NCHRP 8-43 – Tutorial 1: Data Conversion

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

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

1. Introduction

Data sources may be collected in a variety of formats using a variety of classification systems for industrial activity or commodity, and units of measurement for output, production, or other indicators. Some conversions of units are easily carried out by multiplying a factor, such as from pounds to kilograms, but others require complex conversion matrices (also called "bridges") that map one category to another. In cases where category definitions or geographic delineations do not overlap exactly, interpolation will be required. In relational database parlance, this is known as a one-to-many relationship.

In this case study, practical examples will be used to illustrate how the conversions are carried out in common office productivity tools, and sample ready-to-use conversion matrices will be given for the purposes of reader exercise.

First, the complex unit conversions required for converting agricultural production data to freight volumes will be examined, in the context of common office productivity tools or relational databases capable of executing SQL[1] queries. Secondly, a commodity-based method for converting truck loads to tonnage of freight will be examined as an example of an allocation method. This will be followed by a demonstration of estimation methodologies based on a sampled distribution, in the case of converting vehicle counts to vehicle miles (VMT).

In the latter part of this case study, existing commodity classification schemes will be reviewed and strategies for converting between different schemes will be discussed and illustrated. Finally, the chapter will conclude with a discussion on adapting data collected at different geographic granularities for use in the same database – through a conversion of geographic definitions.

This chapter will focus mostly on data conversion, that is, taking one data source and transforming it in a way that could be more readily used for statewide freight planning. Often, the motivation for data conversion stems from a desire to merge two databases collected by different agencies to enlarge the sample size or to gain second-order insight.

---

[1] SQL stands for Structured Query Language, defined in ANSI standard X3.135-1986. Today's common office productivity tools that implement some variation of SQL includes: Microsoft Access®, FoxBase. Other server-platforms also implement SQL, such as Oracle, IBM DB/2, and Microsoft SQL Server. For more information on how to find a suitable SQL platform, see http://www.opengroup.org/public/tech/datam/sql.htm.

These complex data merges are covered in a different case study – NCHRP 8-43 Freight Data Case Studies, No.5.

2. Some Analytic Examples:

One of the best ways to understand the subtleties of data conversion is to consider a number of actual examples. Those selected for this discussion reflects the following scenarios:

- Converting agricultural data from production units such as heads of cattle, millions of eggs, and bushels of corn, to tons.
- Converting truck count data to tonnage data, using a commodity-based load-to-weight bridge, and vice versa.
- Converting truck count data to vehicle-miles (VMT) using a sampled distribution of lengths-of-haul.
- Converting between different schemes of commodity classification, including NAICS, SIC, and STCC.
- Converting between different geographic granularity, such as SPLC, FSAC, zip, FIPS, and County.

In each of these examples, the source data are identified, along with summary tables and conversion matrices extracted from public domain sources. There is an explanation about how the conversion is accomplished, what the index fields should be, and what steps should be taken to ensure the integrity of data is maintained. Finally, there is a discussion about the issues involved when using the results as inputs to other models.

2.1 Converting Agricultural Data

Agriculture production data is widely available from the U.S. Department of Agriculture (National Agricultural Statistics Service[2]) and other state-level sources, but it can be difficult to process because of the specialist units that are used in agricultural reporting. Agriculture productivity is usually measured in units of production per area of land, and such units of production can vary.

However, the availability of agricultural transportation data has declined in recent years. Some statistics, such as Agriculture freight flow data, is no longer collected, instead replaced by the Census of Agriculture,[3] which, like the Economic Census, is produced every five years. The lack of an annual data source on agricultural transportation has made the processing of agricultural production data on an annual basis even more critical,

[2] See http://www.usda.gov/nass/pubs/histdata.htm for more details.

[3] See http://www.nass.usda.gov/census/ for more details on the Census of Agriculture.

without which there would be no estimate of how much freight traffic is generated by agricultural activities.

2.1.1 Unit Conversions

Why do planners want to make these unit conversions? Livestock production is often measured in terms of "heads" or "birds"; to calculate the weight, it will be necessary to use an average weight per animal. Certain dry grains, especially when unprocessed, are measured in bushels – a unit of volume. It will be necessary to know the density to convert it to weight. The other commodities are reported in terms of weight – pounds, cwt (hundredweights), and perhaps tons. Occasionally, data will be reported in terms of metric units of grams and kilograms, or metric volumes such as cubic meters (m^3) or cubic decimeters (dm^3), especially in customs clearance data from Canada.

There can even be important differences between some units that have the same name. In the U.S., the word 'ton' generally refers to "short tons", which equals 2,000 pounds exactly. In some cases, data is reported in terms of 'gross tons', which equals 2,240 pounds. A metric ton is 1,000 kilograms, which equals 2,200 pounds. For planners trying to estimate freight volumes, these differences may be insignificant (as in the case of bulk freight where number of vehicles may vary with loading parameters and other factors) or significant (when such errors are allowed to remain in the dataset unchecked, their effect could multiply and give rise to a large estimation error in the final result) depending on the application. Where possible, it is always worth checking if the 'tons' given in different data sources are indeed the same 'tons'.

STCC4	Commodity	Units	Pounds per Unit	Tons per Unit	STCC4	Commodity	Units	Pounds per Unit	Tons per Unit
0141	Cattle on Feed	head	750	0.375	1191	Hay Other (Dry)	ton	2000	1
0151	Chickens-Commercial Broilers	bird	3.9	0.00195	1343	Lentils	cwt	112	0.056
0152	Egg Production and Value	1000eggs	120	0.0599	1915	Mustard	lbs	1	0.0005
0141	Hogs & Pigs	head	700	0.35	1133	Oats	bushel	32	0.016
0142	Milk Cows & Production	1000lbs	1000	0.5	1143	Peanuts	lbs	1	0.0005
0151	Turkey Production	bird	26	0.0129	1159	Peas Dry Edible	cwt	112	0.056
1131	Barley All	bushel	48	0.024	1134	Rice All	cwt	112	0.056
1131	Beans All Dry Edible	cwt	112	0.056	1135	Rye	bushel	56	0.028
1199	Beans Garbanzo	cwt	112	0.056	1149	Safflower	lbs	112	0.056
1341	Beans Other Dry Edible	cwt	112	0.056	1136	Sorghum For Grain	bushel	56	0.028
1299	Beans Pinto	cwt	112	0.056	1136	Sorghum For Silage	ton	2000	1
1149	Canola	lbs	1	0.0005	1144	Soybeans	bushel	60	0.03
1132	Corn For Grain	bushel	70	0.035	1149	Sunflower Seed All	lbs	1	0.0005
1132	Corn For Silage	ton	2000	1	1149	Sunflower Seed For Oil	lbs	1	0.0005
1129	Cotton Upland	bal	490	0.245	1149	Sunflower Seed Non-Oil Use	lbs	1	0.0005
1142	Flaxseed	bushel	56	0.028	1137	Wheat All	bushel	60	0.03
1191	Forage Alfalfa (Dry Hay+Haylag)	ton	2000	1	1137	Wheat Durum	bushel	60	0.03
1191	Forage All (Dry Hay+Haylage)	ton	2000	1	1137	Wheat Other Spring	bushel	60	0.03
1191	Hay Alfalfa (Dry)	ton	2000	1	1137	Wheat Winter All	bushel	60	0.03
1191	Hay All (Dry)	ton	2000	1					

Table 1: General Purpose Agricultural Conversion Matrix by Commodity

Table 1 is a conversion matrix detailing the Standard Transportation Commodity Code (STCC), commodity, and weight per production unit constructed from a variety of sources. The table allows planners to convert the incompatible units routinely reported in publicly available agriculture data into a uniform measure of volume: short tons. For

quick conversions, simply look up the appropriate commodity and use the conversion factor listed in the table in a spreadsheet.

2.1.2 Database Work

Occasionally, spreadsheet work becomes overwhelming in the face of extensive data, and a more sophisticated database or computer programs should be used to ensure accurate and reproducible results. In modern relational databases, the work of converting units can be made very easy. A quick example is demonstrated here. For more extensive detail, please refer to database manuals, online sources, and data analysis guidebooks available for many statistical or analytical application packages.

Figure 1 demonstrates the type of problems typically found in real world data. Although Abel and Charlie counties are reporting their agricultural production in recognized units, Baker county has reported its egg production in millions of eggs, and its chicken production in hundredweights. With only three counties and three commodities, these kinds of errors can be corrected manually. On a larger scale, it would be necessary to produce a database query that will automatically generate the count with the correct units.

Inconsistent units can also occur when data from multiple sources are merged together – for example, multi-state coalitions may have to merge coal production data from a number of state and private sources; perhaps states may report coal production in terms of tons but individual mines report them in terms of BTUs (a unit of energy content), or in terms of total revenue at the average minemouth price. This type of conversion method will be useful in that situation also.

County	Commodity	Unit	Count
Abel	Hogs	head	200
Baker	Hogs	head	180
Charlie	Hogs	head	210
Abel	Cattle	head	39
Baker	Beef	lbs	33,750
Charlie	Cow	head	26
Abel	Eggs	lbs	5,300
Baker	Eggs	millioneggs	0.0352
Charlie	Eggs	lbs	6,200
Abel	Chicken	tons	70
Baker	Chicken	cwt	1,780
Charlie	Chicken	longtons	21

Figure 1: Example Data

Commodity	Unit	lbsperUnit
Hogs	head	700
Cattle	head	750
Beef	lbs	1
Cow	head	750
Eggs	lbs	1
Eggs	millioneggs	120,000
Chicken	tons	2,000
Chicken	cwt	112
Chicken	longtons	2,240

Figure 2: Conversion Table

To create a conversion table, first identify how the tonnage estimate might depend on the reported data. In this case, because the average weight of hogs are different from those of cattle – and that the same commodity isn't necessarily reported in the same unit, the conversion factor will depend both on the unit and the commodity. It will be necessary to manually identify the conversion factor for every commodity-unit combination, shown in Figure 2. A planner familiar with SQL databases can use the SELECT DISTINCT statement to find all commodity-unit combinations in a large table.

Care should also be taken that every commodity-unit combination appears exactly once in the conversion table (for example, even though there were three records reporting hog production in terms of heads, the combination only appears once in the conversion table). Duplication will lead to erroneous results in database processing. The commodity-unit combination is called the "joint primary keys" of the conversion table, because a combination uniquely identifies a number for the conversion factor.

Once the conversion table has been created, tonnage of the commodities produced can be found by joining the two tables on the commodity-unit combination. An example SQL statement is shown below:

```
SELECT A.County, A.Commodity, A.Count*B.lbsperUnit as lbs
FROM ExampleData as A, ConversionTbl as B
WHERE (A.Commodity=B.Commodity) AND (A.Unit=B.Unit);
```

Most unit-conversion tasks and other data translation tasks can be quickly handled using similar SQL "joins" (database relationship statements) based on one or more primary keys (or index fields). A further discussion of the index fields can be found in Tutorial 5 on Merging Data.

2.1.3 Data Manipulation Considerations

It is always necessary to consider the accuracy of the resulting dataset when planning data conversion. In particular, planners should understand the purpose of the resulting dataset and ways in which different types of inaccuracies may impact the insights and conclusions.

An example of this dilemma is encountered when trying to determine traffic flow based on industrial activity or commercial transaction data. Traffic flow data is used for a variety of purposes, and each purpose can require a different dimension of measurement. If traffic flow data is needed to assess the risk of a grade-crossing accident, then the number of trucks that a location generates (including empties), and their sizes, are critical dimensions of the analysis. On the other hand, to assess whether a certain high-volume truck flow can be replaced with a once-weekly round trip on the rails, the tonnage number (or even commodity value per ton) may be more important.

Generally, minimizing the number of data translations required to get from observed counts to a target dataset will result in the greatest possible accuracy. In the case of traffic flow data, if the source is DOT "loop counters"[4], then every effort should be made to extract discrete vehicles counts by class – which will give an accurate count in terms of vehicles per day. If the source is production data in production units, then it should be directly converted into either loads or tons (depending on the need) using a conversion matrix similar to ones demonstrated above. Planners should avoid converting the production data from units to tons first, followed by a conversion from tons to vehicles

[4] Automated vehicle-counting devices are used by highway maintenance departments to monitor traffic volumes. Those devices are typically inductive loops buried in the pavement, which are capable of counting vehicles and axles, although more recently other technologies have been used. Collectively such devices are termed "loop counters".

using such measures as the average vehicle weight; it is usually better to determine, for example, on average how many heads of cattle are carried in a single truck, without reference to the product weight.

2.2 Converting between Vehicles and Tonnages

1.1.1. Converting from Vehicles to Tonnage

Converting from vehicles to tonnage is generally not a helpful exercise. If the source of the vehicle count data is highway loop counts, then a critical dimension in converting from vehicles to tonnage will be missing – commodity. If on the other hand, the vehicle counts come from a different source (such as a shipper survey, or a commodity-flow database), it is usually possible to gather the data directly in terms of tons. Commodity flow databases are generally derived from commerce or production statistics, and tonnage numbers are frequently available.

However, if the conversion must be done, and the commodity information is not available, it is generally accepted that an average American truck carries 14 net tons.[5] If more detailed information is available, such as FHWA class, number of axles, or type of body, then it will be possible to calculate the tonnage more precisely. However, beware of empty return equipment. Although a chemical tanker can carry a maximum of 22~25 tons payload,[6] almost all trucks carrying "dirty" chemicals[7] return empty, and therefore the average net weight is actually half the maximum value.

On the whole, it is very difficult to calculate tonnage carried purely from the truck counts, unless there is some supplemental information indicating what the trucks might be hauling. Another point worthy of note is that when planning highway maintenance, gross weight of trucks should be used instead of net weights. Conversely, when attempting to use engineering data for planning purposes, this should be taken into account and the weight of the chassis should be subtracted where possible, based on an index such as the vehicle type, FHWA class, or some other classification scheme. To understand how a distribution of ranges could be used to provide a better conversion, see Section 2.3 on distributions.

---

[5] TRANSEARCH® 2001 estimated value, based on national average values reported by participating carriers.

[6] Typical chemical tanker capacity converted from gallons to tons, using typical transported fluid densities.

[7] "Dirty" chemicals is the nickname given to chemicals which soil the transporting tanker to the point where tanker re-use for a different product is impossible unless heavy cleaning is performed. Generally, food-grade tankers are not re-used on a return trip, neither are those carrying dirty chemicals (e.g. bunker oil, sewage). Some commodities are more compatible with others, and thus mixed fleets can have a higher backhaul ratio.

1.1.2. Converting from Tonnage to Vehicles

The tonnage-to-vehicle conversion matrix is an element of the commercially available TRANSEARCH® freight flow database. The database also addresses storage and equipment requirements for each commodity covered through its equipment type classifications. An example matrix based on the FHWA's Freight Analysis Framework is given in Table 2.

STCC2	Description	Tons /Load	STCC2	Description	Tons /Load
20	Food and Kindred Products	18	31	Leather Products	11
21	Tobacco Products	17	32	Concrete, Clay, Glass & Stone	14
22	Textile & Mill Products	17	33	Primary Metals	20
23	Apparel Products	13	34	Fabricated Metals	14
24	Lumber & Wood Products	21	35	Machinery	11
25	Furniture	11	36	Electrical Equipment	13
26	Pulp or Paper Products	19	37	Transportation Equipment	11
27	Printer Matter	14	38	Instruments	10
28	Chemical Products	17	39	Misc. Manufactured Goods	15
29	Petroleum or Coal Products	22	41	Misc. Freight	16
30	Rubber & Plastics	9	50	Secondary Traffic	8

Table 2: Tonnage-to-volume conversion matrix by STCC2

Note that in general bulk trucks (such as those carrying farm products, sand & gravel, lumber, paper, scrap metal, and chemical feedstocks) tend to carry higher tonnages, while dry vans, especially ones involved in warehouse-to-store distribution, tend to carry lighter tonnages. The economic importance of heavy trucking to some areas is so important that there are specific exemptions[8] that permit trucks heavier than the Interstate System standard to travel on local and state routes. Thus, if a railfreight diversion analysis involves diverting a heavy bulk commodity, the true average weight of trucks in the locality should be carefully studied and understood – elsewise the analysis will overestimate the benefits of diverting to rail, particularly if national average truck weights are used.

2.3 Estimates based on Distributions

1.1.3. Conversion from Vehicles to Vehicle-Miles (VMT)

There are two basic ways to convert vehicle count data to vehicle mileage data, which is often used in the assessment of operating costs and externalities. For example, pollution from particulate matter, fuel consumption, and driver cost are often of interest to the public planner, and are best estimated using VMT (as well as vehicle usage pattern) figures.

[8] Due to the importance of the paper and lumber industry to northern New England's economy, the State of Maine and New Hampshire allows trucks up to 100,000 lbs gross to traverse its state highways and part of the Maine Turnpike. This is also consistent with the Canadian standard for truck weights. The Eisenhower Interstate System limits truck weights to 80,000 lbs. Thus, vehicle count impacts for heavy-truck railfreight diversion in northern New England could be 20% less than in other parts of the United States.

A more accurate but also more complicated way of generating VMT data requires knowledge about the actual origination and termination points of each truck. In this instance, individual mileages are found using a shortest-path routing model[9].

If the vehicle count data has been obtained from an inductive highway loop, the commodity origin-destination is unlikely to be available. For this situation, it may still be possible to convert vehicles to vehicle-miles by using a regionally or commodity appropriate length-of-haul distribution.

The method for converting vehicles to vehicle-miles is essentially one of expansion to the universe based on sample observations. Table 3 represents a general distribution of mileage ranges, derived from FHWA's Freight Analysis Framework Database.[10] This table can be used to find the average length of haul, the number of vehicles likely to fall within each strata, and the likely VMT impact from the vehicles observed. The total VMT is the sumproduct of total vehicle count, the average miles driven, and the percentage of trucks in that strata.

Strata	Average Miles Driven for Truck in Strata	Percentage
0~50 miles	20	37%
50~100 miles	66	20%
100~250 miles	165	21%
250~500 miles	350	12%
500~1,000 miles	703	6%
1,000~2,000 miles	1,352	3%
Over 2,000 miles	2,410	1%

Table 3: U.S. Truck Length of Haul Distribution, 1997

It must be emphasized that the resulting vehicle-mile data represents an *estimate* – it is no longer as reliable as the loop count data. The resulting VMT should be used with extreme caution, particularly if any of the following circumstances exist:

- If the daily truck count is less than 30 vehicles in one direction (60 in both directions) for a given segment of highway;
- If observation or traffic-cam reveals that the same trucks are consistently using the route on a daily basis (the commuter-truck phenomenon);
- If there are substantially parallel routes where it is suspected that traffic will segregate by lengths-of-haul – for instance, long distance traffic is much more

[9] Uses of the Oak Ridge National Laboratory shortest-path network model is covered in the NCHRP 8-43 Virginia I-81 case study by Reebie Associates. Other shortest-path tools are available on the internet at http://www.mapquest.com/ and at http://maps.yahoo.com/. Professional truckers operating in unfamiliar territory use the same tools for their route planning.

[10] Data was collected from 1995~1997, but the distribution of lengths-of-haul are unlikely to have changed between then and now. More recent and more detailed data similar to this is available from Reebie Associates under the TRANSEARCH® brand name.

likely to use I-81 rather than the adjacent U.S. 11, parallel routes through most of western Virginia;

- If VMT were to be used in air quality studies, in which case the result would be sensitive to both the VMT and the truck operating characteristics and conditions.

Although the mileage distribution for trucks generally follows a similar pattern nationally, these statistics can vary substantially depending on the area's local geographical situation, proximity to other centers of business activity, and also the type of industries that dominate the local economy. If possible, origin-destination intercept surveys should be used to supplement the distribution data given here, or to provide a cross-check for accuracy.

To illustrate the biases that arise when high concentrations of single-commodity movements occur (e.g. in heavily farming or mining regions), Table 4 gives the length-of-haul distributions by commodity groups. If the commodities carried by the trucks observed in the loop count data is known, such as through the use of a traffic cam or a "traffic counter's " direct observations, it will be possible to use Table 4 to estimate VMT. However, the statistical accuracy of this method cannot be guaranteed, and analysts should use the results carefully.

Commodity	0~50 miles	50~100	100~250	250~500	500~1k	1k~2k	2k+ miles
Farm Products						3%	7%
Processed Foods	3%	4%	9%	12%	16%	19%	15%
Forest Products		4%	8%	9%	9%	6%	7%
Paper Products				3%	5%	6%	7%
Metals			3%	6%	11%	11%	13%
Minerals	70%	64%	36%	21%	8%	7%	3%
Chemicals			4%	7%	13%	16%	9%
Petroleum Products	4%	5%	9%	11%	11%	12%	2%
Equipment				3%	6%	8%	16%
Misc. Manufactured				2%	5%	7%	15%
Secondary Traffic (Distribution)	16%	17%	24%	25%	13%	5%	4%

Table 4: U.S. Truck Commodity Distribution > 2% by Length of Haul, 1997

Reading vertically down the first column reveals that 70% of all tonnages between zero and 50 miles is attributable to aggregates, and only 16% to secondary traffic. (However, the number of discrete loads attributable to secondary traffic is likely to be much higher because aggregates trucks tend to be heavier). Reading horizontally across the "minerals" column reveals that the proportion of tonnage attributable to aggregates tends to decrease as lengths-of-haul increase. Minerals rarely travel more than about 500 miles by truck, but some do, and those are likely to be precious ores which cannot be processed locally, and lack access to suitable rail service.

Note that the proportion of minerals traffic declines dramatically with increasing lengths of haul, while secondary traffic (warehouse-to-warehouse or warehouse-to-store movements by retailers and distributors) tends to average around 250 miles (one day's round-trip drive). The totals do not add up to 100% because not all commodity groups were included in this table, and percentages of less than 2% were omitted because they are not considered reliable. Notably, drayage or transloading movements were also not included – thus, truck movements from mineheads to rail or barge loading pads will not show up, explaining the lack of coal and ore traffic.

1.1.4. Converting from Tons to Ton-miles

To convert from tons to ton-miles, a shortest path model is often used. Without knowing the origins and destinations of the shipment, a length-of-haul distribution can be used, similar to the procedure for converting loads to vehicle miles (as outlined in Section 1.1.3 above). If the vehicle type is known, then length of haul distributions can be further disaggregated based on vehicle type (for example, dry vans travel more distance on average than open-dump sand & gravel trucks). If lengths-of-haul distributions by other factors are available based on surveys or from available sources, these could also be used to help improve the accuracy of the results.

Essentially, the task of converting vehicles and tons to VMT and ton-miles without using a shortest path model consists of a database-join between the length-of-haul distribution and the observed vehicles or tonnages. This join can be done without any indexing criteria based on a general length-of-haul distribution, but the results are likely to be much more accurate if at least some of the characteristics of the trucks observed or tons observed, or commodities observed are known.

2.4 Conversion between Commodity Classifications

2.4.1 Introduction to Commodity Classification

There are several different methods of classifying commodities that are transported. Railroads have generally used the Standard Transportation Commodity Classification (STCC) developed by the Association of American Railroads (AAR) in the 1960s, which is a code up to a maximum of seven digits that can identify the transported goods very precisely. For example, STCC 11 212 11 classifies "Coal, bituminous for pulverized or granular injection into a blast furnace"[11]. Machine-readable files and printed sources detailing the STCC definitions are available from Railinc, Inc.[12]

In transportation analysis, it is much more common to use STCC codes aggregated at a higher level, for example STCC2 (2-digit level) or STCC4 (4-digit level). For example, STCC 11 is simply "Coal", and STCC 1121 is "Bituminous Coal". STCC codes can be aggregated this way because similar commodities are generally grouped together in numeric order. However, derivative products may be assigned a number from a different range, for example, while forest products are categorized as STCC 08, dimensioned lumber is classified under STCC 24 and paper is classified under STCC 26.

Although STCC is commonly used for transportation-related analyses, a different system is normally used for economic analyses, including the Economic Census. Prior to 1987,

---

[11] From the "STCC book", or Standard Transportation Commodity Code STCC 6001-Y issued by the Association of American Railroads (AAR), Operations & Maintenance Department, Agent. This book can be purchased from the AAR.

[12] Railinc Inc. is at http://www.railinc.com/.

the system was called Standard Industrial Classification[13] (SIC). During the 1997 Economic Census, a new system was introduced and termed the North American Industry Classification System[14] (NAICS), to better document trade activity resulting from the North American Free Trade Agreement (NAFTA).

Statistics Canada[15] has generally reported its transportation analysis in terms of Standard Classification of Transported Goods[16], which was designed to handle a greater diversity of commodities than STCC alone offered. Some data, notably import and export data collected by customs officials and steamship lines, may be reported in the Harmonized System (Harmonized Commodity Description and Coding System, or "HS" for short), which has its origins in determining tariff charges when goods cross national borders.

2.4.2 Mapping between Different Systems of Classification

There is rarely a one-to-one correspondence between systems of classification, because the systems were developed for different purposes by users focusing on different aspects of analysis. Generally, mapping from STCC at the two-digit level to SIC or NAICS at the two-digit level can be done quite precisely. However, beyond the two-digit level, mapping is difficult and may require estimation. Also, since SIC and NAICS are industrial classification systems, whereas STCC is focused on classifying transported products, there will be SIC codes for which there is no corresponding STCC code, especially in the service industry such as hospitals, legal, and analytical businesses (SICs 80, 81, etc).

Typically, those industries that produce a commodity product have the same 2-digit STCC code as its 2-digit SIC code. For example, metal mining are industry classification SIC 10 and transported metallic ores are classified as STCC 10. This parallel generally works only for producers of a specific commodity, not consumers. Thus, while power plants consume large volumes of coal (STCC 1121), they produce electricity (SIC 4911). Since electricity is transmitted and not transported, there is no corresponding STCC code for electric power.

---

[13] SIC is still the classification system of choice used in some governmental functions and private survey work. Many private economic databases are coded in terms of SIC; only those which concern the NAFTA trade will generally have been converted into NAICS. A full SIC manual is available on the Department of Labor's website at the following address: http://www.osha.gov/pls/imis/sic_manual.html.

[14] More information about NAICS, including machine-readable lists of classification definitions, are available at the NAICS website at http://www.census.gov/epcd/www/naics.html.

[15] Statistics Canada is the Canadian authority for compiling national statistics, similar to the U.S. Department of Census and other statistical bureaus. Statistics Canada has a website at http://www.statcan.ca/.

[16] More about STCG and other classification schemes used by Statistics Canada is available here: http://www.statcan.ca/english/Subjects/Standard/sctg/sctg-intro.htm.

SIC of Establish-ment	Primary Transport STCC	Commodity	Pcnt of Freight for Primary STCC	Scndry Transport STCC	Commodity	Pcnt of Freight for Scndry STCC
2011	2011	Meat, Fresh or Chilled	95%			
2011	2011	Meat, Fresh or Chilled	66%	2013	Meat Products	9%
2011	2011	Meat, Fresh or Chilled	63%	2015	Dressed Poultry, Fresh	12%
2011	2011	Meat, Fresh or Chilled	47%	2026	Processed Milk	41%
2011	2011	Meat, Fresh or Chilled	43%	2032	Canned Specialities	15%
2011	2011	Meat, Fresh or Chilled	63%	2033	Canned Fruits, etc.	21%
2011	2011	Meat, Fresh or Chilled	67%	2034	Dried Fruits, etc.	11%
2011	2011	Meat, Fresh or Chilled	38%	2037	Frozen Fruit, etc.	13%
2011	2011	Meat, Fresh or Chilled	65%	2038	Frozen Specialities	8%
2011	2011	Meat, Fresh or Chilled	36%	2042	Canned Feed	54%
2011	2011	Meat, Fresh or Chilled	63%	2051	Bread & Bakery Pdts	7%
2011	2011	Meat, Fresh or Chilled	25%	2093	Nut or Veg. Oils	13%
2011	2011	Meat, Fresh or Chilled	41%			
2011	2011	Meat, Fresh or Chilled	40%	2086	Soft Drinks	32%
2011	2011	Meat, Fresh or Chilled	27%	2031	Canned Seafoods	29%
2011	2011	Meat, Fresh or Chilled	39%	2036	Processed Fish	46%
2011	2011	Meat, Fresh or Chilled	49%	2099	Food Prep, NEC	11%
2011	2011	Meat, Fresh or Chilled	76%			
2011	2011	Meat, Fresh or Chilled	69%			
2011	2011	Meat, Fresh or Chilled	60%	2654	Sanitary Food Containers	11%
2011	2011	Meat, Fresh or Chilled	79%			
2011	2011	Meat, Fresh or Chilled	40%	2841	Soap & Detergents	19%
2011	2011	Meat, Fresh or Chilled	77%			

Table 5A: Transport requirements for a sample of industries whose primary SIC classification is 2011 (Meat Packing Plants)

As demonstrated in Table 5A, many industrial facilities produce more than one product, their assigned primary SIC code will correspond only to their primary product but not any secondary products. If specific data is not available on the quantity of primary product versus byproduct that was produced at a particular facility, it may be possible to use information similar to that contained in Table 5A[17] as a basis for an allocation of production to primary and secondary products.

Such an allocation will necessarily require some assumptions to be made. It is possible to treat SIC 2011 as one single industry and assume some percentage of its transportation requirement will consist of STCC 2011, and also of the other STCCs shown, using some type of weighting scheme. However, Table 5A is based on national survey data, and the results will not necessarily be accurate for either a specific area or a specific plant – at least, not at the four-digit STCC level. It is therefore recommended that a SIC-STCC conversion or vice versa be made at the 2-digit level – as evidenced by the fact that only a small minority of SIC 2011 plants have transportation requirements in commodities other than those listed under STCC 20.

[17] Information similar to Table 5A can be obtained from a number of private sources. Reebie Associates' Freight Locater® is one such product. Third-party information services such as InfoUSA, Harris, and others, also make such a database available for marketing and analytical purposes.

SIC of Establish-ment	Primary Transport STCC	Commodity	Pcnt of Freight for Primary STCC	Scndry Transport STCC	Commodity	Pcnt of Freight for Scndry STCC
2429	2429	Misc. Sawmill/Planing Mill	92%			
2429	2429	Misc. Sawmill/Planing Mill	77%			
2429	2429	Misc. Sawmill/Planing Mill	61%	2491	Treated Wood Pdts	24%
2429	2429	Misc. Sawmill/Planing Mill	71%	2499	Misc. Wood Pdts	8%
2429	2429	Misc. Sawmill/Planing Mill	68%	2621	Paper	7%
2429	2429	Misc. Sawmill/Planing Mill	74%	2821	Synthetic Fibres	8%
2429	2429	Misc. Sawmill/Planing Mill	53%	2952	Paving Blocks or Mix	32%
2429	2429	Misc. Sawmill/Planing Mill	78%			
2429	2429	Misc. Sawmill/Planing Mill	76%			
2429	2429	Misc. Sawmill/Planing Mill	66%			
2429	2429	Misc. Sawmill/Planing Mill	78%			
2429	2429	Misc. Sawmill/Planing Mill	71%			
2429	2429	Misc. Sawmill/Planing Mill	75%			

Table 5B: Transport requirements for a sample of industries whose primary SIC classification is 2429 (Special Product Sawmills, not Elsewhere Classified)

Table 5B demonstrates another situation where primary SIC may be different from the STCC code of the commodities requiring transportation. Most sawmills manufacture wood products, but others have diversified into making paper or garden mulch, extracting fibers, or even making paving blocks. Such sawmills are necessarily somewhat specialized and local knowledge or direct contact will be required to determine whether such a condition exists in the study area.

Another important factor is that an industry's primary product may not correspond with its transportation requirements. Typically, producers of raw materials tend to have large volumes of primary product, while producers of high-value products may find that waste materials or production byproducts dominate their transportation requirements. This is particularly true of ore processors, whose transportation requirements for spoils and slag often far exceed those for the extracted metal.

Generally, the need to convert between different systems of classification occurs because two datasets from different providers needs to be merged or joined. For more discussion on how the datasets can be connected together to a yield fresh insight or larger database, consult NCHRP 8-43 Case Study 5.

2.4.3 The Fragmentation Problem

With the ease of bulk information processing and database manipulation work, many analytical procedures previously considered impossible are becoming routine. However, it is important to consider output data integrity when converting between discrete classification systems based on percentage distributions. Such distributive processing featuring many steps involving one-to-many mappings can quickly create unmanageably large databases with unrealistic results.

In the SIC to STCC conversion example, each four-digit SICs converted into a primary four-digit STCC and also a number of secondary four-digit STCCs. Assuming that there were 500 SIC codes in the study area, and each SIC code generated on average five

distinct STCCs, the result will contain 500 primary STCC codes (with substantial tonnage) in addition to 2,000 secondary STCC codes (with lesser tonnage). This phenomenon is termed "flow fragmentation" – because whole flows based on SIC codes are split into many smaller flows based on their STCC classification.

Some fragmentation in the data during the disaggregation process may be helpful, as it allows the analyst to focus more narrowly on specific commodities or geographic regions. However, beyond a certain level, the flows become too fragmented and difficult to work with, and also will not carry much meaningful data. For instance, if the merged database shows 499.5 loads of lumber and 0.5 loads of paper being shipped annually from a local sawmill, it is almost certainly an artifact of the conversion process – somewhere there may be a sawmill which produces paper products also, but it is highly unlikely that the sawmill products to paper ratio would be 1000 to 1 in terms of annual production volume. An algorithm that prevents fragmentation[18] should be implemented to ensure that the traffic is distributed in a realistic fashion.

2.5 Conversion of Geographic Definitions

2.5.1 Concepts of Transport Geocoding

There are many ways to code transportation geography, but the boundaries for conversion do not always overlap. A boundary problem occurs when zones coded by one scheme do not share boundaries with zones coded by a different scheme – i.e., when one zone could map to part of two or more different zones by another coding, and conversely parts of two or more zones may map to the same zone. An example of this is found in Postal zip code boundaries and county boundaries: they do not always overlap. In some cases a zip code area will straddle two adjacent counties, and most counties contain many zipcode zones, with some zip codes assigned to individual buildings that process bulk quantities of mail.

Another key issue is point coding versus zone coding. A scheme based on point coding (such as latitude and longitude) identifies a unique point in space, whereas a zone identifies an area covering many different points in space. Thus, shipments from the same warehouse may show different latitude and longitude records depending on the dock at which the truck was loaded and the accuracy of the GPS device used to obtain the lat/long information, while they will always show the same zipcode because the warehouse has one single unique zip code (even though it may be located in multiple county jusridictions!).

A key concept concerning zone coding is that zones will necessarily denote a finite area of space, and the area of space will necessarily have a *centroid*, which is a term used to

[18] There are many ways to prevent fragmentation; one is to simply reject all flows that show tonnages or loads under a certain threshold figure. A more complicated way involves using an integer program to optimize the load distribution, requiring a solution that is expressed in terms of whole number of trucks or tons. These algorithms are beyond the scope of this paper.

describe the point in the center of the zone where all traffic could be assumed to flow to, if the distribution of traffic to every part of the zone is uniform. In other words, some traffic will terminate at the extreme corners of the zone, but the difference in mileage balances out if they are all assumed to terminate at the centriod[19].

There are two types of centroids: weighted and unweighted centroid. The unweighed centroid assumes that the geographical center of the zone is the centroid, which assumes traffic flows to all areas of the zone in equal proportions. The weighted centroid takes into account the traffic flow may be heavier to one end of the zone versus the other. As an example, if the defined zone is the State of Illinois, the unweighted centroid would be close to Springfield, Ill., whereas the weighted centroid would be immediately southwest of Chicago, perhaps near Joilet, Ill. This is because volumes in the greater Chicago region dominate traffic flowing into and out of the State of Illinois. In general, weighted centroids are a better way to geocode freight data, since distribution of freight activity within a zone is seldom uniform.

2.5.2 Different Ways to Code Transport Geography

A sample list of different ways to geocode freight data is given below. This list is not intended to be exhaustive.

- **Standard Point Location Code (SPLC):** A Standard Point Location Code (SPLC) is assigned to all stations registered by rail carriers. Between six to nine digits, this numeric code is used to specify the physical location of a station. Presently a minimum of six digits is required with three zeros. SPLCs are assigned through Railinc's Business Services Division.[20]
- **Zip Code:** ZIP Code identifies a specific geographic delivery area for the U.S. Postal Service. On July 1, 1963, the U.S. Postal Service implemented the ZIP (Zoning Improvement Plan) Code to improve the sorting and delivery of mail and ease the way toward better, faster automated processing of letters and packages.[21]
- **Federal Information Processing Standard (FIPS):** Federal information processing standards codes (FIPS codes) are a standardized set of numeric or alphabetic codes issued by the National Institute of Standards and Technology (NIST) to ensure uniform identification of geographic entities through all federal

---

[19] To understand the concept of centroids, consult any standard textbook or software manual for a geographic information system (GIS) package.

[20] A machine-readable database of SPLC codes may be obtained from Railinc, Inc. at http://www.railinc.com/.

[21] A database containing all the zip codes and mappings to town and counties can be obtained from the U.S. Postal Service by calling the National Customer Service Center at (800) 238-3150. There are also third-party providers that sell the same data, including Reebie Associates, and other commercial outlets such as zipinfo.com.

government agencies.[22] The entities covered include: states and statistically equivalent entities, counties and statistically equivalent entities, named populated and related location entities (such as, places and county subdivisions), and American Indian and Alaska Native areas. Generally, FIPS codes at the five-digit level represent counties and independent cities in some jurisdictions (e.g. Virginia).

- **Freight Station Accounting Code (FSAC):** Rail station records are uniquely identified by combination of the Standard Carrier Alpha Code (SCAC) field and Freight Station Accounting Code (FSAC) in the Centralized Station Master (CSM) file. The file is a geographic location database, which contains data about rail and motor carrier points for North America and international areas. This file is primarily used by railroads to help plan freight movements from origin to destination in an efficient and timely manner. Generally, SPLCs are preferred to FSACs, because data processing for multiple rail carriers is easier with SPLCs rather than FSACs.
- **Longitude and Latitudes (lat/long):** Lat/long is a definitive way of uniquely identifying a point in space on earth. Many Global Positioning System (GPS) devices output locations in terms of latitude, longitude, and altitude, which corresponds to a measure of distance from the equator on a north-south axis, from Greenwich in England on a east-west axis, and from average sea level on a vertical axis respectively.
- **Census Tract:** This is the smallest geographic unit used by the U.S. Census to enumerate population and demographics data. Data may also be available on a local level at other granularities such as Quarter Sections, street address, and other non-standard methods of coding geographic data.
- **Mile Markers:** For ongoing barge operations, mile markers or mileposts along the river may be used to identify the loading and unloading locations uniquely. Not all barges dock at recognized ports, and some ports define their activity in terms of all loading/unloading taking place on the riverbank between a lower and upper limit of mile markers. When utilizing marine data, a marine atlas showing mile markers[23] and correlating it to another measure (such as lat/long) can be very helpful.

---

[22] A FIPS database in machine-readable format can be obtained from the U.S. Geological Survey at http://geonames.usgs.gov/fips55.html. There are also third-party companies who will provide bridges between FIPS and zip, and other geographic coding zones. Standard GIS packages such as ArcView and TransCAD should also be able to perform the required conversion from any geospatial coding to FIPS.

[23] The Tennessee Valley Authority (TVA) has detailed data for activities taking place along the Tennessee River by mile marker. The Army Corps of Engineers are custodians of data for the other inland waterways, but the data may not be publicly available. For coastwise shipping and international shipping, origins and destinations are generally defined in terms of the port, or dock name within a large port. Some port authorities have this information and may be prepared to release them for state studies.

2.5.3 Typical Method for Constructing a National Freight Database

Generally, rail waybill sample can be obtained in terms of SPLCs, although some commercial sources of waybill information now come FIPS coded. Motor freight data tend to be coded in terms of lat/long (if a GPS data-collection device is used), or zip code (if accounting or customer contact data was used). Inland marine data may be coded in terms of mile markers as previously mentioned, or ports. To further complicate matters, marine freight (and to some extent rail freight) tend to be drayed some distance from the terminating point, which may or may not fall within the same zipcode or county/FIPS zone.

It is typical to convert all data into a FIPS-coded or zip-coded database for statewide freight planning – which may involve a national database as states will need to consider the impact of through freight and terminating freight resulting from interstate commerce. Another geocoding scheme often used is the Business Economic Area or "BEA", which are collections of counties defined by the Bureau of Economic Analysis based on trade patterns. The U.S. Census Bureau defines Consolidated Metropolitan Statistical Areas (CMSA) which are large, multi-state metropolitan areas, but it does not cover parts of the country that are not considered 'metropolitan' and thence is rarely used in national freight analysis.

Conversion from one geocode scheme to another is best accomplished by using a one-to-one mapping matrix, and in cases where there is a one-to-many correspondence use weighting and/or distribution methods based on some measure of economic activity. For instance, if one zipcode zone straddles the boundary between two counties, it is likely that the traffic is proportional to the percentage of land area occupied by the zipcode zone within the two counties.