NCHRP 8-43 – Tutorial 5: Merging Data

J. A. Blair* & A. Lu
Reebie Associates, 101 Rogers St., Suite 403
Cambridge, Mass. 02142-1068

* Corressponding Author, e-mail *jblair@reebie.com*

1. Introduction

Today's planners have a staggering array of data available for analysis. Seldom however, do the characteristics of the data available exactly match the needs of the study. While it is sometimes reasonable to utilize available data as a proxy for the desired information, it is also sometimes possible to merge two or more distinct data sources into a common file that contains more or all of the desired elements.

Such a process can be conducted in a statistically rigorous fashion to provide detailed and accurate results, or it can merely provide a "direction finder", providing information that is not statistically accurate, but nevertheless insightful.

Reebie Associates has conducted numerous analyses that demand the construction of customized databases. One of the most common examples is the merge of commodity flow data from various sources to build more complete databases. Another is the construction of regionally detailed data by merging national transportation data with local population or industrial activity metrics.

1.1 Merging Data

Merging data often means trying to combine data from different sources, collected by different methods, covering different geography, and reporting different features. Each of these issues adds to the complexity of merging data, but do not make such constructs impossible. The successful merging of data requires that planners understand several basic principals of data analysis.

1.1.1. Principal Elements:

1.1.1.1 Data granularity

The first of these principal elements is *data granularity.* Data granularity represents the level of detail embedded in the dataset. A dataset might report to cover North America, or Cook County, IL. It might claim to include all rail traffic but only selected portions of truck traffic. These elements of data granularity are critical to understanding the data merge process.

An effective data merge requires that all data sets reflect similar granularity. Very often base datasets need to be adjusted in preparation of a merge. One dataset might have to be made less detailed, or perhaps only a portion of the data can be used in the merger. It is often the case that an interim dataset must be constructed to allow a data merge.

How closely matched are the two datasets in terms of their specificity often determines the believability of the merge. It is not reasonable to merge single county-level results with national volume data.

Merging often reduces data specificity. The frequent result of merging two independent datasets is that the overall result is less specific than the original inputs. This makes intuitive sense as the compounding of error in statistical analysis. The more independent measures combined, the less likely the combined results are correct. This is not to say that data merges are invalid, but rather to highlight the fact that there is a statistical erosion of accuracy at a detailed level for combined datasets.

Does the merged data provide the necessary specificity to conduct the analysis? It is possible that in order to merge databases that the granularity is no longer usable for the analysis. This can be the case where in order to merge to datasets; the aggregation of data makes the final results meaningless. For example, merging industrial data using Standard Industrial Codes (SICs) and North American Industrial Classification System (NAICS) codes, requires that data be aggregated to a 2-digit SIC level. If the analysis seeks to understand the nuances of manufacturing sector trends, the 2-digit SIC level does not provide sufficient detail to identify man of the sub-classes of the manufacturing sector.

1.1.2. __Index Criteria__

The index criteria are the elements in ***both*** files that are the same that can be used either for check totals or as the basis for the merge. Here as with the data granularity issue, the index criteria must match at some level in order to make the merge feasible. Planning the data merge requires that analysts consider what field or fields can be matched between the two data files, and then insuring that the files are aggregated or disaggregated in such a was as to permit comparison. As an example, a zip code level dataset cannot be readily merged to a county-level dataset because there is no common geographic boundary between zip codes and counties: counties extend beyond zip code boundaries, and zip coded beyond county boundaries. Area codes are silmilarly diffcicult to combine with any non-area code based dataset because there is no common boundary between area codes and any jurisdictional borders. Conversely, a county-level dataset and a state level dataset can be more easily combined because counties do not cross state boundaries.

The more index criteria that can be established between the two data sources, the more likely that the merged data will provide meaningful insight. Often data merges are conducted with but one common index criteria. Much preferred however, is a multi-factor merge, using two or more index criteria between two dissimilar datasets. For example, commodity flow data and industrial data can readily be combined using a common geographic index, but the combination is likely to provide significantly more insight and accuracy when geography and industry classifications are utilized. Such a multi-factor merge reduces the likelihood of finding “coal going to New Castle” or Christmas trees shipped from Arizona.

In practice, it is not always possible to develop multi-factor data merges. As an alternative, planners can often utilize check totals and a detailed review to measure the validity of the data combination.

1.1.1.2 Data Integrity

Before investing the sometime substantial time to conduct the data merge, analysts should consider how "believable" the combined dataset would be to the audience. Merging the results of a poorly conducted survey with detailed U.S. Census data does not add legitimacy to the survey results. If the source data are spurious, the resulting data merge will not be credible.

Maintaining the integrity of the data in a data merge requires that; (1) the source data is viewed as credible; and (2) that the data merge process is credible.

The validity of the source data determines the validity of the merged data. As outlined above, the merge process frequently reduces the validity of the original source data. If that data was improperly collected, the merge results are similarly tainted. It is important for analysts to fully vet the data they intend to utilize for a merge process. Understanding the source data's strengths and weaknesses allows planners to know what portion of the results are defendable, and what portions are suspect. Knowing the source data also allows analysts to plan their merge around any weaknesses. Using a weak portion of the source data an index criteria casts doubt on all results, no matter how well the process was designed and executed.

The validity of the merge process determines the validity of the combined data. Merging data is often not a statistical process. There are assumptions and choices made throughout the data merge process that can undermine the legitimacy of the results. It is often wise to discuss the planned merge process with experienced data manipulators such as Reebie Associates. Such consultation may identify better methods or data sources for the proposed merge.

It is also important to maintain process records for the data merge. Methodological decisions and assumptions are easily forgotten in the excitement of a successful merge. The ability to reproduce the results insures that operator or machine error did not corrupt legitimate data processing techniques.

2. Some Analytic Examples:

One of the best ways to understand both the capabilities and the methodologies of data merging is to review a series of actual examples. Those selected for this discussion reflect the following scenarios:

- Merging commodity flow and import/export data.

- Merging commodity flows and detailed truck configuration data.

- Merging regional survey results with detailed Census data.

In each of these examples, the source data are identified, along with the specific data elements from each that were used in the analysis. There is an explanation about the source data granularity and index criteria chosen and what steps had to be taken to match the different data sources to affect the merge. Finally, there is a discussion about the issues involved in the data merge, and the output results of the process.

1.2. Inland Flow of International Goods Database

1.2.1. Goal of the Analysis

In this project, a private sector transportation provider approached Reebie Associates to develop an estimate of international trade cargo flows by port, moving to inland points by truck and rail intermodal. The client was seeking to identify specific market opportunities to provide inland distribution services to international steamship carriers.

The transportation provider wanted specifically to identify the volume of containerized international traffic that was moving in individual domestic trade corridors, and further, to identify those ocean shipping companies that were dominant in each lane of the corridor trade. This dataset would then be used to target sales and marketing opportunities among international carriers. The client sought to identify the market shares of the multiple carriers operating in common lanes, and the total volume of international trade that could be combined with the provider's current domestic rail and truck flows.

1.2.2. Source Data

Unable to identify a ready source of data that would identify international container volumes by origin and destination, mode, and container line, Reebie Associates staff set out to construct a custom dataset that modeled these features. Data would need to provide individual origin-destination flows by selected modes, segmented for international and domestic freight. In addition, the international freight flows would need to be further segmented by market share for individual steamship carriers. While no single dataset provided all these features, the staff at Reebie Associates was able to identify multiple source data files that contained the critical elements. If these data could be merged effectively, the combined dataset could provide the level of insight sought by the client for his marketing efforts.

The source data selected for this analysis consisted of two primary research datasets: Reebie Associates TRANSEARCH 2003 domestic database; and a partial year extract of international container freight from Commonwealth Business Media's PIERS dataset[1].

[1] The client, to gain market insight in the international container trade, had previously purchased the partial year dataset.

Reebie Associates' TRANSEARCH database contains origins and destinations of freight, the quantity and commodity mix of that traffic, and the distribution of modes. TRANSEARCH reports to cover U.S. domestic, NAFTA trade, and major elements of inland seatrade activity. TRANSEARCH generally does not however, uniquely identify international data from domestic production.

In this analysis, TRANSEARCH provided several critical elements:

- Commodity detail – used to separate "containerizable" freight from "non-containerizable" freight.
- Mode detail – to separate rail intermodal and truckload volumes
- Origins and destinations – to identify inland origin/destination linkages for international freight volumes.

The Port Import Export Reporting Service (PIERS) database is maintained by Commonwealth Business Media, publishers of the Journal of Commerce. The PIERS database provides the most comprehensive body of international trade data for ports in the United States, Mexico and South America2. But because the data is primarily intended to reflect port-to-port trade activity, only small portions of the data reflect inland origins and destinations. Likewise, primary data sources capture port of entry and exit detail, but cannot accommodate the surfeit of rebill, reconsignment, transload, and break-bulk techniques employed in international trade.

In this analysis, the PIERS dataset provided several critical elements:

- Volume – total international container volumes U.S. port of entry/exit
- Geography – container volumes by U.S. port of entry/exit
- Container shipping line – volume of containers per container shipping line for each U.S. port of entry/exit

The available PIERS dataset contained information for eight months. These volumes were annualized without regard to seasonality[3].

[2] http://www.piers.com/

[3] The PIERS database was made available for the analysis, but was not an ideal match to the TRANSEARCH dataset. There was only limited overlap in the time period covered, and the available PIERS data contained only eight months worth of data.

Had a PIERS and TRANSEARCH datasets been acquired specifically for this data merge, the specifications for the two datasets would have been matched for time coverage, and a better result anticipated.

1.2.3. Data Granularity

2.1.1.1 Geography

In this example, both the PIERS and the TRANSEARCH datasets were "national" in scope. In other words, both reported to reflect the totality of international traffic moving into and out of the United States. But while the PIERS data identified containerized volumes at an origin and destination port level, the Reebie TRANSEARCH data identified origins and destinations at a county (FIPS) level. The inconsistent granularity of the data meant that wherever two port facilities were located in the same county, the relevant origin and destinations could not be individually separated. This occurred in only one instance -- Los Angeles County, California, which contains two container ports.

While TRANSEARCH reports truck volumes at a county level, rail intermodal volumes can be reported only at a state-portion of BEA[4] level. This broader geography meant that inland rail distribution would be modeled after these larger geographic areas, potentially increasing the risk that domestic traffic patterns would undermine any identification of international freight movement. This factor impacted many more regions, including Northern New Jersey, New York, Philadelphia, Norfolk, Los Angeles, San Francisco, Seattle, and several others.

2.1.1.2 Volume

The available PIERS dataset reflected containerized volumes only: no break-bulk cargoes were included. Conversely, the TRANSEARCH database contained all manufactured commodities, most raw materials, and many unprocessed agricultural products. The merging of the two datasets required that they be made compatible with respect to covered commodities.

The TRANSEARCH data reports flows in tons by commodity, mode, and geography. The PIERS data reports volumes in twenty-foot-equivalent (TEU) container loads by geography. Bringing the two datasets into alignment for merging required that a common volume metric be selected [forty-foot–equivalent (FEU) container loads was adopted], and all data be converted to this format.

The PIERS data was converted from FEUs to TEUs, by dividing reported volumes by a factor of two5. TRANSEARCH data was converted to FEUs based on historical weights per

[4] BEA descriptive footnote.

[5] Since volumes were reported in "equivalent units" it was deemed appropriate that a 2:1 conversion factor was reasonable. Had volumes been reported in actual twenty-foot or forty-foot container volumes, the conversion process would have been more vexing. Because of their lower net weight limit per foot, one forty-foot container frequently cannot accommodate heavy product contained in two twenty-foot containers. It was assumed that the conversion of reported tons to TEUs, averaged the net weights of all container volumes over the total reported, thus providing a blended product density that could be effectively loaded in forty-foot containers.

cubic foot for each of commodities identified. Eliminating that volume of traffic that is not generally containerized6 further filtered these FEU totals.

Acknowledging the limitations of the commodity and geography conversions, and the risks of data infection resulting from the presence of domestic traffic in the database, the client elected to continue the modeling exercise.

1.2.4. Index Criteria

With the PIERS and TRANSEARCH datasets finally matched in terms of geographic and volume metrics, there was a need to identify a common factor in both datasets that could be used to link them together. While multiple index factors are always preferable, there was identified only a single element in common between the PIERS and Transearch datasets: the location (county FIPS or BEA) of the port-of-entry/port-of-departure.

In the PIERS dataset, the location of the port served as the origin (port-of departure) or the destination (port-of-entry) of an international container movement. In the Transearch dataset, the location of the port served as the destination (port-of-departure) or the origin (port-of-entry) for an inland move. By merging the two databases, we could suggest the possible volumes, modes, originations, and destinations of the inland movement of international freight. Furthermore, by identifying the portion of the traffic at the port of entry or departure attributed to a specific ocean carrier (PIERS), we could approximate that carrier's market share in individual inland trade corridors.

Securing this information was the ultimate goal of the client, who sought to market distribution products and services to these ocean carriers in specific inland trade corridors

1.2.5. Merging Process and Results

Utilizing the port location data (specific port names converted to county [FIPS] or BEA [state-portion BEA for rail]) found in the PIERS dataset, inland origins and destinations were developed in proportion to the trade patterns identified for the combined domestic and international volumes found in the TRANSEARCH dataset. The total volumes distributed were those reflected in the PIERS database, and provided an estimate of international trade volumes in individual corridors. Residual volumes were logically attributed to domestic traffic sources. In no instance did the PIERS reported volume of international freight exceed the reported total of international and domestic freight in TRANSEARCH[7]. The resulting data merge provided a national database of flows, segmented by mode, for international and domestic data. Within the international data,

[6] This is accomplished using a proprietary "containerization screening" technique developed by Reebie Associates that seeks to identify that proportion of a commodity classification grouping that can be transported in containers, or dry-van type trailer equipment.

[7] This might have occurred, since freight can often move locally, from the port to a location in the same county. As Transearch does NOT (ALWAYS) capture intra-county movements, such volumes would be missing from the dataset, and thus cause a total domestic and international volume to be lower than the reported international total.

the individual flows were further broken out to provide market shares by individual steamship lines.

The merged database allowed the client to evaluate and prioritize inland lanes for the introduction of new service products geared to international steamship carriers, and to identify the specific carrier to whom these services would be offered. They were further able to estimate what portion of the freight would favor rail intermodal products, and what percentage preferred truck-based solutions. The merged dataset was able to provide significant insight into the opportunity to position service products for inland trade without the expense or difficulty of a detailed survey process.

1.3. FHWA Congestion Database

1.3.1. __Goal of the Analysis__

In this project, the Federal Highway Administration (FHWA) approached Reebie Associates to develop a model to estimate the composition of truck traffic on various segments of the National Highway System (NHS). FHWA was familiar with the data contained in Reebie Associates TRANSEARCH, but sought to take this data further. They wanted to identify the composition of the truck traffic by a number of standard classifications including: singles (straight trucks); semi-trailers, double trailer and triple trailer combinations in order to develop detailed highway congestion forecasts. The FHWA also wanted to sub segment the information in the "Congestion Database" by axle configurations and gross vehicle weights (GVW) to permit better highway maintenance planning performance.

1.3.2. __Source Data__

The desired data elements for the Congestion Database were not available in any one database, but could be developed using multiple databases. Reebie Associates staff undertook to develop a merged dataset containing elements of commodity flow *and* detailed truck characteristics. The final Congestion Database would need to provide individual origin-destination highway flows, routed across the national highway network, segmented by truck body types and configurations. If these data could be merged effectively, the combined dataset could provide the level of insight sought by the FHWA to improve highway maintenance planning performance.

The source data selected for this analysis consisted of two primary research datasets: Reebie Associates TRANSEARCH 1998 domestic database; and the U.S. Census Bureau's Vehicle Inventory and Use Survey (VIUS) 1997 edition.

Reebie Associates' TRANSEARCH database contains origins and destinations of freight, the quantity and commodity mix of that traffic, and the distribution of traffic by modes and submodes. TRANSEARCH reports to cover U.S. domestic, NAFTA trade, and major elements of inland seatrade activity. Other than identifying truck traffic as For-hire Truckload (TL), Less-than-truckload (LTL), and Private Truck (PVT), TRANSEARCH does not routinely provide any information on truck body types or configurations.

In this analysis, TRANSEARCH provided several critical elements:

- Commodity detail – used to identify likely truck body types and possible configurations based on the characteristics of individual commodities.

- Mode detail – to separate TL + PVT from LTL volumes to help define truck body types and configurations.

- Origins and destinations – to identify highway routing across the national network of highways.

"The Vehicle Inventory and Use Survey (VIUS) provides data on the physical and operational characteristics of the Nation's truck population. This survey is conducted every 5 years as part of the economic census."[8]

"The 1997 VIUS is a probability sample of private and commercial trucks registered (or licensed) in the United States as of July 1, 1997. This survey excludes vehicles owned by Federal, state, or local governments; ambulances; buses; motor homes; farm tractors; unpowered trailer units; and trucks reported to have been sold, junked, or wrecked prior to July 1, 1996. A sample of about 131,000 trucks was surveyed to measure the characteristics of nearly 75 million trucks registered in the United States"[9]

In this analysis, VIUS provided several critical elements:

- Vehicle registration and operation – VIUS links truck characteristics to state vehicle registrations, and provides estimates on in-state and out-of-state operations.

- Truck body types and characteristics – VIUS presents detailed information about truck body types (dry van, reefer, flatbed, auto carrier etc) and configurations (three axle tractor, two axle trailer, double trailer, straight truck etc.).

- Gross and net weight estimates – VIUS presents average gross and net weights by trailer body type and configuration.

- Average loaded and empty miles – VIUS provided estimates of average loaded and empty vehicle miles annually. These figures help determine truck utilization figures and repositioning patterns.

[8] U.S. Census Bureau

[9] Ibid.

1.3.3. <u>Data Granularity</u>

2.1.1.3 Geography

In this example, both the VIUS and the TRANSEARCH datasets were "national" in scope. In other words, both reported to reflect the totality of domestic truck traffic. But while the VIUS data identified vehicle characteristics at the state level, the Reebie TRANSEARCH data identified origins and destinations at a county (FIPS) level. As TRANSEARCH requires this FIPS–level geography to facilitate highway routing, the data needed to remain at a county-to-county level for processing even though the merge process could only take place at a statewide level.

2.1.1.4 Definition of "Truck"

With its emphasis on for-hire and private truck movements, and its exclusion of government, farm, and recreational vehicles, VIUS provides a fairly comparable scope to TRANSEARCH freight flow data. In several areas however, VIUS provides more comprehensive coverage for vehicles not routinely included in TRANSEARCH.

For example, the VIUS database includes passenger vehicles in commercial activity (station wagons, SUVs or light trucks in local and Postal delivery service). TRANSEARCH does not report to capture these volumes in any comprehensive fashion. Thus to merge these datasets without accounting for this different definition of a "truck" would risk allocating a substantial volume of freight activity to these small delivery vehicles, and to understate the activity of larger commercial vehicles on the highways. Thus, the definition of "truck" in VIUS was adjusted to more closely reflect the body of traffic captured in TRANSEARCH.

In addition, VIUS includes coverage of numerous vehicles that are not involved in the transportation of "freight" as defined in TRANSEARCH. VIUS data covers vehicles transporting municipal waste, utility poles, local appliance and furniture delivery, etc. These movements are not routinely captured in TRANSEARCH, and had to be removed from the VIUS data prior to merging.

1.3.4. <u>Index Criteria</u>

With the VIUS and TRANSEARCH datasets finally matched in terms of geography and coverage, there was a need to identify common elements in both datasets that could be used to link them together. Both VIUS and Transearch contain commodity information; VIUS at a level that approximates 2-digit SIC and Transearch at 4-digit STCC level. By using a series of conversion tables, we were able to process both datasets at a 2-digit STCC level.

Second, while it is not a regular part of TRANSEARCH distributions, Reebie Associates maintains a series of truck body-type algorithms that attribute TRANSEARCH truck tonnages to different truck types. While not as detailed as those contained in VIUS, these body-types provided an important link between the two datasets.

Third, there is a geographic link between the two datasets. VIUS reports truck characteristics based on the state of registration, or more importantly, the base state of operation. Inferring that the trucks are based in a state to service the needs of the local economy, we assumed that the base state truck profiles were a reasonable proxy for the types of trucks making pick-ups in that state. Thus the base state served as an index field to TRANSEARCH's origin state volumes, and provided a critical geographic linkage that made the merging of the databases possible.

Using the various index criteria, Reebie Associates was able to link the two datasets to construct a national goods movement database that provided the detailed truck body-type and configuration detail sought by the FHWA for its congestion analysis.

1.3.5. <u>Merging Process and Results</u>

For this analysis, truck tonnages were extracted from Reebie Associates 1998 TRANSEARCH database, including Truckload Tons (TL), Less-than-Truckload tons (LTL), and Private truck tons (PVT). This volume represents approximately 70% of the total tons in the TRANSEARCH Database. The TL and PVT tons were combined as representative of total truckload traffic, while the LTL tons were kept separate for special processing.

2.1.1.5 Allocation of TRANSEARCH Tons to Body Types and Configurations

Reebie Associates then allocated truck tonnage to seven basic truck body types, including: Dry Van, Reefer, Flat, Automobile, Bulk (including hoppers and open-top gondolas), Tank, and Livestock. For each four-digit STCC found in TRANSEARCH, Reebie Staff developed a distribution of traffic to the specific body types. Some STCC's were attributed to one body-type (STCC 2516 – Children's Furniture – allocated 100% to Dry Van), while others were distributed to two or three different body types type (STCC 24396 – Structural Wood Products -- allocated 60% to Dry Van, 40% to Flat). This allocation was developed using the commodity information in the VIUS database, and the Reebie Associates proprietary model.

The second step involved developing distributions by configuration for each body-type. For this part of the analysis, Reebie staff again used the VIUS database to develop the various allocations. For the truckload portion of the analysis, we classified the body-types into twelve different truck configurations based on number of trailers and axles. These configurations were Straight Trucks, CS3 (Combination Vehicle, Semi-Trailer, 3 total axles), CS4, CS5, CS6, CS7, DS5 (Double, Semi-Trailer, 5 total axles), DS6, DS7, DS8, DS9+, TS7+ (Triple, Semi-Trailer, 7 or more axles). Reebie constructed configuration distributions for each body-type based on the state of origin of the traffic and two distributions for Dry Van based on state of origin and commodity handled.

For the LTL volumes in the TRANSEARCH database, Reebie developed a separate configuration distribution based on information obtained from VIUS on trucks dedicated to LTL service. This too was based on state of origin, but did not include the secondary analysis of commodity in that such an analysis would have been inappropriate for LTL (mixed commodity) traffic. Short haul, secondary, and very light density traffic was likewise analyzed separately, and distributions were constructed that considered the unique aspects of these types of traffic.

At this point, the database contained origin and destination FIPS, STCC4, Body-type, configuration, and allocated tons.

2.1.1.6 Development of Loads and Empties

To derive loads from the file -- which at this point contained only tons – Reebie staff developed a series of average weight distributions for each body type, STCC and configuration using data derived from VIUS, and several Reebie proprietary models. These factors served as divisors for the tons contained in the TRANSEARCH file.

To develop empty movements,. Reebie staff developed a series of backhaul percentages based on truck body-type and configuration derived from data in VUIS. These represented the relative percentage of empty movements versus loaded movements for each body type and configuration alternative. These were then applied to each of the loaded movements as an estimate of empty truck trips. Empty routings reflected the reverse of the loaded movement.

2.1.1.7 Final Processing

The file was then "rolled-up" based to meet the specified criteria of FHWA's Congestion Analysis and included Origin and Destination FIPS, STCC4, Straight Truck Loads, Semi-trailer Loads, Double Trailer Loads, Triple Trailer Loads, and Empty Loads. This file was then routed across the national highway network to produce a detailed estimate of the composition of trucks (including their body types and configurations) on various highway segments across the nation.

The modeled results were tested against actual visual truck count data in several regions across the country, and found to be a reasonable proxy for detailed visual truck counts.

1.4. Downtown Package Express Database

1.4.1. Goal of the Analysis

In this project, the Chicago Department of Transportation (CDOT) and the Chicago Transit Authority (CTA) approached Reebie Associates to develop a model to estimate the volume of downtown package air freight that might be attracted to a premium service transit-based freight transportation product. The analysis would seek to identify regional package volumes, and then to sub segment the volumes into small city-block sized geographies. The results would be used to determine the size of the current downtown market for package air freight, and to determine – based on the characteristics and concentrations of the freight – what portion of the traffic could be diverted to a subway-based "Package Express" service between downtown and the City's two airports

1.4.2. Source Data

The desired data elements for the Package Express database were not available in any one database, but could be developed using multiple databases. Data was available on the number of establishments in the downtown district by zip code, and package air freight volumes were recorded at the Airport level. The Reebie Associates consulting team undertook to develop a custom dataset containing elements from each of these two datasets. The final Package Express database contained detailed information on air package freight volumes and provided this data as city-block-level geography.

The source data selected for this analysis consisted of two primary research datasets: Dun and Bradstreet's (D&B) Million Dollar Database of company information and U.S. Census Bureau 2000 employment data.

Dun and Bradstreet's *Million Dollar Database* contains information on nearly 1,600,000 businesses in the U.S. and Canada[10]. The information contained in the database includes: industry classification information (based on Standard Industrial Classifications or SICs) number of employees at the location, annual sales, ownership profile, principal executives, and contact information such as telephone and fax numbers. For this project, a database containing the total number of businesses by SIC and zip code was purchased, with full detail provided for a random and representative sample of these businesses.

For the Package Freight database, the *Million Dollar Database* provided several critical elements:

- Candidate business names – The D&B data used to identify firms that would be contacted telephonically to determine their air package shipping patterns.
- Business SIC – The identification of businesses by industry groupings and zip codes allowed Reebie Associates to interpolate air package results for similar businesses in the Chicago region.
- Business employment – Business employment levels provided a basis for scaling the air package results developed from the telephonic surveys.

The *Million Dollar Database* was used directly to provide control totals for the census data at a zip code level. It's primary purpose was however, to provide contact information for a telephonic survey of local businesses to determine to what extent these businesses were utilizing overnight and second-day package services. The survey developed some 300 responses which were then categorized based on industry SIC and employment levels to determine the scalability of the results to the other (unsurveyed) businesses in the region. It was ultimately these survey results that were merged with the Census data, while the original summary file provided zip code level employment and establishment control totals.

[10] www.dnbmdd.com

In this analysis, the U.S. Census data by Quarter Section[11] provided several critical elements:

- Number of Business and employment at a sub-zip code level – The Census data allowed us to identify businesses by type at a sub-zip-code level. This was necessary in that the rate of service adoption decreased as walking distance from the downtown terminal increased. Increments of two city blocks reflected about a five-minute walk, and thus allowed Reebie Associates to estimate service adoption in five-minute-walk increments – similar to the methodology used to estimate transit ridership levels.

- Business Industry Groupings – The Census data categorized businesses in North American Industry Classification System (NAICS) codes. We converted these codes to SICs to provide us with a critical link between the Census and the D&B datasets.

- Total number of businesses in the region – Ad the D&B database does not purport to capture all establishments in the region, the Census data provided us with a basis for scaling the D&B results to the Census reported total establishments in the region.

1.4.3. ***Data Granularity***

2.1.1.8 Geography

In constructing the Package Express database, Reebie Associates was careful to match geographies between the two datasets. For this reason, the consulting team selected zip-code-level geography for data gathering and matching totals, while the analysis was done at a quarter-section level.

A Census quarter section comprises a ½ mile square area in a metropolitan area. For Chicago, there are 1,083 quarter sections within the City's boundaries. These geographic segments can be overlaid against Zip Codes, but there is no direct correlation between the two geographies. Rather, Reebie Associates utilized individual quarter-section employment levels and zip code based business establishment and employment figures to associate the two. Through this process the consulting team was able to ascertain business establishments at a quarter-section level, and identify the volume of package freight available at half-mile intervals from the proposed downtown terminal location.

[11] The Census data was obtained through two sources: the Northern Illinois Planning Commission (NIPC) who produces current and forecasted population, household, and employment data at the quarter section level; and the U.S. Census bureau which produces data on the number of establishments at a zip code level. An important subprocess of the analysis was the creation of a merged database of Census data that provided employment and number of establishments by North American Industry Classification System (NAICS) codes, and by quarter section.

2.1.1.9 NAICS versus SIC Industrial Classifications

Whereas the D&B data reported establishments in the common Standard Industrial Classifications (SIC), the Census data was reported in North American Industry Classification System (NAICS) codes. These two classification schemes do not translate directly – in other words there is not a single SIC classification that matches a single NAICS code. To resolve this situation, Reebie Associates aggregated both the SIC and NAICS codes to permit reasonable conversion, and then matched employment levels to insure that the subtotals (and hence the conversions) were comparable. Some data manipulation was necessary to insure that all establishments and employment was included. The final analysis was conducted at a two-digit SIC level, and proved sufficiently detailed for the rigors of the analysis.

1.4.4. <u>Index Criteria</u>

To merge the information in both the Census and the D&B data files, a common field needed to be established. In the construction of the merged Package Express database, geography – specifically zip codes – was selected as the index criteria for the two files. Reebie Associates undertook significant processing to insure that a common list of zip codes was included in both databases, and that the comparable figures – such as total employment by zip code – were consistent. Taking the extra time at this point insured that the merging process would not produce un-assignable records, or inconsistent totals.

Although not used as an index field, the employment numbers between the two datasets were used as match totals – a double check against the accuracy of two datasets collected at different times, for different purposes, via different means.

1.4.5. <u>Merging Process and Results</u>

For this analysis, the survey results developed using the D&B data were merged with the Census data by SIC and employment levels, using zip codes as the common denominator.

The survey data tallied packages shipped by different industry types (SIC). These ratios were assigned to all similar industries in that zip code, and compared to similar industries in other zip codes. Likewise, the survey data was correlated against the firm employment level, and matched against similar forms in the Census data to provide a double check of totals.

When the total package volumes estimated for each zip code were relatively consistent between both methodologies, the consulting team had a higher confidence level in the overall results.

The final database contained: (1) estimated overnight, air freight, and air courier package volumes; (2) number of establishments by SIC; and (3) employment by SIC for each Census quarter section in the study region. This database was the foundation of a subsequent made choice analysis that allowed the consulting team to estimate the attractiveness of the proposed service based on the distance of the freight from the downtown terminal in two city block increments.