Method to Synthesize a Full Matrix of Interdistrict Highway Travel Times from Census Journey-to-Work Data

W. THOMAS WALKER

The 1980 census Urban Transportation Planning Package (UTPP) is a large-scale source of observed travel times derived from the journey-to-work data. Plans for the 1990 census indicate that work trip times will again be collected. These data are incomplete, however, in that most cells in the interdistrict matrix have no census travel time observation. This paper explores the accuracy and usefulness of census travel time data versus survey- and network-based methods and airline distance approaches at the minor civil division (MCD) level of analysis. A promising technique to extract and edit the census data and to synthesize the missing observations is developed.

The need for accurate estimates of a matrix of highway travel times between a fine-grained system of zones arises in many transportation activities, including market analyses for retail business locations, vehicle routing and scheduling operations, highway needs studies and cost-benefit analyses, transit service planning, highway accessibility studies, and validating the travel time data input to travel forecasting models. The large number of cells contained in realistic travel time matrices generally precludes exhaustive interdistrict estimation with floating car survey methods because of high costs. Travel time matrices are usually generated by building minimum paths through computerized networks, where the links may contain measured speeds, but more often, because of survey costs, contain generalized averages of measured speeds, perhaps stratified into a number of route types or area type and time-of-day categories. The large geographical area associated with commercial census data bases and many highway-related activities, such as scheduling long-haul truck deliveries often precludes network methods because of the large number of links required. In such cases, travel time estimation based on some form of airline distance analysis is the norm.

The 1980 census Urban Transportation Planning Package (UTPP) is a large-scale source of observed travel times derived from the journey-to-work data. Plans for the 1990 census indicate that work trip times will again be collected. These data are incomplete, however, in that most cells in the interdistrict matrix have no census travel time observation. This paper explores the accuracy and usefulness of census travel time data versus survey- and network-based methods and airline distance approaches at the Minor Civil Division (MCD) level of analysis. A promising technique to extract and edit the census data and to synthesize the missing observations is explored. Comparisons between census and network-based travel time estimates are made for a four-county portion of New Jersey that was covered by a recent travel time survey, as are comparisons between census and airline distance estimation methods. As a case study, a census-based travel time matrix is generated at the MCD level for the portion of the Northeast corridor shown in Figure 1, including southeastern Pennsylvania and the entire state of New Jersey (816 districts).

Delaware Valley Regional Planning Commission, The Bourse Building, 21 South 5th Street, Philadelphia, Pa. 19106.

BACKGROUND

Over the last quarter century there has been a great deal of work on estimating highway speeds and travel times. The relationship between highway speed, vehicular volume, and roadway capacity on individual links was studied intensively in the early 1960s under the auspices of the FHWA and the Highway Research Board. This work culminated in the 1965 Highway Capacity Manual (*1*). Because of the importance of speed in determining roadway service levels for design purposes, interest in this subject has continued to the present day (*2*). This work, however, is focused on individual highway links with little attempt to estimate interdistrict travel times.

Travel-demand forecasters, however, have considerable interest in travel times because they are used as a principal measure of interzonal highway service levels for purpose of trip distribution and modal split. These analyses, however, tend to be focused on the impact of errors in various network methods on the accuracy of the travel demands generated for specific transit and highway facilities (*3,4*). In this work, accuracy in trip travel time estimates is often sacrificed to obtain more accurate demand estimates or network assignments, viewing the resulting time as an impedance or composite measure including other factors. To date, there has been little attempt to use or enhance the census journey-to-work travel times except perhaps to validate network-based travel time results, because direct census observations fill only a small portion of the highway travel time matrix.

METHODS FOR ESTIMATING INTERDISTRICT TRAVEL TIMES

This section explores three methods for synthesizing travel times between a moderately dense system of analysis districts; synthesized census work-trip times, minimum path network

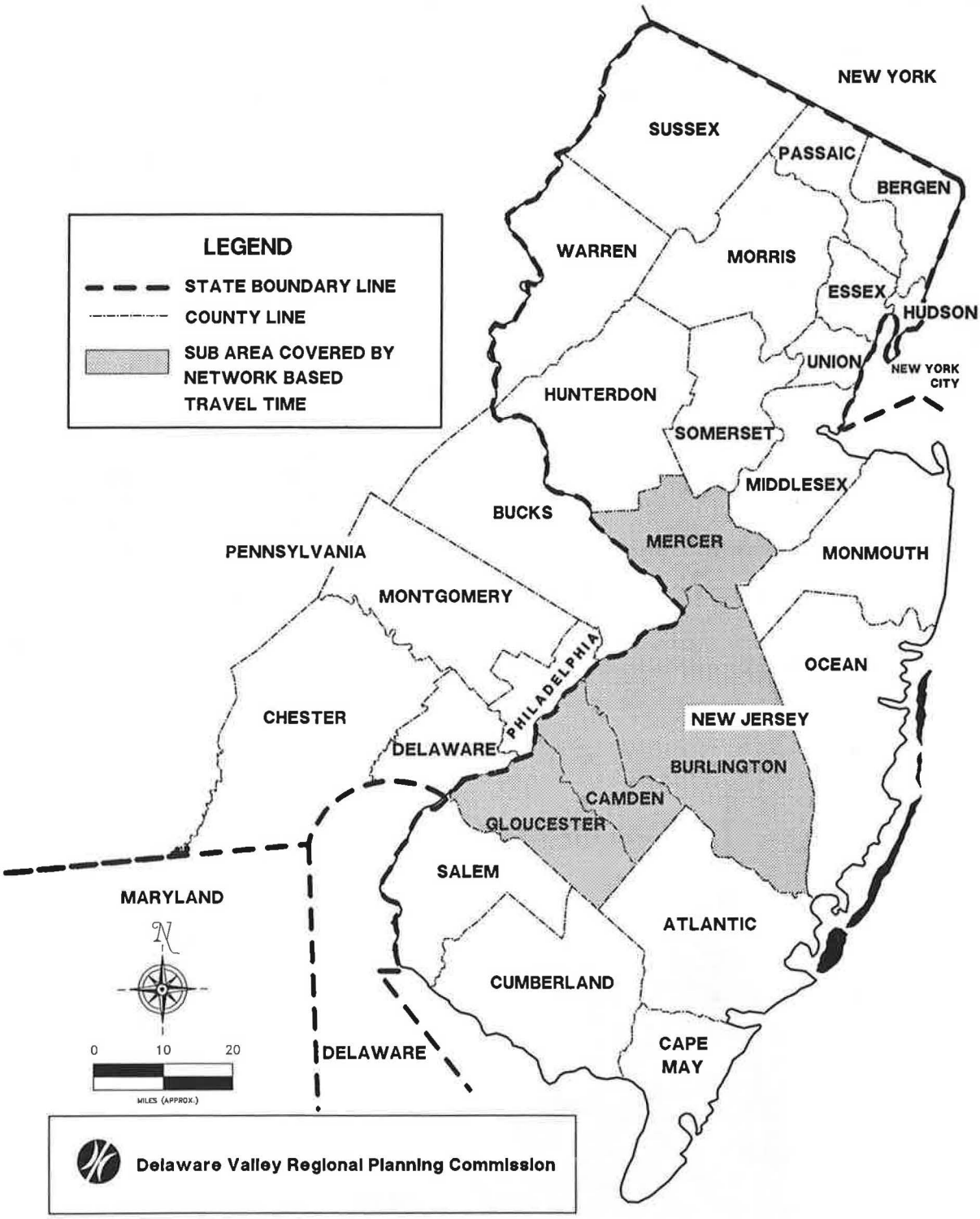

FIGURE 1 Case-study area for travel time estimation.

methods, and airline distance-based time calculations. The district system used is based on MCD within southeastern Pennsylvania and New Jersey. In total, these comprise some 816 districts with an average area of about 12 square miles, although there is considerable variation in individual district sizes ranging from small boroughs to large rural townships. The City of Philadelphia is technically one MCD. Philadelphia was broken up into 12 subdistricts in this analysis.

Questions 23 and 24 of the 1980 Census long form asked each employed resident to indicate where he worked, the work-trip time, and what travel mode was used. These data were processed, tabulated, and averaged at the census tract level for just under eight percent of all employed residents and distributed in Part IV of the UTPP (*5*). Both the Delaware Valley Regional Planning Commission and the New Jersey Department of Transportation purchased the 1980 UTPP. These data files provide a very large number of observations of interdistrict travel time. The drive-alone travel mode was chosen because it gives the most direct estimate of over the peak-period road-travel times. Carpooling travel times were rejected because they included an additional travel time associated with picking up passengers. In total, almost 160,000 census tract level observations of drive-alone travel times were recorded in the combined UTPPs, some 15,000 of which were duplicates caused by overlap of the UTPP areas. When averaged to the MCD or district level of analysis, these observations cover almost 40,000, or six percent, of the 666,000 possible interdistrict interchanges. On average, there were about 4.2 census tract travel time observations per MCD interchange covered by the sample before editing. These observations were not evenly distributed, however. Some major centers, such as the Philadelphia central business district, had hundreds of observations, while small MCDs may have had 10 or fewer observations. On average, there were almost 47 travel time observations per MCD.

This UTPP travel time data is also available at the census tract and county levels of aggregation. Although more precise in terms of areal definition, census tracts are too numerous for the resulting travel time matrix to be adequately filled (3.2% full) with the UTPP sample. The large number of tracts involved makes the methods used to synthesize missing observations less reliable and more costly in terms of computational effort. Conversely, the county level of aggregation is too coarse to adequately support most transportation applications requiring travel time estimates. The MCDs provide a good compromise between travel time data need and data availability. Although of some use as received from the Census, this MCD data set must be expanded to full rank before it is appropriate for most scheduling and accessibility applications.

Synthesizing Missing Travel Times from Census Data

The basic motivation for the method used to synthesize missing travel times derives from the fact that census travel times are not randomly dispersed through the interdistrict travel time matrix. Rather, these observations follow commutation patterns as expressed by the manifest behavior of the employed resident who filled out the 1980 Census. Generally, these persons work in employment centers, and these centers are served by the highway system with workers following a sort of minimum path from home to work. It is in the nature of work trip length frequencies for observed work trips to be clustered in the shorter travel time ranges (Figure 2); there-

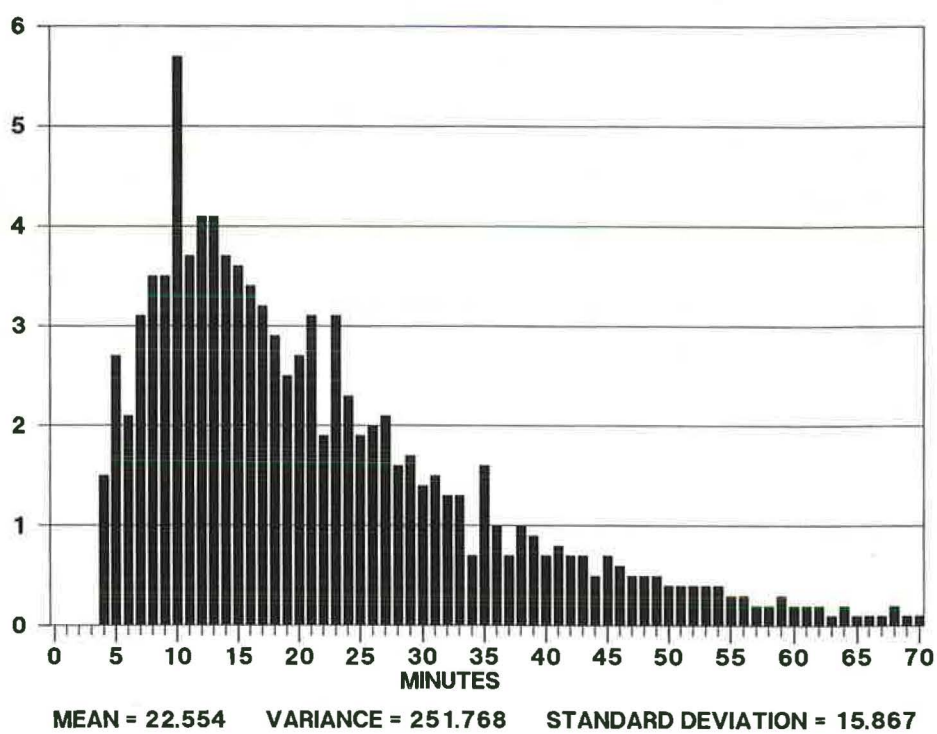

FIGURE 2 Trip-length frequency distribution of home-based work trips.

fore, it is the longer interdistrict interchanges that tend to be missing from the census data.

However, these missing observations may be synthesized from the census data by the creation of a network using each census observation as arc or one-way link. These arcs are then chained and summed into paths using a minimum path algorithm, thereby creating a synthetic time for most, if not all, missing observations. This process is similar to the spider network applications that were popular in travel analysis some years ago.

Figure 3 presents a visualization of this process. The Census has provided a drive-alone travel time estimated from MCD A to MCD B, and from MCD B to MCD C, but not from A to C. However, by chaining and adding observations A-B to B-C, one can synthesize a travel time estimate from A to C. This method, however, imposes special accuracy requirements on the census data in that spurious responses must be removed before synthesizing the missing travel times. Standard statistic techniques involving averaging individual responses do not apply in this application.

Editing the Census Travel Time Responses

The principal concern in editing the raw census travel time responses is to remove unrealistically small travel times from the file. Such a response, given the minimum path techniques used to synthesize missing travel times, could cause whole regions of the resulting matrix to have unrealistically small time estimates. For example, one employed resident from northern New Jersey reported that he made it to work in Honolulu, Hawaii, in 30 min. Even this extreme example is not necessarily a spurious response. Generally speaking, the census responses were contentious and quite good. This person may maintain a temporary or secondary residence in Hawaii 30 min from work. However, a 10-min estimate of travel time from, for example, Newark to Atlantic City, New Jersey, will have a disastrous impact on a synthetic matrix generated by minimum path methods.

The method used to edit the UTPP drive-alone travel time observations involved calculating an effective speed based on the airline distance between the home and work MCD. Those observed times that fall outside of an acceptable speed range were rejected. The airline distances were generated from grid coordinates of an arbitrary point located within each MCD, generally the geographic centroid. The edit procedure took each intercensus tract observation, calculated an effective airline speed based on the MCD of origin and destination, and rejected those observations that had an effective speed greater than 50 airline miles per hour (mph), or less than 4.9 mph. This corresponds to an over-the-road speed of 55 to 60 mph and 5 to 6 mph respectively, depending on the degree of circuity involved. The lower limit could be raised somewhat. However, many short commutes of 5 miles or less involve rather low speeds because much of the trip takes place on local and collector streets and relatively little on higher speed highway facilities. In any case, a minimum path algorithm can build paths around arcs with too low an effective speed. The

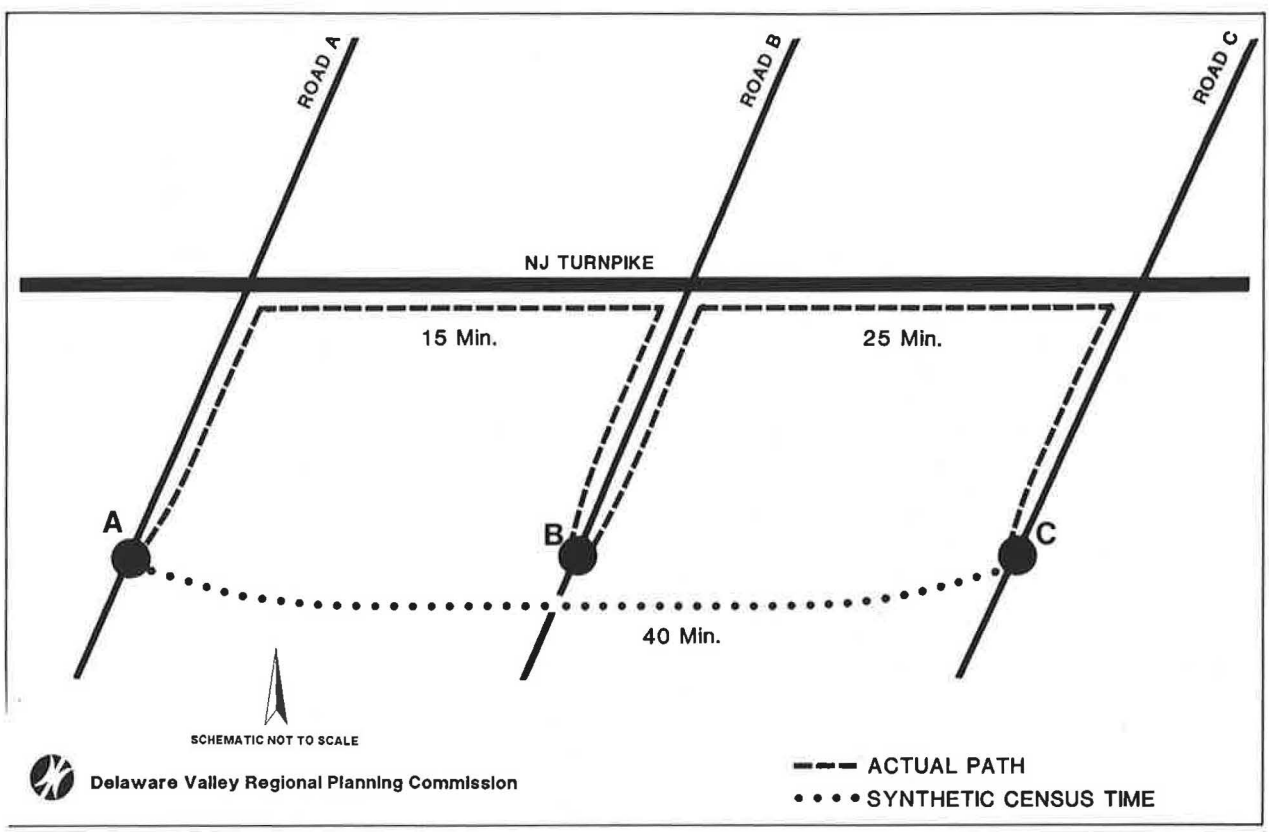

FIGURE 3 Hypothetical example of synthesized census travel time.

critical part of the edit is to remove census observations with excessive operating speeds. After the edit at the census tract level, the surviving travel time observations were compressed to the MCD level of detail, averaging the times in cases where multiple observations between two given MCDs existed. This edit procedure rejected 8,559 of 158,476 tract level travel time observations, or about 5.4 percent of the census data. After averaging to the MCD level of aggregation, 37,855 inter-MCD travel times were available to the path building algorithm.

A critic of this approach might contend that the synthetic travel times generated will tend to be over estimated because implicit local street access and terminal times will be duplicated for each arc added to the chain spanning the origin and destination MCD. In order to minimize chain length, all long-distance as well as short-distance travel time observations must be included in the calculation. Individual MCDs may have tens or even hundreds of travel time observations emanating from them. Through these long arcs, the average number of arcs per chain in the synthetic matrix was held to 3.75. The price we pay for this, however, is that the resulting network is nonsparse. Network sparseness refers to the percentage of the maximum number of one-way arcs (nodes × nodes) included in the network or the percentage of the node-link incidence matrix that is nonzero. Most packaged computer programs restrict the number of arcs out of a node to 4 or 8 to maintain sparseness and are therefore useless for synthesizing missing census travel times from UTPP data. By way of comparison, a traditional FHWA highway network may be about 0.02 to 0.3 percent full, whereas the spider network corresponding to the census travel time observations is 5.69 percent full. This manyfold increase in the fullness of the network has significant implications for the computer science techniques used to implement the Moore minimum path algorithm.

Minimum Path Building in Nonsparse Networks

The theorems defining the mathematical properties of the Moore algorithm apply to all levels of sparseness. The number of arcs per mode principally impacts the techniques used to implement this algorithm. It may be useful to consult the work of others on related problems, although most of the operations research literature on path building deals with sparse networks that have had much more computational success.

The Moore algorithm uses dynamic programming techniques to construct a tree or ordered sequence of arcs that connect the MCD of origin with all destination MCDs via a minimum time path. This tree is built one arc at a time, starting with the origin, always adding the minimum time arc emanating from the subset of active nodes. A node is considered active when it has been reached by the tree, but no arc emanating from the node has been added. Nodes not yet added to the tree and nodes that already have an outbound arc in the tree are considered inactive and therefore ineligible for link selection. This list of eligible arcs is often called a sequencing table and the problems caused by nonsparseness relate to the computational effort needed to maintain and update this table, which can grow quite large during the path building process.

Generally, these problems involve optimizing the tradeoffs between the computational savings derived from minimizing the size of the sequencing table and the computational costs required to edit its contents. The use of largely standard edit and merge techniques to maintain the sequencing table produced a method that is computationally practical, albeit much more expensive than tree building through sparse networks. The 816 district case study took a time equivalent to about 1.5 hours on an IBM 4381 computer to build minimum paths and skim the times and number of arcs over each path.

Variations on the methods described here to construct a spider network from census travel time observations are possible. For instance, a two-way link could be constructed from each home-to-work observation by attributing the same time to the reverse movement. This alternate approach would reduce the tendency to overestimate travel times because of path circuity, but would almost double the fullness of the resulting network, thereby exacerbating the tree-building problems discussed previously. In any case, there is no evidence in the following corroborative data sets that synthetic census times generated by the recommended approach are overestimated.

Network Method to Estimate MCD Travel Times

Travel times between MCDs may also be estimated by traditional network methods. Each MCD is given a centroid designation and a series of network links are defined, each representing a given street or highway segment with additional nodes placed at street intersections. Centroids are then connected to this street system via a series of approach links whose travel times are representative of approach travel over local roads (Figure 4). Travel speeds over the street links should in theory be estimated by floating car survey techniques. This technique is commonly used in traffic studies throughout the United States and Europe.

The principal problem associated with the travel time estimation with network methods vis-a-vis the census work trip times is the high cost of coding the network and measuring the speeds. The grain or density of network links should be at least equal to that of the district system boundaries. For MCDs, this implies that all freeway and major arterial facilities must be included with minor arterials and collector roads added as needed. For the 816 district example given above, this network would be large and expensive to survey.

Network and survey methods can be efficiently applied for smaller study areas, however. In 1986, such a travel time survey and network was prepared by the Delaware Valley Regional Planning Commission (DVRPC) for the portion of southern New Jersey, including Mercer, Burlington, Camden, and Gloucester counties (Figure 1). Although only a small portion of the 816 district system is covered by the census data, this network provides alternate MCD travel time estimates for comparison purposes. The network had the following characteristics: 107 districts, 800 nodes, and 1107 two-way links.

These districts were for the most part MCDs, although some small MCDs were missing, and cities such as Trenton and Camden were subdivided into smaller zones. For comparison with census data, zonal times were averaged to an MCD value and missing MCDs were dropped from the calculations. The minimum paths through this network produced travel-time

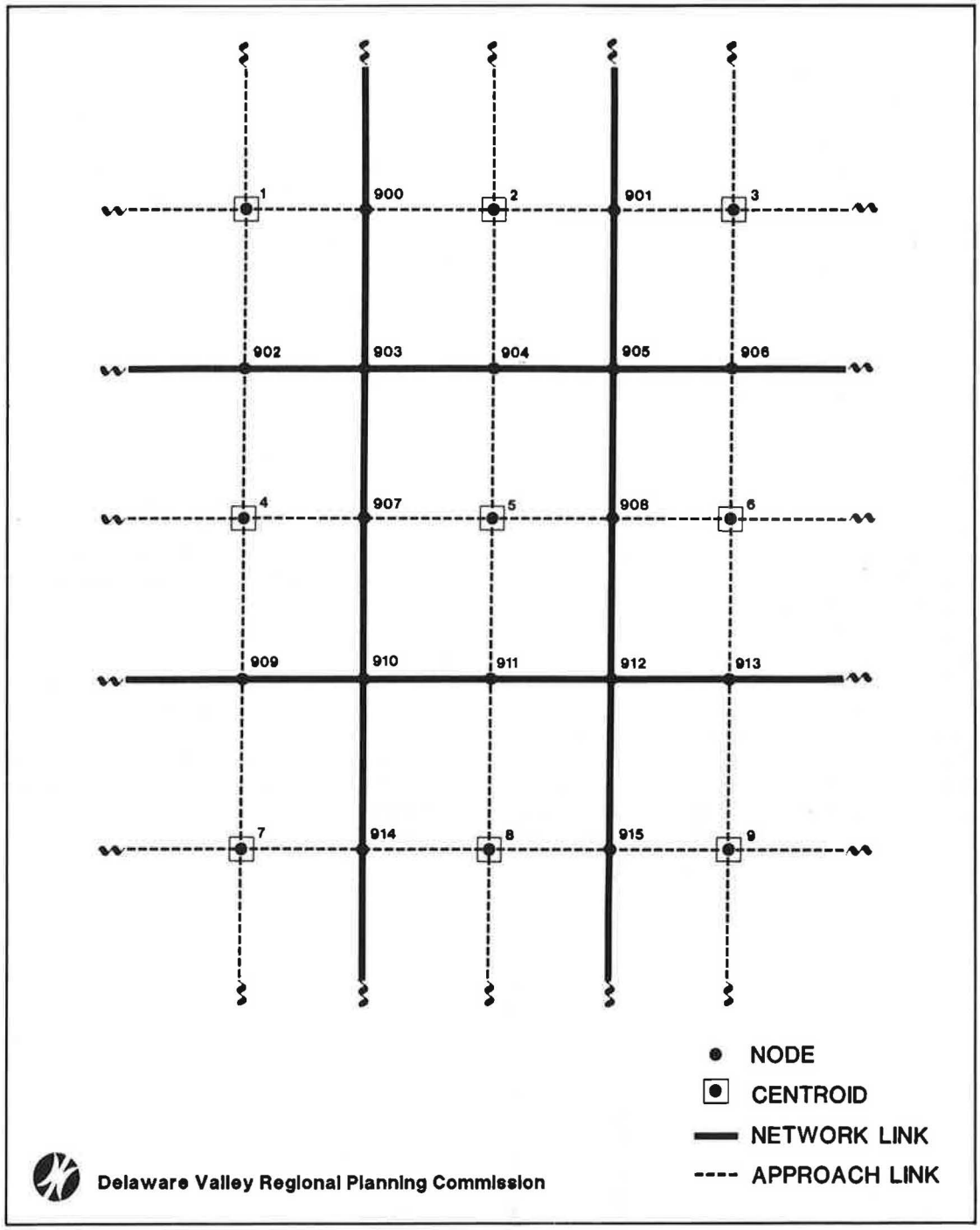

FIGURE 4 Traditional highway network.

estimates between 94 of the 112 southern New Jersey MCDs. In total, 30 percent of these links were provided with a measured speed. Unsurveyed links were given an average speed by route and area type from the survey (*6*). Peak and off-peak speeds in this survey were almost the same for most roads. The average peak and off-peak speeds were 27 and 28 mph, respectively. Some suburban highways now have slower speeds in the off peak.

Airline Distance–Based Method for Estimating MCD Travel Times

The most common method to estimate travel times in use in vehicle scheduling and market analysis studies is to use airline distance as a proxy for travel time. These distances can be calculated easily from grid coordinates by triangulation. With this technique, distances can be made specific, for example,

to the individual warehouses or fast food restaurants. This specificity is in general not possible in the census or network methods because of sample requirements and network size. However, we pay a price for this specificity in the accuracy of the times obtained. It is obvious that travel speeds are not the same in all directions from a given point.

For purposes of comparison with census times, the average speed (27 mph) from the network-based travel time survey was used after adjusting for the average circuity of the street system. A factor of 1.09 calculated by dividing total over-the-road network by total airline distance reduced this speed to 24.8 mph for the area covered by the network and travel time survey.

VALIDATION OF THE SYNTHETIC CENSUS RESULTS

Although it is possible to synthesize easily a complete travel time matrix from the drive-alone census journey-to-work travel times with the methods outlined previously, the accuracy of this synthetic data is unknown. In determining this accuracy, one cannot follow the usual approach of comparing synthesized with actual times because these two subsets of the travel times for the most part do not overlap. Synthetic census times exist for interchanges, with real census observations only in cases where a faster chain has been built around the census time. Rather, an indirect validation method must be used.

We will first compare the observed census travel times with the network-based and airline distance–based estimates discussed previously. The statistical indicators calculated from these comparisons will then form a bench mark for comparing the synthesized census matrix with these alternate travel time estimates. If the synthetic census times have comparison statistics similar to the observed census observations, then this is indirect evidence that the synthesized travel times are also reflective of actual conditions. When interpreting these comparisons it is important that only the observed census time data resulted from a random sample and therefore in theory has known statistical properties. The network and airline distance travel times are also synthetic. For this reason the comparisons between the observed census, network, and airline distance estimates are of interest in their own right.

The Bureau of the Census recommends the following variation on the standard error equation for use in estimating errors in census means

$$\sigma_{\bar{x}} = C\sqrt{\frac{\sigma^2}{n}}$$

where C is an adjustment factor that compensates for processing errors in coding and tabulating the census responses. C is not given for travel times; however, the value of this parameter varies from 0.8 to 3.4 for other items in the census. Another problem is that many travel time responses were previously averaged to census tracts, perhaps reducing the variation in the sample. For these reasons it is not possible to calculate the standard error for the census data.

A histogram of intertract census drive-alone travel time responses by minutes of total time is shown in Figure 5. These census times have been averaged in cases where multiple observations between tracts existed in the raw census data. Since these multiple observations tend to be concentrated in the lower time intervals, this averaging tends to shift the center of gravity of the histogram to the right compared to the network time–based trip length distribution shown in Figure 2. Although averaged, there are pronounced peaks in this

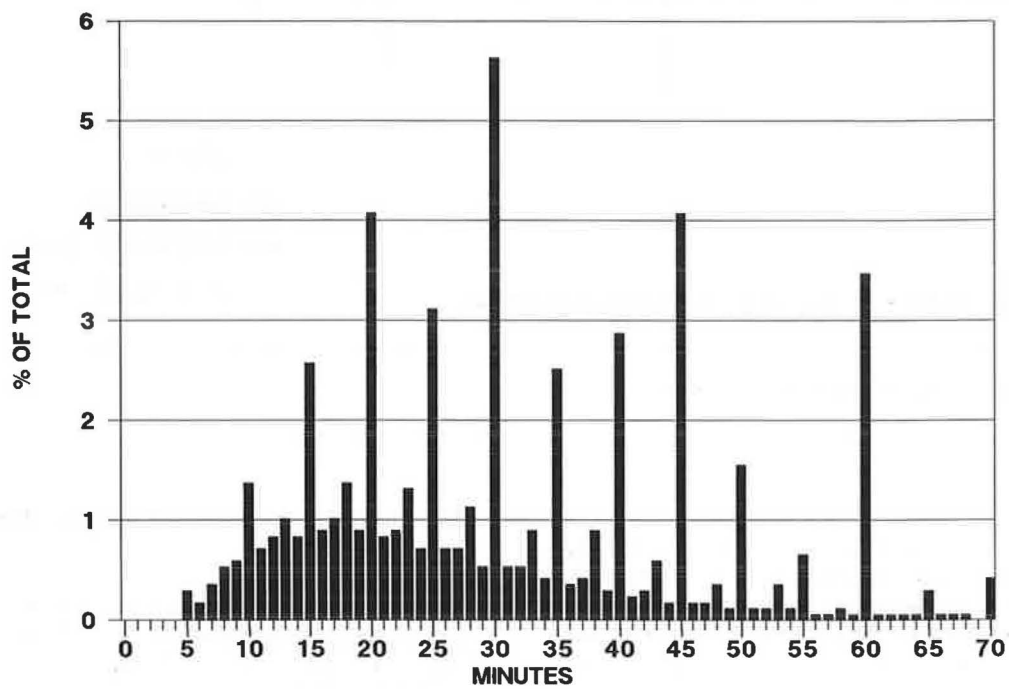

FIGURE 5 Observed census drive alone travel times.

data at even 5-min intervals and somewhat larger peaks at 15-min intervals. Respondents tended to round their journey-to-work times to the nearest 5 min and, to a greater degree, 15 min. This behavior is understandable given the probable daily variations in travel times. The journey-to-work time is known accurately by most people, given the importance of the work trip and the need to arrive at work on time, although individual variation in driving habits may be expected. The aggregation to the MCD level is also a source of travel time variation, especially in rural areas where MCDs are large. Approximately 10 percent of the travel time observations represent workers allocated to the geographic work place by the census because of incomplete address. Since reported travel time was a primary basis for determining allocation peer groups, this procedure will not introduce significant errors into the observed census travel time data (*7*). Rather, allocated observations will be associated with interchanges that already have similar times in the census data.

A statistical comparison between the observed census travel times and corresponding results of the network and airline distance methods are presented in Table 1. The R^2 measure of each is similar (0.67 and 0.70) and there is little difference between the census, network, and airline distance–based mean values. There is a substantial difference between the standard deviation of the airline distance and census travel times that is not found in the network/census comparisons. The airline distance method with its constant average speeds tends to overestimate the travel time for long trips and underestimate shorter trip times. The network-based times had an almost 8 min (30%) root mean square (RMS) difference. The corresponding RMS difference for airline distance times was somewhat larger—9 min or 35 percent.

Also included in Table 1 are three useful decomposition measures of mean square difference proposed by Theil (*8*). For the Theil tests, UM measures to the fraction of mean square error attributable to differences in the means, US refers to the fraction of this error resulting from difference in standard deviation, and UC estimates the corresponding fraction resulting from incomplete covariation (scatter). The Theil tests in Table 1 confirm that 98 percent of the observed difference between network and census times is attributable to scatter. The airline distance–based times had some 16 percent of the difference attributable to the relative standard deviations and 84 percent to scatter.

Figure 6 compares the prediction-realization diagrams for the observed census versus network times with the synthetic census versus network. From these diagrams it is apparent that the synthetic part of the census data tends to fill in the longer travel times of the matrix. As previously discussed, one might argue that the chaining method used to synthesize the missing values would overestimate the travel times for long trips by counting the terminal and local times for each link internal to the chain. The slight upward tilt of the data in Figure 4 shows that this has not happened. On average, long synthetic census times are slightly less than their network counterparts. This phenomenon is also reflected in the slight

TABLE 1 OBSERVED CENSUS VERSUS NETWORK AND AIRLINE DISTANCE–BASED TRAVEL TIMES

	Observed Census Versus	
	Highway Network	Airline Distance
Number of Observations	2,565	3,225
R^2	0.67	0.70
Difference between -		
Means (minutes)	- 1.07 (- 4.2%)	- 0.13 (0.5%)
Standard Dev. (minutes)	0.44 (3.4%)	3.58 (28.0%)
RMS Difference (minutes)	7.89 (30.7%)	9.00 (35.4%)
Theil Tests -		
UM	0.02	0.00
US	0.00	0.16
UC	0.98	0.84

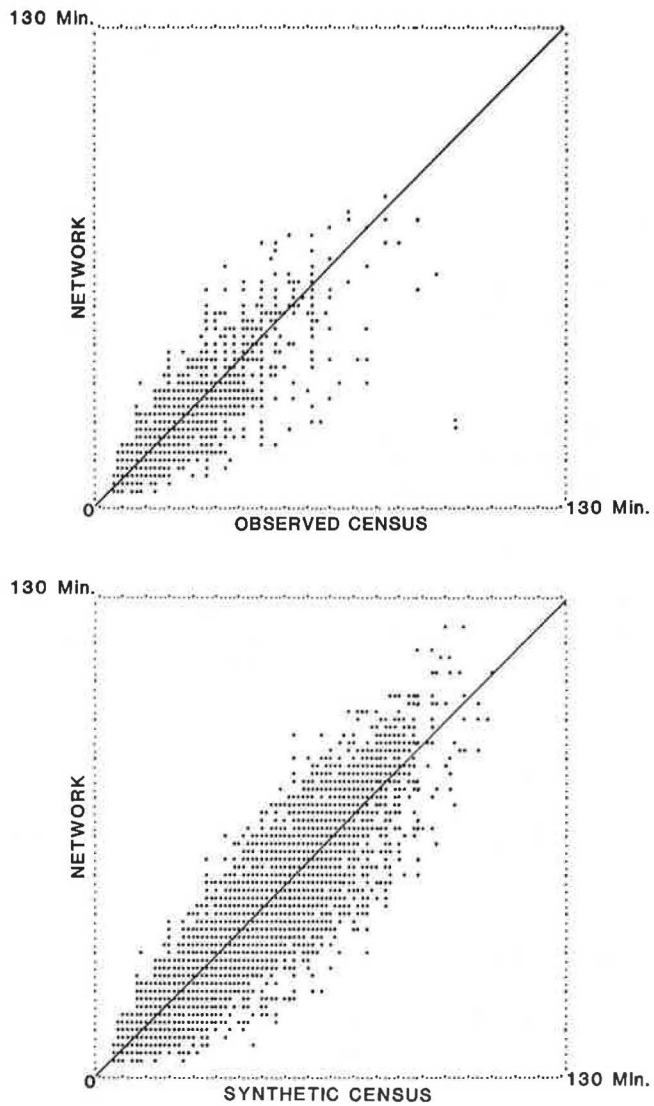

FIGURE 6 Observed census and synthetic census versus network-based travel times.

TABLE 2 SYNTHETIC CENSUS VERSUS NETWORK AND AIRLINE DISTANCE-BASED TRAVEL TIMES

	Synthetic Census Versus	
	Highway Network	Airline Distance
Number of Observations	8,742	12,320
R^2	0.84	0.85
Difference between -		
Means (minutes)	- 0.64 (- 1.6%)	6.38 (15.5%)
Standard Dev. (minutes)	1.97 (10.4%)	9.61 (50.8%)
RMS Difference (minutes)	8.49 (20.9%)	14.65 (35.7%)
Theil Test -		
UM	0.01	0.19
US	0.05	0.43
UC	0.94	0.39

increase in the US Theil statistic given in Table 2. This apparent overestimation of travel times by the network may result from a tendency to survey links in developed areas where network grain is finer and links more numerous, thereby imparting a small downward bias in the speeds shown in the table for unsurveyed links.

Except for the standard deviation, the statistical measures of the difference between the synthetic census and the network are essentially the same as the comparisons with the observed census data. This is strong indirect evidence that the accuracy of the synthetic part of the census matrix is comparable to the observed census travel times.

The comparisons between airline distance and synthetic census show a significant increase in the difference measures principally as a result of the strong tendency of the airline technique to overestimate long-distance travel times. A constant effective airline speed is inadequate. Effective airline speeds should be higher for long-distance travel and lower for shorter trips. Another use for the synthetic census data might be to estimate more accurate effective airline speeds. Such a speed could be calculated for the appropriate MCD interchange from the census time and airline distance matrices. This speed might then be applied to the precise airline distance in areas where MCDs are too large to supply the required geographical precision.

CONCLUSIONS

With the chain-building techniques explored here, one can synthesize a full matrix of MCD travel times from the 5 percent of these interchanges provided by the 1980 Census in the detailed journey-to-work data. Although incompatible with most transportation-planning software batteries because of the density of the spider network representing observed census times, missing census times can be synthesized through minimum path-building techniques in a cost-effective and accurate manner.

Comparisons between synthetic census and traditional transportation network travel time survey methods indicate that the synthetic census results are generally reasonable and representative of over-the-road travel times. Airline distance-based travel time estimates were found to be inaccurate because of a constant airline speed used to convert distance to over-the-road time. Overall speeds are slower for short movements and faster for longer trips, and vary significantly by direction of travel.

The process explored in the paper is capable of producing accurate, fine-grained travel time estimates from census data for large areas of the United States without the expense and difficulty associated with coding realistic highway networks and estimating link speeds through floating car surveys. This process, however, requires that adequate interdistrict connectivity be available in the census travel time observations. For the roughly 8 percent sample from the 1980 Census, work times must be aggregated to the MCD level (12 square miles on average) to achieve this. The sample size and the geographical coverage of the 1990 Census journey-to-work data will determine the grain and extent of the 1990 travel time matrix that can be generated with these methods.

REFERENCES

1. Highway Capacity Manual. *Highway Research Board Special Report 87*, Washington, D.C., 1965.
2. Highway Capacity Manual. *TRB Special Report 209*. TRB, National Research Council, Washington, D.C., 1985.
3. A. Talvitie and Y. Delghani. Comparison of Observed and Coded Network and Travel Time Measurements. *Transportation Research Record 723*, TRB, National Research Council, Washington, D.C., 1979, pp. 46–51.
4. D. E. Boyce and R. W. Eash. The Effect on Equilibrium Trip Assignment of Different Link Congestion Functions. *Transportation Research, Part A: General*, Vol. 15a, No. 3, pp. 223–232.
5. *Urban Transportation Planning Package: 1980 Census Technical Documentation for Summary Tape*, U.S. Bureau of the Census, Washington, D.C., 1983, p. H-4.
6. 1971–73 to 1986 Highway Travel Time and Speed Comparisons in the New Jersey Portion of the Delaware Valley Region. *Transportation Data Bulletin No. 9*, Delaware Valley Regional Planning Commission, Philadelphia, Pa. 1987, p. 10.
7. P. N. Fulton. Procedure Used for Allocating Incomplete Place of Work Responses in the 1980 Census Urban Transportation Planning Package. *Transportation Research Record 981*, TRB, National Research Council, Washington, D.C., 1984, pp. 15–20.
8. H. Theil. *Applied Economic Forecasting*. Rand McNally, Chicago, Ill., 1966, pp. 29–36.

Publication of this paper sponsored by Committee on Transportation Data and Information Systems.