THE CREATION OF POINT-BASED HOUSEHOLD AND JOB GEOGRAPHIC DATASETS FOR TRAVEL MODELING: DISCUSSION OF THE DEVELOPMENT PROCESS AND OF MODELING BENEFITS

Erik Sabina, P.E., Regional Modeling Manager
DeVon Culbertson, Information Technology Services Manager
Jill Locantore, Demographic Analyst
The Denver Regional Council of Governments
4500 Cherry Creek Drive South, Suite 800
Denver, CO 80246
303-455-1000 (voice)
303-480-6790 (fax)
esabina@drcog.org
dculbertson@drcog.org
jlocantore@drcog.org

Jeremy Papuga, Transportation Planner
Pima Association of Governments
177 North Church Avenue, Suite 405
Tucson, AZ, 85701
(520) 792-1093 (voice)
(520) 620-6981 (fax)
jpapuga@pagnet.org

Abstract

The Denver Regional Council of Governments is conducting a project to develop a disaggregate, activity-based travel modeling system for use in regional planning. Many of the advantages of such models are derived from their disaggregate approach to decision-modeling, and the disaggregate demographic data required to support this approach. Population synthesis permits disaggregate specification of household and person characteristics. However, household and job location in the model typically are known only at the TAZ level. This limitation leads to aggregate specification of numerous key variables, such as transit walk access distance and walk mode distance, or requires the use of complex, non-intuitive approaches to reducing aggregation error through zonal subdivision. The project team concluded that it was practical to assign an x-y location to each household and job in the region, and that this approach would produce a variety of benefits both in model estimation, and in model application. Example benefits include simpler, more accurate calculation of short-distance trip lengths and transit access distances/times, which in turn significantly improve mode choice model outcomes for transit and non-motorized modes. Model scenario analysis also is greatly improved, as it is possible to effectively depict, for example, the transportation effects of meso-scale land use approaches such as transit-oriented development, and especially the extent to which such developments contribute to transit and non-motorized mode use. This paper describes the process of developing these x-y values, including the necessary input data employed to combine disparate input datasets into a reasonable point location distribution of households. Deficiencies in the input data are discussed, together with methods of mitigating those deficiencies. Summary results of the distribution assignment process are described. Model application benefits resulting from this approach are discussed. A comparison of this approach with an analogous parcel-based approach in another similar model project is briefly outlined. Finally, potential enhancements to this approach are discussed, with possible application to future projects.

The Integrated Regional Model Project

The Denver Regional Council of Governments, together with project partners the Regional Transportation District (RTD) and the Colorado Department of Transportation (CDOT) is conducting a project to replace its aggregate socioeconomic and transportation models with a next-generation, integrated system of models, including an activity-based travel modeling system. DRCOG is being assisted in the project by a consultant team led by Cambridge Systematics, with Parsons Transportation Group, and with Mark Bradley, John Bowman, and Connetix. In the project Vision Phase, DRCOG selected a model design that may be characterized as belonging to the “San Francisco family of models.” These models are, in a sense, descendants of the activity model built for Portland METRO, enhanced for the San Francisco County Transportation Authority and again for the Sacramento Area Council of Governments.

The project was “kicked off” in December, 2005, and currently is expected to be complete in June, 2008. Key tasks and milestones include:

- Model estimation data – begun in the Spring of 2006, and completed in Winter 2007. Estimation data was created at the most disaggregate level possible. For example, point locations were acquired for all schools in the region, both public and private, together with enrollment at each. However, standard methods of producing skims (highway and transit) from travel model networks are still aggregate, from zone centroid to zone centroid.
- Model estimation – this task is now underway, and expected to be completed by October, 2007.
- Software development – model architectural design was completed in 2006, and software development has been ongoing since that time. The core software architecture is now complete in draft form, and several draft components have been built and integrated with this system for testing purposes. Modifications to the TransCAD GISDK code that embodies DRCOG's current four-step model also are in process, and will be used for a variety of steps such as transit path building, skimming, assignment, etc.
- Model calibration/validation – this task is scheduled to commence in October 2007, when model estimation and software development are scheduled to be complete. A six-month process is anticipated, leading to a finished model in late Spring 2008.

A more complete description of model development status is provided elsewhere.[1] The IRM project is being conducted in the context of a broader initiative at DRCOG, toward the development of a comprehensive set of disaggregate data for the region, describing its development patterns, both business and demographic, and its travel characteristics. Known as the Regional Data Model (RDM), the disaggregate data used and produced by the IRM will integrate with the RDM in a seamless fashion, through the use of industry-standard software that connects the IRM to the RDM, stored primarily in Microsoft SQL Server.

Model Design Motivation – Regional Planning Initiatives

During the IRM Vision Phase that informed the IRM model design, attention was focused on the major planning initiatives, both current and anticipated, that were considered of high importance in shaping the region's future. The Vision Phase evaluated likely activity-based model design alternatives and their modeling capabilities in the light of these key planning issues. IRM design decisions were made to maximize the model's ability to address these issues, consistent with practicalities of model design, construction, and operation. Key planning issues included:

- Highway toll facilities. The Colorado Tolling Enterprise (CTE) was established three years ago by the state legislature, and has been working to identify a set of corridors with the potential for toll facility establishment. The CTE has identified about six such corridors in the Denver area and is conducting its own evaluation of these corridors. DRCOG has been told to expect these, or some subset of them, to be submitted to the regional planning process for inclusion in the regional plan. These efforts also have caused planners conducting several Environmental Impact Statements in the region to take a harder look at toll options in their alternatives analyses.
- The effects of MetroVision Urban Centers and other transit-oriented developments. Support of such development patterns is intended to foster a more balanced transportation system, reducing the number and lengths of trips, foster additional bicycle/pedestrian use, etc. The MetroVision 2030 Update developed in 2004 included approximately 70 such

centers, and the evaluation of the effects of these centers is a key aspect of the regional model's usefulness. These will be evaluated again during the MetroVision 2035 process now underway.
- Effects of the MetroVision Urban Growth Boundary. The extent of the Urban Growth Boundary/Area currently is set at approximately 750 square miles for the year 2030, and the extent to which it may need to be expanded for 2035 will be a key part of the MetroVision 2035 process.
- Re-examination of lower-density development, referred to as "semi-urban." Issues include defining semi-urban, estimating how much of it now exists, how much should there be, and what are its transportation and air quality effects.
- The FasTracks ballot initiative of 2004. Passage of this initiative kicked off a project to build about 130 miles of rapid transit to all parts of the region by the year 2017. The ability to evaluate the effects of such a system will be critical over the next decade.
- Air quality. As always, evaluation of the effects on air quality of various policy/transportation initiatives will continue to be a key issue in regional planning.
- Highway project planning. This also will continue to be a core focus of the planning process in the region.

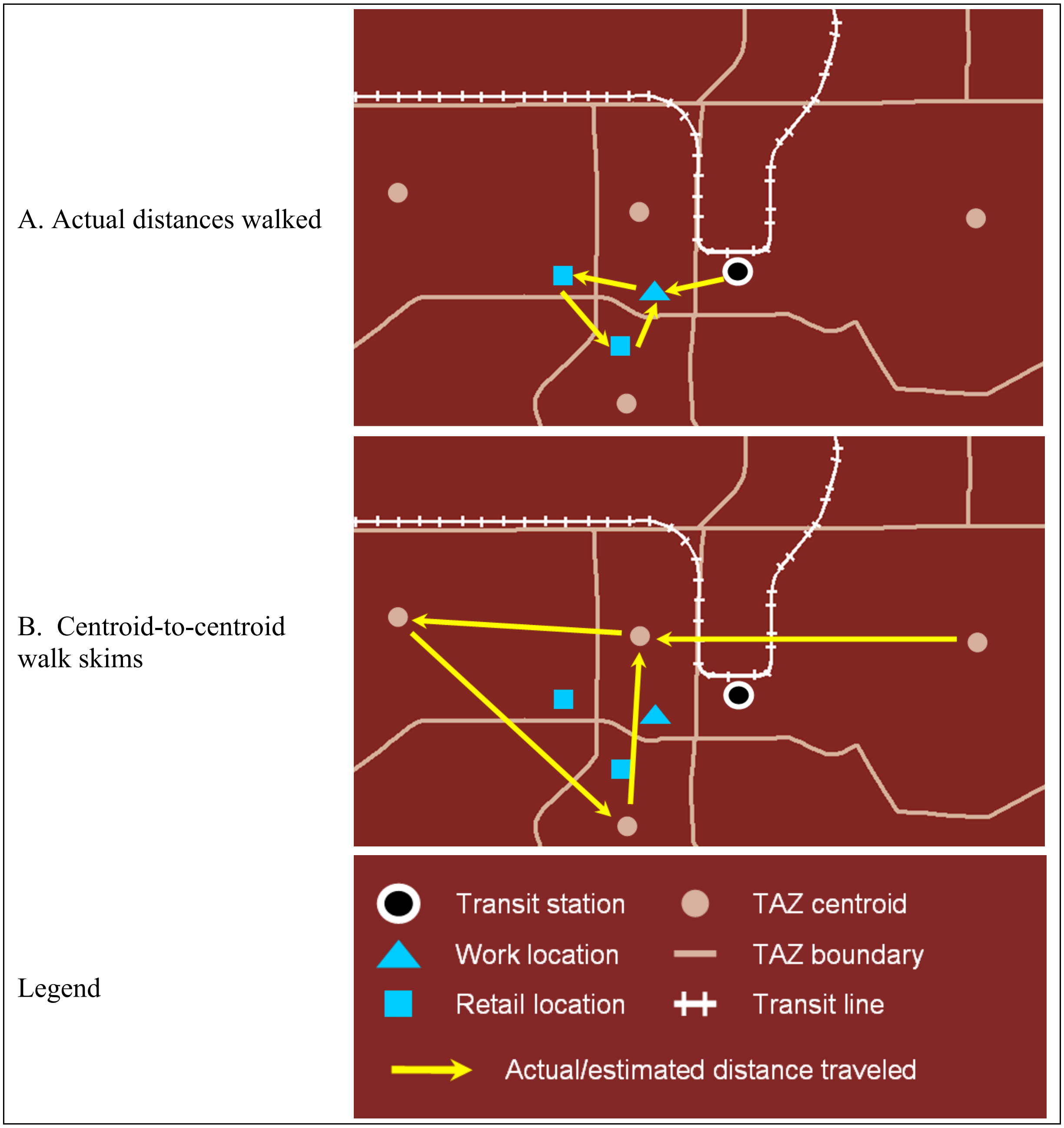
Shortcomings of Aggregate Modeling Approaches to Walk Trips

Zone-based skims, the standard output of the skim processes of all major travel modeling software packages, calculate time, distance and other variables between artificial "centroids" for each zone pair. Figure 1 shows an example of the error inherent in calculating centroid-to-centroid walk skims, showing the zone system in a planned urban center on a planned passenger rail line in Lakewood, a suburban town west of Denver. A hypothetical example is shown of a person arriving at the train station, walking to his/her place of employment, and then making a walking errand tour from there over lunch. As the figure shows, a centroid-to-centroid approach to skim development can greatly overstate the total walk distance for such a tour.

The errors inherent in the centroid-to-centroid approach are small (on a percentage basis) when the distances are long (usually the case with auto trips) and the zones are small (usually the case in the urban core.) However, walk trips typically are short, and therefore poorly handled by such a system, even in areas of small zones, as the errors caused by centroid-to-centroid skimming are often a large fraction of the true skim value. Walk access/egress trips (transit or auto) suffer from the same problem, and worse: it is likely that centroid-to-transit stop, walk-to-transit skims suffer from bias, rather than simply from large, but random errors. This is because, especially in areas of large zones, the distribution of destinations accessing/egressing transit often is not smooth across the zone, but rather skewed toward the transit stop. In such cases, the centroid-to-centroid skims are longer than the true mean skims for the zone.

One approach to mitigating these bias effects is to estimate models on skims built in the same way that the operational model will build them. This would mean using centroid-to-centroid skims in the estimation dataset, including skims for walk and walk-to-transit trips. However, it is common for a household survey dataset (the typical basic dataset for model estimation) to have survey places (household location, work location, other place locations) geocoded to the xy location level. Estimation using centroid-to-centroid skims means, in effect, throwing away this detailed and very valuable data.

Figure 1. Error associated with centroid-to-centroid walk skims



A. Actual distances walked

B. Centroid-to-centroid walk skims

Legend

Finally, zone-based modeling dramatically reduces the model's ability to evaluate the effects of "meso-scale" development pattern planning policies such as transit-oriented development (TOD.) Location/distribution of TOD development is extremely crude in a zone-based system, with locations limited to the centroids, together with the assumption that all development in the zone is located at that one point. The overall effect of these limitations is to do a poor job of evaluating TOD (and other similar development patterns), very possibly understating their effects to an unknown degree. Overstatement of walk distances means underestimating walk and transit trips, which means overestimating roadway VMT, which means overestimation of air pollutant emissions, which means underestimation of the benefits of TOD and urban centers.

The IRM project team investigated several methods of mitigating these effects. Two that were investigated in detail were:

- To further subdivide TAZs into smaller “virtual” TAZs for walk and transit walk access skimming. This approach essentially has the same deficiencies as the zone-based approach, but less severe.
- To place xy destinations in “bands” of distance from transit stops. XY locations in the estimation dataset would be assigned to these bands (1/4 mile, _ miles, etc.), and development patterns in scenarios run with the operational model also would be distributed in these bands. Such a system would reduce the transit walk access distance error inherent in a zone-based approach. Unfortunately, another approach would be required to manage walk-only trips, raising the likelihood that the walk trip approach would be inconsistent with the transit walk access approach, leading to potentially damaging inconsistencies in model estimation and operation.

In the project team’s efforts to find a way out of these difficulties, we began to consider the point-based development data that DRCOG has been developing in recent years. DRCOG has been developing point data for employment locations for some years. DRCOG also develops point data for housing units. However, additional work would be required to transform housing unit point data to household point data.

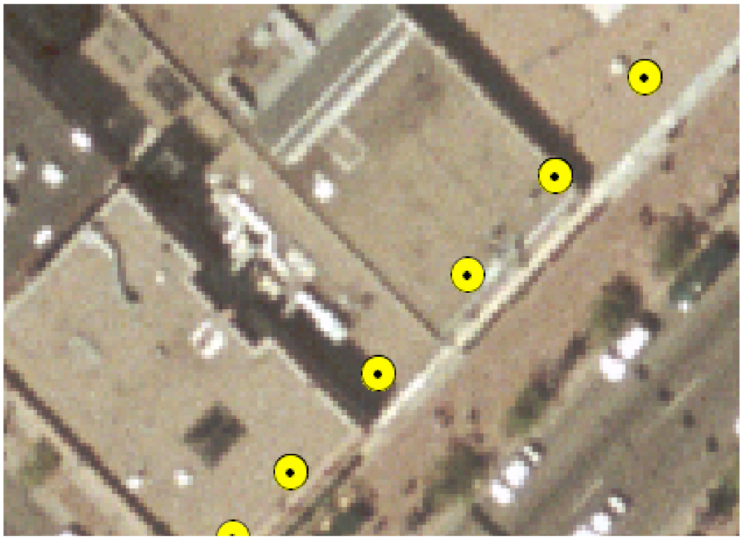
Employment Points

The development of XY coordinates representing the location of firms and employees within the Denver region has been relatively straightforward, if tedious and time-consuming in some cases. In fact DRCOG has been developing annual point-level employment data since the year 2000. The primary source of this data is Quarterly Census of Employment and Labor (QCEW), a cooperative program involving the Bureau of Labor Statistics of the U.S. Department of Labor and the State Employment Security Agencies. The QCEW program produces a comprehensive tabulation of employment and wage information for workers covered by state unemployment insurance laws and federal workers covered by the Unemployment Compensation for Federal Employees program.[2] DRCOG obtains detailed QCEW data, including the street address, zip code, and number of employees associated with each firm in the Denver region, from the Colorado Department of Labor each year.

DRCOG is therefore able to generate an XY coordinate for most firms simply by geocoding the address and zip code reported in the QCEW. Government employers, including schools, post offices and local governments, require some additional data manipulation, however. The QCEW typically includes consolidated records for these employers – e.g., one address and employment number for an entire school district. In these cases, DRCOG surveys the relevant agency to obtain more detailed information – e.g., the location and number of employees at each school within the district. The QCEW also includes both consolidated "parent" records for large private firms and individual "child" records representing different locations of the firm. The parent records must be filtered out to avoid double counting of employees.

DRCOG's 2000 employment point database includes XY coordinates for approximately 85,000 firms, representing approximately 1.4 million jobs. Figure 1 below shows a small sample of the employment points displayed over aerial photography, illustrating the spatial accuracy of the data.

Figure 1. Illustration of Employment Point Accuracy



Household Points

The development of XY coordinates representing the location of households in the Denver region was more complicated, requiring the blending of data from a variety of sources. Point-level data indicating the location of residential utility hook-ups is available from the various utility companies that serve the Denver region, including XCEL Energy, United Power, Inc., Poudre Valley Electric, Mountain View Electric Association, Morgan County REA and Intermountain REA. However, these data suffer from two main shortcomings. First, the earliest data available is from the year 2002, whereas DRCOG's model estimation base year is 2000. Second, the data do not differentiate between vacant and occupied housing units; obviously only occupied housing units are of interest within the context of travel modeling.

The IRM team was able to overcome these shortcomings by drawing upon three additional data sources: DRCOG's annual population and household estimates, 2000 US Census data (Summary File 1), and TIGER/Line files indicating the spatial location of Census tracts and blocks. DRCOG's annual population and household estimates start with the 2000 Census and grow the population each year based on building permit data obtained from local governments, a quarterly vacancy rate survey sponsored by the Apartment Association of Metro Denver[3], and assumptions regarding average household size. From 2002 forward, DRCOG has used point-level residential utility hook-up data to determine the spatial distribution of new housing units within each jurisdiction, producing final estimates at the Census tract level.

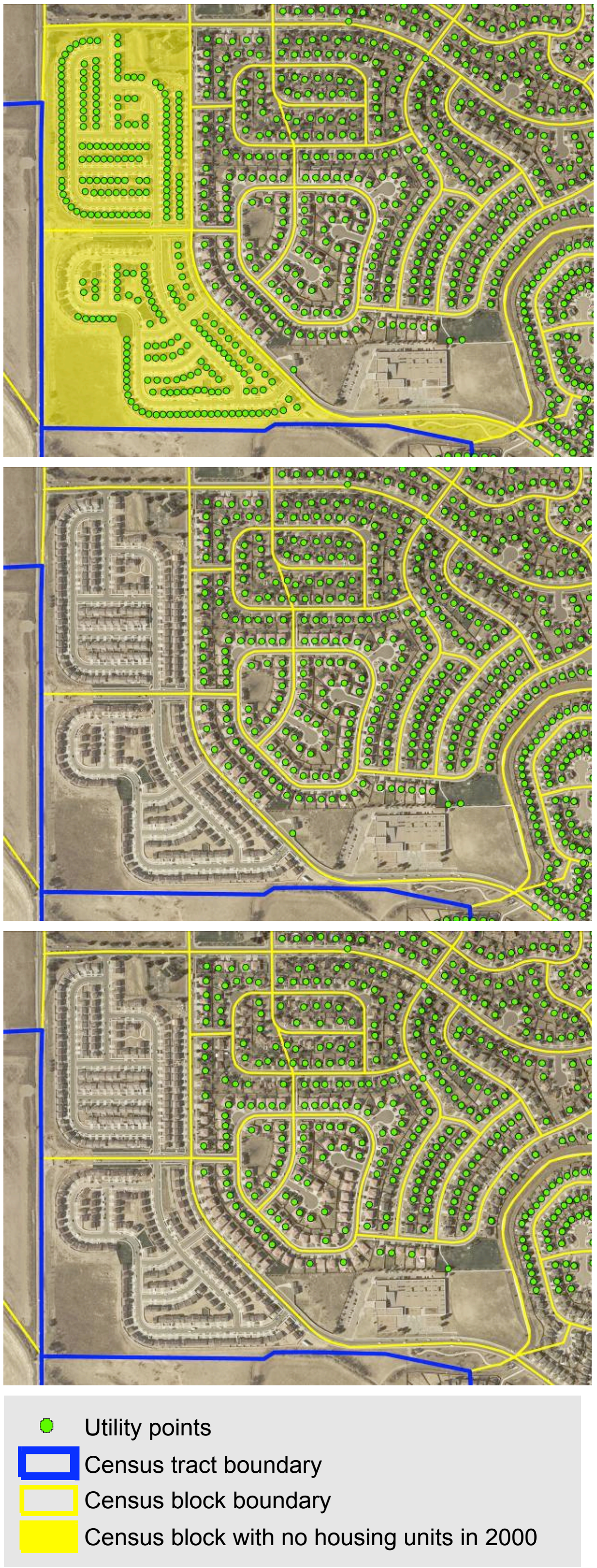
The first step in creating point-level household data for the year 2000 was adjusting the 2002 residential utility hook-up data to match housing unit control totals from the 2000 Census. We compared DRCOG's 2002 tract-level housing unit estimates with the 2000 Census to identify tracts where the number of units differed between the two years. If a tract contained more housing units in 2002 than 2000, we identified blocks within the tract that had no housing units in 2000 and removed all of the utility points located within those blocks (see Figures 2A and 2B). Such blocks represent areas of new housing construction between 2000 and 2002. If the total number of remaining utility points located within the tract still exceeded the number of housing units in 2000, we randomly removed additional points so that the total number located inside each block within the tract matched the 2000 Census control total (see Figure 2C). If a tract contained fewer housing units in 2002 than 2000, a random subset of points within the tract were duplicated so that the total number of points equaled the 2000 Census control total.

The second step was to adjust the utility points to represent only occupied housing units. We again used the 2000 Census to identify tracts where the vacancy rate was greater than zero, and randomly removed utility points located within those tracts so that the total number of points equaled the total number of occupied households.

One shortcoming of our methodology relates to small scale spatial inaccuracies in the TIGER/Line data from the Census. A comparison of the TIGER/Line data to aerial photography suggests that at the tract level, the line-work fairly accurately corresponds with the roads and other dividing lines that form tract boundaries. Accuracy is less consistent at the block level, however; Figure 3 illustrates a particularly egregious example. Such inaccuracies may cause the appearance of a mismatch between the number of utility points and the number of Census housing units within a given Census block, when in fact none exists. In the absence of an alternative data source, an immediate solution to this problem is not obvious and therefore our final 2000 household point data contains a small but tolerable level of noise. DRCOG will be working closely with the US Census on future updates to the TIGER/Line data, to ensure a higher level of spatial accuracy.

Other organizations considering the development of point-level household data could reduce the amount of work required and improve the quality of the data by changing at least two aspects of our methodology. First, using utility point data from the same year as the model estimation base year would eliminate the need to grow or shrink the utility point data to match the model estimation base year. Second, obtaining or developing more accurate spatial data indicating the location of Census tracts and blocks would decrease errors resulting from a mismatch between Census line-work and the actual location of streets and housing units.

Figure 2. Adjustment of 2002 Utility Points to Match 2000 Census Control Totals



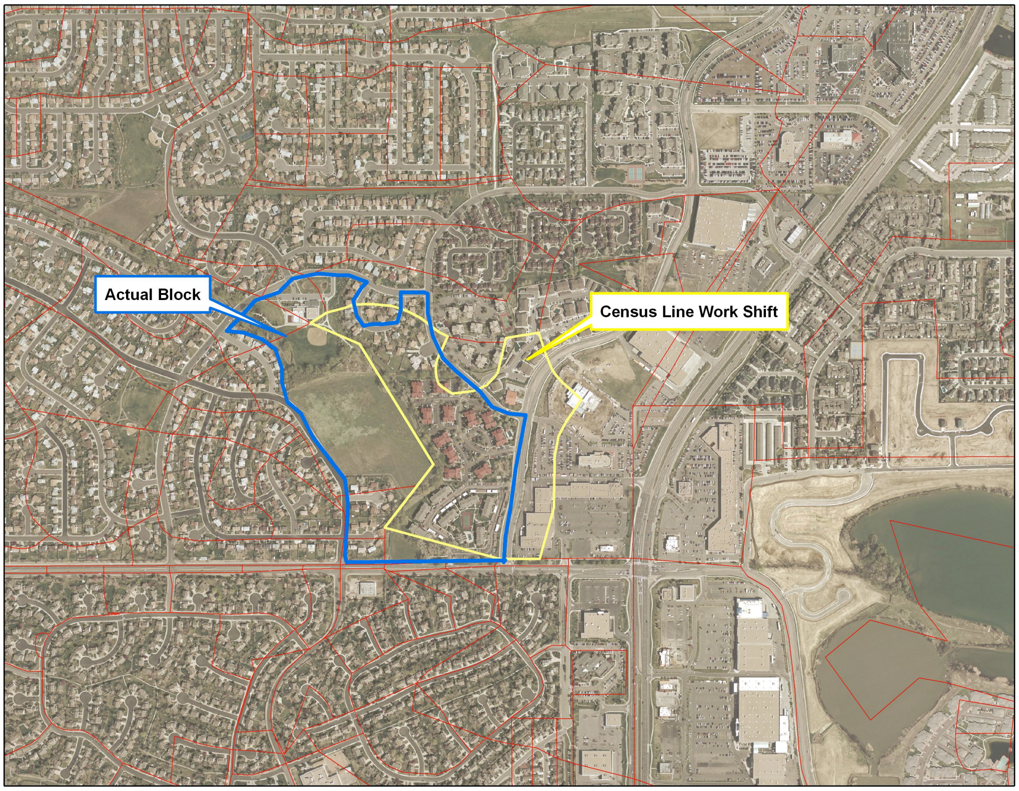
A. Identification of Census blocks with no housing units in 2000

B. Removal of utility points from Census blocks with no housing units in 2000

C. Random removal of additional utility points to match 2000 Census block housing unit control totals

Legend

Figure 3. Inaccuracies in Census TIGER/Line Data



Use of the Point-Based Development Data

Using point-based data for the model's development dataset solves many problems "at a stroke":

- It permits model estimation using the XY point data available in the household survey, removing consistency problems between model estimation and operation.
- It permits detailed, disaggregate calculation of walk distances, greatly enhancing model performance.
- It enables full detail in the depiction of development pattern scenarios such as TOD.
- It opens the door to further improvements. For example, while choice models in the IRM still will be run at the zone level (with tour locations assigned an xy location in the zone through a Monte Carlo process), point data makes it possible to run the choice models themselves at the point level at some point in the future. The connection of a point-based model to simulation software also is theoretically enhanced, as the highly detailed depiction of traffic patterns is much better supported.

Generally speaking, point-based data deemphasizes the use of zones in the model, and depiction of development in this way is represents a large step toward a fully disaggregate model system. However, many difficulties remain to be addressed. For example, effective methods of distributing point data for future year scenarios are neither obvious nor settled. Example issues include:

- How should existing (base year) development point data be managed in a future year scenario to take redevelopment into account, as redevelopment implies the relocation or elimination of some existing points?
- What information is available to, in effect, “build” a suburban neighborhood for a future year scenario, in a presently undeveloped area (recognizing that the xy pattern of households is framed around a road network and other details of neighborhood design.)
- What ratio of single versus multi-family households should be included in the dataset, and how should the geographic pattern of these two types of households be developed?

These and many other problems should be kept in perspective, however. Many such “problems” truly are opportunities in reality: modelers have not had the opportunity in the past even to worry about how they would depict future scenario development patterns in full detail. It should also be recognized that existing zone-based models *already* give an XY point to every job and household in the dataset: that point is the location of the centroid of the zone. This is a very low standard, and easy to improve upon. The use of point-based data, with its strong fidelity to a true representation of development location, opens the door to a nearly unlimited field of model improvements, which can be pursued incrementally, in response to planning needs as they arise.

References

[1] Cambridge Systematics, Inc., John Bowman, and Mark Bradley. *DRCOG Model Design Plan*. January 2007.

[2] U.S. Department of Labor Bureau of Labor Statistics, “Quarterly Census of Employment And Wages (ES-202) Program Overview,” http://www.bls.gov/cew/cewover.htm

[3] Gordon Von Stoh, "Denver Metro Area Apartment Vacancy and Rent Survey,” http://dola.colorado.gov/cdh/vacancy/metrodenverintro.htm