cross-sectional surveys conducted in different years but with the same set of variables. If individual records from different surveys match sufficiently, then they can be linked and treated as the same respondents, therefore providing insights into how the individual attributes, such as travel behavior, change over time. (Miller, et al., 2014). Since most travel demand models focus on explaining current travel behavior, rather than forecasting future scenarios or "temporal transferability", recent research has utilized pseudo-panel dataset to generate models with better predicting power. For example, Habib et al. (2014) calculated "temporal factors" through logarithmic and polynomial formulations to pool TTS data from 1996, 2001 and 2006. Then, the evaluation of commuting mode choice preferences over a 10-year period was examined by comparing individual year-specific and pooled models, and it was concluded that the pooled model outperforms year-specific models in terms of temporal transferability. Salem et al. (2015) used the same datasets to construct a Meta model for activity-travel generation processes and yielded similar results.

6 DATA FUSION VALIDITY

Data fusion validity is a set of criteria for judging the quality of a data fusion procedure (Kiesl & Rässler, 2006). Four levels of validity and methods for checking them are presented in this section; a high-level summary is provided in Table 4.

The first and the lowest level of validity is to preserve marginal distributions. Data fusion is usually said to be successful if the empirical marginal and joint distributions of the common variable Z and specific variable Y in the fused file is similar to that in the donor file. Preliminary approaches involve comparing average values of common variable Z in donor and fused file, average values of original variable Y and imputed Y', and the correlation between Y and Y' for each common variable Z. Hypothesis tests such as the χ^2-test or the t-test are required to compare the empirical distributions or moments. (Kiesl & Rässler, 2006)

The second level is to preserve correlation structures. This can only be achieved based on the untestable assumption that specific variables are conditionally uncorrelated. Current approaches of checking this level of validity are lacking; Kiesl et al. (2006) stated that more research is needed to overcome this assumption.

The third level is to preserve joint distributions, which is subject to the Conditional Independence Assumptions (CIA), as explained in 4.2.1. This level might be achieved provided that the common variables possess a very high explanatory power. (Kiesl & Rässler, 2006)

The fourth and the most specific level is to preserve individual values, which means all the true values of the imputed variable Y' are reproduced in the fused file with certainty. This outcome is very unlikely to be achieved, and it is not necessary to attain this level of certainty, as imputed Y' is not meant to reflect the real values. (Kiesl & Rässler, 2006)

As suggested by Kiesl et al. (2006), more research should be focussed on designing or improving methods that are able to ensure the preservation of marginal distributions and correlation structures (first two levels), as achieving the last two levels are constrained mathematically and can be potentially misleading.

TABLE 4: DATA FUSION VALIDITY AND ADDITIONAL NOTES (KIESL & RÄSSLER, 2006) (RÄSSLER, 2002)

Levels of validity	Description	Comments
Preserving marginal distributions	• The empirical marginal and joint distributions in the donor file are preserved in the fused file. • If specific variable Y is imputed from (X, Z), where Z is the common variable, then $f'_Y = f_Y$ and $f'_{Y,Z} = f_{Y,Z}$.	Compare average values • Common variables Z in donor and receptor files • Imputed Y' and Y in donor file Compare correlations: • Y' and Y (for each common variable Z) Compare distributions using χ^2-test or t-test.
Preserving correlation structures	Cov'(X, Y, Z) = Cov(X, Y, Z), a measure of association of variables	The fused file should have the same moments and correlation structure as the actual population of interest. More research is needed to overcome the assumption such that specific variables are conditionally uncorrelated.
Preserving joint distributions	$f'_{X,Y,Z} = f_{X,Y,Z}$	If Conditional Independence Assumptions (CIA) (as described at the end of 4.2.1) is testable.
Preserving individual values	When the true value of imputed Y variable is reproduced: $y'_i = y_i$.	Unlikely to be achieved, maybe for the rare case of exact matching.

7 CASE STUDIES

Data fusion in the context of travel surveys is still a new concept; therefore, applications are rare. Although a few case studies in the literature have used the techniques from previous chapters, most of them did not specify the details of their fusion process. In other cases, researchers simply called the process "fusion" if two or more datasets were considered for data analysis. This is particularly evident for passive datasets due to their lack of socio-demographic variables and differences in data format, accuracy, and bias. As a result, most research has focused on **mining** passive datasets with the help of data from traditional travel surveys or census, instead of directly fusing them. Moreover, data fusion is highly *contextual*, meaning the method employed is directly related to the specific objectives after fusion. Applications from past literature typically fit into the 4 objectives listed below:

- ***Survey Integration***: Combining surveys to correct sampling bias using weighting procedures. These surveys are usually related: either one provides more detailed respondents' attributes than the other, or each survey targets specific groups of populations with similar questions.
- ***Data Enrichment***: Using a traditional travel survey or socio-economic variables from census to enrich a passive dataset, or combining passive datasets to yield more travel information.
- ***Data Matching***: Using rules or statistical methods to match records from two datasets.
- ***Pattern Identification***: Identifying trip chains and purposes using rules defined by land use information, points of interest (POI), or other surveys.

The purpose of this section is to identify gaps between data fusion techniques and current practices and provide relevant contexts for future research.

7.1 Survey Integration

Survey integration is a process of using weight adjustment to combine related surveys. For example, Nakamya et al. (2007) combined a time use survey and a household travel survey with common sociodemographic and travel behavior variables. Verreault & Morency (2016) combined web and landline travel surveys. Both studies investigated the impact of survey integration on trip characteristics such as trip rates and duration, and both yielded slightly higher number in the fused datasets, indicating a possible reduction of underreport (Nakamya et al., 2007) and proxy bias (Verreault & Morency, 2016).

Nakamya et al. (2007) discussed the differences of sample design between the time use and travel surveys. As they were conducted a year apart, a harmonization step was required to resolve the temporal lag; however, the authors only stated that the surveys were "cleaned", with weights calculated from census by gender, age, education level, and marital status at the personal level. Factors were also applied for the trips on different days of the week and in different months of the year. For data fusion, the **micro-approach** of the statistical matching procedure was used; however, the authors did not make clear how the common variables were selected and how the matching step was carried out. Instead, the description provided indicates that the method was likely a direct **concatenation** of datasets. For comparison, a fused, but unweighted dataset was also prepared. During the analysis stage, the distribution of socio-demographic data between the two surveys and fused dataset was compared, with weighted and unweighted options. It appears that the fused **and** weighted dataset consistently approximates the distribution of census data, the reference population. There, however, was no criterion provided on how the determination was made of when a distribution was close *enough*. The author also indicated that the fused dataset corrects the deficiency of underreported trips in the travel survey; however, how much of this deficiency is resolved and where the "ground truth" lies remain unclear.

Verreault & Morency (2016) conducted similar research on web and landline surveys, aiming to correct the under-sampling bias of 20-29-year-old students through a web survey. The landline-based travel survey only contacts participants whose numbers are in the phone book. The web survey aims to include students that are unreachable by the landline survey to avoid double counting (e.g. students living in the dormitory, or their landline numbers are

not included in the phone book). The method to ensure these two sample frames are independent is not trivial – it involves significant work **after** data collection process. Thus, the important part of this research is that it informs that travel surveys targeting to different populations should be designed and coordinated to ensure independent sample frames in the first place. This idea is applicable for the design of core and satellite surveys in TTS. In terms of data fusion phase, this paper only discussed the weight adjustments but did not specify what type of fusion technique was used.

In the context of core-satellite survey integration, the weighting procedure is well-documented in previous research, but the fusion process remains ambiguous. One could argue that fusion techniques from Section 4 were never used in the case studies; however, it is recommended that the authors be contacted to acquire further details on the process. The conduct of a real-world example, such as fusing 2016 TTS and 2015 StudentMoveTo data, is also recommended.

7.2 Data Enrichment

The latest research on data enrichment was conducted by Regt et al. (2017), focusing on combining smartcard and Global System for Mobile Communications (GSM) datasets to investigate transit user mobility patterns versus the overall travel demand. Specifically, smartcard data with a tap-in and tap-out information[11] was aggregated to the same level as GSM data[12], which accounts for the overall population. OD matrices for both datasets were generated hourly. To quantify the variation of travel demand, a mean percentage error (MPE) measure was introduced to track the hourly change of the OD matrices of smartcard and GSM separately. The data fusion step, as the author described, was to relate the normalized MPEs of both datasets and plot them against each other in a graph. **Figure 11** shows how the public transit usage and general flow of population have changed in each zone and different time periods. For example, during 11:00 – 12:00, the general population in this zone has increased by 40%, but the transit usage has decreased by 50%. This raises an interesting question for transit planners on why this could be possible, and how to address it. Overall, Regt et al. (2017) combined two datasets graphically to offer an interesting insight into the relationship between transit and overall travel demand. However, the authors also admit that direct fusion was impossible due to semantic incompatibility between datasets, and low spatial resolution of GSM data.

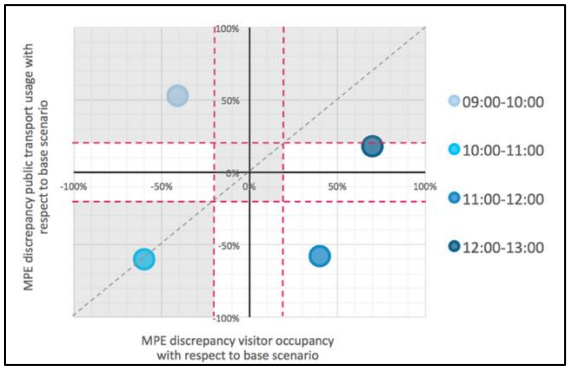


FIGURE 11: MPES OF SMARTCARD AND GSM DATA IN A GIVEN ZONE (REGT ET AL., 2017).

7.3 Data Matching

For matching records from different datasets, Spurr et al. (2015) offer a valuable attempt to match smartcard data records with data from household travel survey (HTS). To resolve temporal incompatibility, these smartcard

[11] Smartcard in many faring systems don't have tap-out option. For example, PRESTO card only contains boarding data, alighting information must be inferred. Various algorithms exist in the literature for inferring alighting information.
[12] GSM data was provided by one of the largest network providers of Netherland (Vodafone). It is aggregated and expanded to overall population based on census data.

data were extracted only for the duration when the HTS was conducted. Since the HTS only required respondents to recall their trips on a "typical average" day, the smartcard data distribution was first analyzed, and then a day that best represented a "typical average" day was selected. To overcome semantic incompatibility, HTS public transit trips were converted to smartcard transaction sequences through a series of translation dictionaries. The matching process resembled the **exact matching** procedure based on object characteristics. Here, traveller attributes such as station sequence and boarding locations were defined, followed by an attempt to produce a one-to-one match between HTS and smartcard records by adjusting the resolutions for both the time period and location (named as a spatiotemporal filter). The outcome was promising: about 50% of HTS transit users were matched by the smartcard dataset. Overall, this research shows the possibility of directly matching to verify HTS, but authors maintained that a systematic approach has yet to be devised and could be the focus in the future.

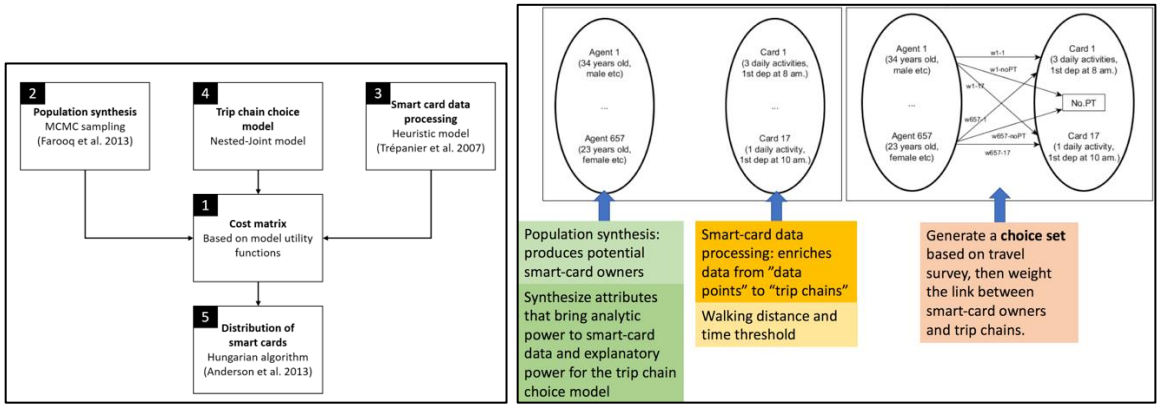


FIGURE 12: LEFT: A HIGH-LEVEL OVERVIEW OF THE SIMULATION APPROACH ON MATCHING SYNTHETIC POPULATION AND SMARTCARDS; RIGHT: A MORE DETAILED ILLUSTRATION OF THE MATCHING PROCESS. BOTH FIGURES ARE EXTRACTED FROM GRAPPERON ET AL. (2016).

Matching in a more advanced form based on a series of systematic steps was provided by Grapperon et al. (2016), who attempted to attach socio-demographic information to smartcards. They conceived this problem as solving a maximum weighted bipartite matching problem, with smartcards on one side of the graph and potential smartcard users on the other side. Five steps, as shown in Figure 12, are elaborated in the remainder of this section.

The first step is to conducting population synthesis using the Monte-Carlo Markov Chain (MCMC) method; this generates a collection of potential smartcard holders (synthetic agents) for the entire city. The data sources include the household travel survey and the census. The household travel survey is diary-based, collecting info on all the trips conducted (including the ones using smartcards) within one day for participants and their family members. The census was used to expand the survey data, and enrich it by adding missing variables such education, income and marital status.

The second step is to enrich smartcard data with alighting and activity locations, hence producing "trip chains". A rule-based methodology by Trépanier et al. (2007) was used to assume that public transit trips are chained: if the boarding location of the next trip is within 1 km of the previous trip, then it could be inferred that the alighting location of the previous trip is within this area. For activity locations, 30 mins between consecutive boardings is a threshold to differentiate a "transfer" from a "real activity" (Devillanine et al. 2012).

The third step is to develop a behavioural choice set model. The decision maker of this model is the synthesized population in the first step. It contains socio-demographic characteristics, which can be used to construct the utility function. The decision object is the trip chain, which contains attributes (e.g. time of the first/last departure) estimated from the smartcard data. A list of mode choice alternatives based on the travel survey and associated trip chains are organized as a nested joint model.

Results from the first three steps are incorporated into the construction of cost matrix using the utility values of the alternatives, which is in-turn optimized by the Hungarian Algorithm. This algorithm produces the closest association between the synthesized population and smartcard data. The result of the matching process is evaluated by comparing the marginal distributions of socio-demographic variables between the synthesized agents and data in travel survey.

In applying these steps, some discrepancies were observed, but the authors stated that the result could be improved by incorporating land use information in the choice set model, and refining the heuristics for inferring home and alighting locations. Overall, this was the first example that involves a systematic approach to data fusion and deserves more attention.

7.4 Pattern Identification

Several studies have used travel surveys and passive datasets to identify travel patterns, such as trip purposes. Kusakabe & Asakura (2014) explicitly mention a data fusion process for estimating trip purposes for smartcard holders. They used data from a household travel survey as a basis for constructing a naïve Bayes probabilistic model, which classified trip purposes based on the time interval between the arrival and next departure times at the same station. The accuracy of the model depends on the actual trip purposes ("ground truth") – some purposes have similar characteristics and are, hence, difficult to decipher. Other dynamic factors such as seasonal changes, special events, or policy changes can affect travel behavior. Therefore, they must be taken into account in the modelling process. Other studies such as inferring trip purposes through GPS data (Shen & Stopher, 2013), or analyzing job-housing relationships via smart-card (Long & Thill, 2015) have treated household travel survey as a "ground truth" to both develop rules and to fuse other datasets. Although even using "static" travel survey data to find patterns in the "dynamic" passive datasets can still yield a high rate of pattern identification, atypical travel behaviors due to special events such as road closure may not be identified.

8 HYPOTHETICAL CASES FOR TTS 2.0

A core-satellite survey structure is recommended for the next TTS project in 2021; data sources that need to be included, and the methods to link them together are still under consideration. Data fusion is an inevitable process of integrating datasets; however, it is highly contextual – the content of the datasets should be fully understood before choosing appropriate fusion techniques. The purpose of this section is to offer context relevant to the TTS and identify research opportunities to provide a motivation for data fusion.

8.1 Household Income Imputation

In 2016 TTS, an optional income question was added, allowing for respondents to select an income range. Preliminary analysis showed that the overall response rate from three survey options (mail, landline, and web) was around 80%, with 20% values, therefore, missing (Lo, 2017). This leads to several avenues of potential analysis:

- ***Determine the validity of existing responses***: First, investigate the pattern of non-response and determine if it is related to other survey variables. For example, there is a possibility that respondents with very low or high *actual* incomes refuse, or randomly report their numbers. Income surveys such as Canadian Income Survey (CIS) can be used as a reference to impute the missing values in TTS survey (Statistics Canada, 2017). However, it should be noted that CIS also contains imputed values, and the accuracy must also be evaluated.
- ***Expansion to GTHA population***: Use the 2016 census to calculate expansion factors so that the income information can be expanded to the entire population.
- ***Relationships between income and travel patterns***: Different income groups may exhibit different travel behaviors: such as travel time or modes. Linking TTS travel and income questions offer an opportunity to investigate any correlations.
- ***Year-over-year trend***: The correlations between income and travel behaviors can be investigated in different years. Constructing a pseudo-panel would allow for the examination of trends. Since the income variable was completely missing in previous TTS, past CIS data can be used to impute the income variable.

8.2 PRESTO Card Data Analysis

During summer 2017, all PRESTO data to date was acquired by UTTRI for further analysis. Overall, the data lacks spatial accuracy, as the stopping location is determined with varying degrees of accuracy (most agencies use scheduled rather than actual positions of vehicles in determining tap locations). GO Stations are not labelled by GPS coordinates but by zones. Moreover, Metrolinx is still working to expand the PRESTO service to all transit agencies (Hollis, 2017). By the next TTS in 2021, data accuracy issues will hopefully be fixed, and PRESTO data may produce representative data of transit ridership in GTHA. The following projects can be carried out:

- ***Generate OD matrices***: Use boarding and inferred alighting locations to produce OD matrices, then compare with OD matrices from TTS for people with transit passes. In this case, PRESTO data serves as "ground truth" to "correct" the inaccuracies of TTS data once the cardholder and the survey respondent are matched using statistical matching techniques.
- ***Understand the transit system's demand***: PRESTO data offers a continuous record of the temporal change of travel demand. Constructing OD matrices from different "time-points" reflects transit system's demand during all operating hours and days. Comparing relationships between overall travel and transit demand, as elaborated in Regt et al. (2017) can be beneficial for transit planning.

8.3 Time-use Survey

Statistics Canada conducts a General Social Survey on Time Use every 5 years, with the latest one in 2015 (Statistics Canada, 2017). Questions such as distance from home to work, private vehicle ownership, travel mode for work, and frequency of congestions during daily commute are of interest (Statistics Canada, 2015). This survey also contains questions related to accessibility (e.g. disability that limits daily activities/trips), and perception of time (e.g. if respondents feel stressed due to limited time). These are helpful for categorizing respondents for modelling purposes on mode choice. One study can be carried out to access the suitability of this survey for TTS 2.0:

- ***Determine the fusibility between 2015 Time Use Survey and 2016 TTS***: Time use survey possesses rich temporal data whereas TTS has detailed spatial information. Moreover, they have socio-demographic variables (potentially) in common. By accounting for the temporal lag between them, the possibility of statistical matching should be evaluated.

8.4 Cellint Cellular Data

Cellint Traffic Solutions is a leading provider of real-time road traffic information based on cellular data. In 2016, Cellint partnered with Rogers, one of the largest cellular carriers in Canada, to launch traffic information service for car and navigation companies, as well as government agencies (Cellint Traffic Solutions, 2016). Cellint offers technologies such as TrafficSense and NetEyes to provide travel information with street-level accuracy by matching live cellular data with cellular signaling patterns of the routes. The output such as OD has several accessible formats for further analyses:

- ***Acquire and process data from Cellint, match records with census***: Use OD data to infer people's home and work locations, as well as travel mode and trip purpose. Match each record with the census to determine socio-demographic parameters.
- ***Compare spatial and temporal trip distributions between Cellint and TTS, and socio-demographic parameters***: After pre-processing, compare Cellint data with TTS 2016 survey results. It is worth examining the temporal distribution of trips, and spatial distribution of OD pairs to see how well TTS captures the actual trip activities. Moreover, sociodemographic information of both datasets can be compared to determine how well each dataset covers the actual population (use census as a reference).

8.5 StudentMoveTO Data

In autumn 2015, Toronto's four universities conducted a web-based travel survey for their students, aiming to find out students' travel patterns and factors that influence how they schedule work, studies, and daily activities (StudentMoveTO, 2016).

- ***Correcting under-sampling bias of TTS 2.0, if exists***: Due to high coverage to young populations, StudentMoveTo data can be used to correct under-sampling bias, if exists in TTS. Adjusting weighting factors and incorporating into TTS core survey data, as described in Verreault & Morency (2016), may be of interest.
- ***Feasibility of becoming the satellite survey of TTS 2.0***: More in-depth questions can target the student group as a satellite component of the next TTS.

9 CONCLUDING REMARKS

This report examines common data fusion technique and applications, and presents recommendations for several hypothetical avenues of research specifically relevant to the TTS. Applying data fusion techniques in the context of travel survey is motivated by the proliferation of "big data" and the flexible core-satellite structure proposed by Goulias et al. (2013). The end goal of data fusion is to shorten or replace the traditional, cross-sectional travel survey while combining multiple data sources (e.g. passively collected data and census) to generate something as comparable or even superior to traditional travel survey data. However, real-life applications are very rare due to the challenges outlined below:

- Datasets can be incompatible at several levels (e.g. spatial, temporal, and semantic). This is particularly true as most datasets are not designed for integration.
- Methods to resolve data incompatibility are difficult to generalize as each dataset is unique and should be treated on a case-by-case basis. Conceptual ideas in Section 3.1 such as aggregating zones or variable categories are not the optimum solution as this affects data resolution.
- There is a lack of fine measurement of the data fusion outcome. Four levels of validity proposed by Rässler (2002) are a good starting point but they also lack granularity. Statistical measurement such as t-test is only provided for the lowest level of validity (i.e. preserving marginal distributions), but no specific tests are recommended for other levels.

Due to the above reasons, most of the real-life examples as described in Section 7 are not strictly data fusion, though many claimed they are. The study on directly matching smart-card records with travel survey smart card holders (Spurr et al., 2015), and the study on enriching smart-card information with census via simulation (Grapperon et al., 2016), are the ones that best approximate the data fusion processes and, therefore, deserve careful attention. Other studies, such as using travel survey data to define rules to interpret, and to some extent "enrich" passive datasets, do not physically "fuse" data but nevertheless offer interesting results and possibilities on how multiple datasets can be utilized together.

Many traditional data fusion techniques, as described in Section 4, are available, but most are rigorously discussed at the theoretical level and few are put into actual use. Partially caused by the complexity and incompatibility of real-life datasets, there is no specific statement on when a method is the recommended one under certain situations. As a result, further research on evaluating the feasibility and applicability of each method should be conducted.

As for the next TTS in 2021, several data sources that can be incorporated into the core-satellite structure are outlined in Section 8. Most of them should be readily accessible. Conducting data fusion exercises with these datasets, such as fusing TTS 2016 and 2015 StudentMoveTo data, will offer a deeper insight into the potential difficulty/gaps discussed in this report and a clearer direction for future research.

10 BIBLIOGRAPHY

Bachmann, C., 2011. *Multi-sensor Data Fusion for Traffic Speed and Travel Time Estimation,* s.l.: s.n.

Bachmann, C., Roorda, M. J., Abdulhai, B. & Moshiri, B., 2013. Fusing a bluetooth traffic monitoring system with loop detector data for improved freeway traffic speed estimation. *Journal of Intelligent Transportation Systems,* 17(2), pp. 152-164.

Bayard, C., Bonnel, P. & Morency, C., 2009. *Survey mode integration and data fusion: Methods and challenges.* s.l., Emerald Group Publishing Limited, pp. 587-611.

Beckman, R. J., Baggerly, K. A. & McKay, M. D., 1996. Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice,* 30(6), pp. 415-429.

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

Breiman, L., Friedman, J. H., A, O. R. & Stone, C. J., 1984. *Classification and Regression Trees.* s.l.:Wadsworth.

Cellint Traffic Solutions, 2016. *Cellint to Launch Traffic Information Service in Canada with Rogers.* [Online]
Available at: http://www.cellint.com/success-stories/cellint-to-launch-traffic-information-service-in-canada-with-rogers/
[Accessed 19 August 2017].

Cowell, R., David, A., Lauritzen, S. & Spiegelhalter, D., 1999. *Probabilistic networks and export systems.* New York: Sprinter-Verlag.

Data Management Group, 2014. *2011 TTS Design and Conduct of the Survey,* Toronto: s.n.

Deming, W. E. & Stephan, F. F., 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics,* 11(4), pp. 427-444.

Denk, M. & Hackl, P., 2004. Data integration: Techniques and evaluation. *Austrian Journal of Statistics,* 33(1&2), pp. 135-152.

Devillanine, F., Munizaga, M. & Trépanier, M., 2012. Detection of activities of public transport users by analyzing smart card data. *Transportation Research Record: Journal of the Transportation Research Board,* pp. 48-55.

D'Ambrosio, A., Aria, M. & Siciliano, R., 2007. Robust Tree-Based Incremental Imputation Method for Data Fusion. *IDA,* pp. 174-183.

D'Orazio, M., 2013. *Introduction to Statistical Matching.* [Online]
Available at: http://en.eustat.eus/productosServicios/datos/Seminario_55_Slides.pdf

D'Orazio, M., Di Zio, M. & Scanu, M., 2006. *Statistical Matching: Theory and Practice.* s.l.:John Wiley & Sons.

Fellegi, I. & Sunter, A., 1969. A theory for record linkage. *Journal of the American Statistical Association,* Volume 64, pp. 1183-1200.

Goulias, K., Pendyala, R. & Bhat, C., 2013. Keynote - total design data needs for the new generation large-scale activity microsimulation models. *Transport survey methods: best practice for decision making,* pp. 21-46.

Grapperon, A., Farooq, B. & Trepanier, M., 2016. *Information fusion of smart card data with travel survey,* s.l.: (No. CIRRELT-2016-49).

Habib, K., Swait, J. & Salem, S., 2014. Using repeated cross-sectional travel surveys to enhance forecasting robustness: Accounting for changing mode preference. *Transportation Research Part A: Policy and Practice,* Volume 67, pp. 110-126.

Hollis, R., 2017. *PRESTO quarterly report.* [Online]
Available at:
http://www.metrolinx.com/en/docs/pdf/board_agenda/20170628/20170628_BoardMtg_PRESTOQuarterly_Report_EN.pdf
[Accessed 19 August 2017].

Howell, D., 2007. *The Analysis of Missing Data,* s.l.: The Sage handbook of social science and methodology.

Kiesl, H. & Rässler, S., 2006. *How valid can data fusion be?*. [Online]
Available at: http://doku.iab.de/discussionpapers/2006/dp1506.pdf
[Accessed 25 July 2017].

Kusakabe, T. & Asakura, Y., 2014. Behavioural data mining of transit smart card data: A data fusion approach. *Transportation Research Part C: Emerging Technologies,* Volume 46, pp. 179-191.

Leulescu, A. & Agafitei, M., 2013. *Statistical matching: a model based approach for data integration,* s.l.: Eurostat European Commission.

Lin, T., Hasnine, S. & Habib, K., 2017. *Influence of latent attitudinal factors on the level of multimodality of post-secondary students in Toronto,* s.l.: s.n.

Lo, A., 2017. *Impact of multiple survey frames on data quality of household travel surveys: the case of the 2016 Transportation Tomorrow Survey,* s.l.: s.n.

Long, Y. & Thill, J. C., 2015. Combining smart card data and household travel survey to analyze jobs-housing relationships in Beijing. *Computers, Environment and Urban Systems,* Volume 53, pp. 19-35.

Mahmoud, M. S., Habib, K. & Shalaby, A., 2015. *Application of RP/SP data to the joint estimation of mode choice models: lessons learned from an empirical investigation into cross-regional commuting trips in the Greater Toronto and Hamilton Area,* s.l.: Transportation Research Board 95th Annual Meeting.

McCullagh, P. & Nelder, J. A., 1989. *Generalized linear models*. London: Chapman and Hall.

Miller, E. et al., 2014. *Changing Practices in Data Collection on the Movement of People,* s.l.: Transportation Research Board.

Ministry of Transportation Ontario, 2016. *About.* [Online]
Available at: http://tts2016.ca/en/about.php
[Accessed 19 November 2016].

Nakamya, J., Moons, E., Koelet, S. & Wets, G., 2007. Impact of data integration on some important travel behavior indicators. *Transportation Research Record: Journal of the Transportation Research Board (1993),* pp. 89-94.

Namazi-Rad, M. et al., 2017. An unconstrained statistical matching algorithm for combining individual and household level geo-specific census and survey data. *Computers, Environment and Urban Systems,* Volume 63, pp. 3-14.

Prichard, D. R. & Miller, E. J., 2011. Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultanrously. *Transportation,* Volume 39, pp. 685-704.

Rässler, S., 2002. *Statistical matching: A frequentist theory, practical applications, and alternative bayesian approaches.* New York: Springer-Verlag.

Regt, K. d., Cats, O., Oort, N. v. & Lint, H. v., 2017. *Investigating potential transit ridership by fusing smartcard and GSM data,* Delft: s.n.

Rodgers, W. L., 1984. An evaluation of statistical matching. *Journal of Business & Economic Statistics,* 2(1), pp. 91-102.

Roszka, W., 2015. Some practical issues related to the integration of data from sample surveys. *Statistika: Statistics and Economy Journal,* 95(1), pp. 60-75.

Rubin, D. B., 1986. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics,* 4(1), pp. 87-94.

Rudra, M., 2014. *Application of Imputation Methods in the Analysis of Freight Trip Generation in the Greater Toronto and Hamilton Area,* s.l.: s.n.

Salem, S. & Habib, K., 2015. Use of repeated cross-sectional travel surveys to develop a Meta model of activity-travel generation process models: accounting for changing preference in time expenditure choices. *Transportmetrica A: Transport Science,* 11(8), pp. 729-749.

Saporta, G., 2002. Data fusion and data grafting. pp. 465-473.

Scanu, M. & Tuoto, T., 2008. *ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data,* s.l.: s.n.

Shen, L. & Stopher, P. R., 2013. A process for trip purpose imputation from Global Positioning System data. *Transportation Research Part C: Emerging Technologies,* Volume 36, pp. 261-267.

Spurr, T., Chu, A., Chapleau, R. & Piché, D., 2015. A smart card transaction "travel diary" to assess the accuracy of the Montrèal household travel survey. *Transportation Research Procedia,* Volume 11, pp. 350-364.

Statistics Canada, 2015. *General Social Survey on Time Use, 2015.* [Online]
Available at:
http://www23.statcan.gc.ca/imdb/p3Instr.pl?Function=assembleInstr&lang=en&Item_Id=217656#qb218233
[Accessed 19 August 2017].

Statistics Canada, 2017. *2015 Time Use Survey Technical Note.* [Online]
Available at: http://www.statcan.gc.ca/pub/89-658-x/89-658-x2017001-eng.htm
[Accessed 19 August 2017].

Statistics Canada, 2017. *Canadian Income Survey (CIS)*. [Online]
Available at: http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5200
[Accessed 19 August 2017].

Statistics New Zealand, 2015. *Data integration manual: 2nd edition*. [Online]
Available at: http://m.stats.govt.nz/methods/data-integration/data-integration-manual-2edn.aspx

StudentMoveTO, 2016. *StudentMoveTo, An overview of early findings*. [Online]
Available at: http://www.studentmoveto.ca/wp-content/uploads/2016/04/StudentMoveTO.Handout_4Uni.v2.pdf
[Accessed 19 August 2017].

Train, K., 2002. *Discrete Choice Methods with Simulation*. s.l.:Cambridge University Press.

Trépanier, M., Tranchant, N. & Chapleau, R., 2007. Chapleau, Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems ,* 11(1), pp. 1-14.

Van der Lann, P., 2000. Integrating administrative registers and household surveys. *Netherland Official Statistics,* Volume 15, pp. 7-15.

Verreault, H. & Morency, C., 2016. Integration of a phone-based household travel survey and a web-based student travel survey. *Transportation,* pp. 1-15.

Willenborg, L. & Heerschap, H., 2012. *Method Series. Theme: Matching*. [Online]
Available at: https://www.cbs.nl/NR/rdonlyres/0EDC70A4-C776-43F6-94AD-A173EFE58915/0/2012Matchingart.pdf

Willenborg, L. & van der Laar, R., 2014. *Object Matching (Record Linkage)*. [Online]
Available at: https://ec.europa.eu/eurostat/cros/content/micro-fusion_en