

Date: August 26, 1998

To: William J. Upton

From: Paul Waddell

Cc: Bill Davidson
Doug Hunt
Rick Donnelly
Pat Costinett
Larry Conrad
Bud Reiff

Re: Technical Memorandum 1, Task 3E:
1980 Data Preparation for Longitudinal Calibration

1 Overview

This technical memorandum is the first of three specifying data needed for input to the Eugene-Springfield longitudinal calibration¹ of the metropolitan model. Based on the scheduled completion of Task 3E by the end of December 1998, data preparation needs to be concluded by the end of September in order to leave adequate time for the longitudinal calibration of the model and preparation of a final report. While this is a very condensed schedule, it is facilitated by the amount of data preparation and assimilation undertaken by LCOG during the spring of this year in anticipation of this task.

Three types of data are needed for Task 3E. The schedule for completion of the data preparation is as follows:

- ◆ 1980 data for input to the model September 25, 1998
- ◆ Interim year data for input to the model October 2, 1998
- ◆ Interim year data to be used as calibration targets October 2, 1998

Three technical memoranda will provide detail specifications for the preparation of these data. We understand that ODOT staff will be preparing the data, with

¹ This task has been previously labeled as 'historical validation'. Discussion at the Peer Review Meeting in New York suggests that this process would be more accurately described as 'longitudinal calibration', since it involved making adjustments to the model parameters and specification in addition to providing information about the longitudinal ability of the model to replicate historical trends.

some assistance from Larry Conrad and LCOG. Larry will be working directly with ODOT staff throughout the data preparation process. Upon completion of each of these three data preparation subtasks the results need to be documented by ODOT and Larry Conrad in memoranda that incorporate the original specifications and provide detailed results.

The data preparation process, then, will be documented by the following technical memoranda, which will form the first sections in the final report for Task 3E:

3E.1	Specifies 1980 data needed as input to the model	August 26, 1998
3E.2	Specifies interim year (1981-94) input data	September 1, 1998
3E.3	Specifies interim year data for calibration targets	September 11, 1998
3E.4	Documents results of work specified in 3E.1	October 9, 1998
3E.5	Documents results of work specified in 3E.2	October 16, 1998
3E.6	Documents results of work specified in 3E.3	October 16, 1998

The following is a description of the input data needed for 1980, and specifies data formats for each of the input files.

The major files needed for 1980 are:

- 1980 Parcel Database
- 1980 Business Establishment database
- 1980 Household data from the census
- 1980 Urban Growth Boundary
- 1980 Land Use Plan
- 1980 Development Cost Estimates
- 1980 Travel Model Outputs

Each of these data files and their specification are discussed below.

2 1980 Parcel Database

A 1980 parcel database has been retrieved from archives by LCOG that appears to contain the appropriate data for creating a 1980 base file. The process of creating the database for input to the metropolitan model involves extensive GIS overlay

processing to integrate environmental constraints into the parcel database². This should be done using Arc/Info polygon overlay operations (union), resulting in sub-parcel polygons with unique combinations of the parcel layer and each spatial overlay. Due to this fragmentation of the parcel polygons the parcel database is specified in two files. The first file is the result of the spatial overlay process in the GIS, and contains the parcel number, the area of each polygon, and the overlays used. This file will often have multiple records per parcel due to the GIS overlay operations. The second file will contain all of the attributes of the parcel, but will be retained in its original format of one record per parcel (not split by the overlays). The data preparation component of the software manages the integration of these files, and allocates housing units and commercial square footage to the sub-parcel polygons in proportion to their area.

The following specification is for the result of the spatial overlay, which should be produced as an Arcview shape file named *Parcel80*, containing the following fields in the dbf attribute table.

Name	Format	Description
Area	Float	Assigned by ArcView
Perimeter	Float	Assigned by ArcView
Parcel_	Integer	Assigned by ArcView
Parcel_ID	Integer	Assigned by ArcView
Parcel	Character	Unique Parcel ID
Xcoord	Float	X Coordinate of Parcel
Ycoord	Float	Y Coordinate of Parcel
Floodplain	Integer	Dummy (1 if floodplain/0 otherwise)
Streambuf	Integer	Dummy (1 if streambuf/0 otherwise)
Wetlands	Integer	Dummy (1 if wetland/0 otherwise)
Slope25	Integer	Dummy (1 if slope>25/0 otherwise)
Transbuf	Integer	Dummy (1 if transbuf/0 otherwise)
UGB	Integer	Dummy (1 if outside UGB/0 otherwise)

The second parcel file, with one record per parcel, is specified as follows, and named *Parceldata80.dbf*.

Name	Format	Description
Parcel	Character	Unique Parcel ID
Lucode	Character	Detailed Land Use Code
PLU	Integer	Planned Land Use
County	Integer	County Code

² We assume that the environmental GIS layers that were prepared for 1994 have not changed since 1980, and propose that the same environmental layers be used in preparing the 1980 base.

City	Integer	City Code
Blkgrp	Char	Census Tract / Block Group (1980)
Zone	Integer	Traffic Analysis Zone
Stnumber	Integer	Street Number
Stprefix	Character	Street Prefix
Stname	Character	Street Name
Stsuffix	Character	Street Suffix
Acres	Float	Land area of parcel
Landval	Float	Total Land Value
Impval	Float	Total Improvement Value
Units	Integer	Total Housing Units
Sqft	Integer	Total Building square footage
Yrbuilt	Integer	Year of Construction
ALU	Integer	Aggregate Land Use Type

ALU values are aggregated from the detailed land use codes in the parcel file (could use either Statclass or Lucode – statclas was principally used in 1994 file). The ALU types used in calibration are:

	ALU	Description
Residential:	1	Single Family
	2	Residential 2-4 Unit
	3	Multi-family
Non-residential:	4	Industrial
	5	Warehouse
	6	Retail
	7	Office
	8	Special Purpose

2.1 GIS Processing

- Each of the overlays needed are available for 1994 and can be used to represent 1980. If the 1980 UGB was different than the boundary in 1994, then a 1980 version should be substituted.
- The spatial extent of the resulting coverage should be clipped to the boundary of the study area (1994 LCOG traffic zones).

2.2 Edit Checking

The following data in the parcel file should be checked, through automated edit-checks done using database or statistical software as follows:

(1) Once the business establishment file is linked to parcels by a maplot number, a series of tests focusing on the business-related parcels should be done:

- a) Employment > 0 and No Square Feet
- b) Square Feet > 0 and no Employment
- c) Emp > 0 and (Sqft/emp < 100 or > 1500)
- d) Emp > 4 and StatClass Unassigned or Residential
- e) Improvement Value > 0 and No Sqft
- f) Square Feet > 0 and No Improvement Value
- g) Improvement Value/Sqft < 10 or > 250
- h) Sqft > 0 and Year Built < 1900 or Missing
- i) Density (FAR) < .1 or > 1

Of these tests, the most significant are the first two. Also, the problem parcels can be sorted in order by size (sqft for test a and employment for test b) to isolate the most significant cases and attempt to resolve their problems first.

(2) For parcels with residential land uses (ALU 1-3), do the following tests:

- a) Units > 0 and No Improvement Value
- b) Land value extreme (< \$1000 /acre or > \$250,000/acre)
- c) Improvement value extreme (0 or > \$500,000)
- d) Density extreme (< .1 units/acre or > 50 units/acre)

(3) Additional aggregate tests should be done at the zone level to capture outlier values:

- a) Average land value per acre by land use type (ALU: Industrial, Warehouse, etc) by zone: check for extreme or unusual values
- b) Check aggregate quantity of square footage of each ALU for reasonableness: using average sqft per employee values, what total employment estimate would be produced? How does this compare to observed employment in study area?
- c) Compare total housing units from parcel file against census or LCOG estimates for study area, by housing type.
- d) Check average improvement values by ALU (per unit for residential and per sqft for nonresidential), for extreme or unusual values at the zonal level.

Once these tests are done, and depending on the results, a strategy for data editing should be developed and implemented that involves elimination of problematic records that are inconsequential in size, and synthesis of missing or problematic data values for significant parcels. It would be impractical to more fully specify this strategy until these results are available. Time and resource constraints will clearly influence the strategy. The sooner we can get feedback on this, the better. The strategy that is developed is to be documented in Technical Memorandum 3E.4.

3 1980 Business Establishment Database

The 1980 business establishment database has been prepared by LCOG, and is geocoded, including a match to parcels using a Master Address File. Reformatting of this data will be necessary to produce a business establishment file *Business80.dbf*, in the following format:

Name	Format	Description
BusId	Char	Unique Business ID
Busname	Char	Business Name
County	Integer	County Code
City	Integer	City Code
Stnumber	Integer	Street Number
Stprefix	Character	Street Prefix
Stname	Character	Street Name
Stsuffix	Character	Street Suffix
SIC	Integer	Standard Industrial Class
Emp	Integer	Employment at Site
Zone	Integer	Traffic Analysis Zone
ALU	Integer	Actual Land Use
ParcelID	Character	Unique Parcel ID
Sector	Character	Sector Name

Sectors are aggregations of the SIC values into large industry groupings. The sectors used in 1994 should be matched in the 1980 file:

Name	SIC (2-digit)
Basic	01-51
Retail	52-59
Service	60-81, 83-89
Goved	83, 90-99

3.1 GIS Processing

The geocoded business file should be converted to an Arcview shapefile, as points.

3.2 Edit Checking

The edit checks described for parcels also deal with employment, so little needs to be added here. One additional check would be useful:

- ◆ Business establishments with missing employment, or with very large employment (> 1,000). No employment businesses should be deleted. Large businesses (> 250 jobs or so) should be scanned for reasonableness. One potential problem to look for in large businesses is potentially unallocated firm-level employment that should be distributed to branch offices. This processing was done by LCOG at the time of the original data preparation, but a quick scan of large businesses would be a useful cross-check.

4 1980 Household Data from the Census

Household data is prepared for the base year using a variation of the Synthetic Baseline Population methodology developed by Dick Beckman at Los Alamos. The data processing needed to prepare the data for loading into this procedure is described below. Two 1980 census data sources are used, and both are currently available: Summary Tape File 3A (STF3A) and Public Use Microdata Sample (PUMS) 5%.

Iterative Proportional Fitting (IPF) is used to take Census PUMS data, which is the *joint distribution* of household descriptions that contain limited geographic information, and STF3A, which is the *marginal distribution* of socioeconomic tabulations on a geographic basis, and combine them to create synthetic households distributed by zone.

Households were classified from the 1990 census by income, age of head, household size, and presence of children, and these are the same characteristics that will need to be used in 1980.

4.1 1980 Census STF3A

The summary tables from STF3A needed for 1980 are the following:

Household Characteristic	STF3A Table Number
--------------------------	--------------------

Household income	68
Household size (# persons)	18
Children present	20
Age of head	88
Moved in last 5 years	110
Housing type (# units in struct)	104

The level of geography is specified by the SUMLEV variable, and only the Block Group level should be selected, for Lane County. The County, Census Tract and Block Group should be retained for each record in Lane County along with the tables identified.

4.2 1980 PUMS

The PUMS data are structured as a series of household record collections, with one record containing household-level data followed by as many person records as are indicated for the household, with further data on individuals in the household. All but one variable needed for the preparation of the household database are on the household-level record. Only the age of the head of the household is not available on the household record, and must be retrieved from the person record for the household head.

The variables needed from the 1980 PUMS for the tabulation are:

Record	Variable
Household	Number of Persons
Household	Units in Structure
Household	Vacancy Status (keep only occupied units)
Household	Year Householder Moved
Household	Household Income
Household	Number of Children
Household	Household Weight
Person	Relationship (keep only household head)
Person	Age (of head)

4.3 GIS Processing

A GIS coverage of 1980 census block groups will be needed in order to cross-classify parcels by zone and block group for allocating households to Traffic Analysis Zones.

4.4 Processing Notes

- ◆ The age of head information should be merged on to the household-level record, so that only one record per household is generated in the resulting file.
- ◆ The classification boundaries will not exactly match between 1980 and 1990, particularly on income. The classifications available in the census should be aggregated to match as closely as possible the class boundaries used in 1990. The household characteristics used to classify 1990 households, and the values used to create classification levels on each of these characteristics are shown below. Note that the household typology results from the combination of each of these four characteristics, so for example, the first household type would be income group 1 (under \$10,000), age of head 1 (under 29), household size 1 (1 person), and with no children.

Income	Age of Head	Household Size	Children
Under \$10,000	Under 29	1	0
\$10,000-24,999	20-49	2	1 or more
\$25,000-49,999	50-64	3	
\$50,000 or more	65 or Over	4 or more	

- ◆ In addition to the household characteristics, we need to retain information about the type of housing occupied by households:

Housing Type

Single-family (1-detached or attached, or mobile home)

Residential 2-4 Unit (2-4 units in structure)

Multi-family (5 or more units in structure)

- ◆ A tabulation of the selected PUMS data using Income Group, Age of Head, Household Size, Presence of Children, and Mobility (moved in last 5 yrs) should be prepared, weighted by the Sample Weight. This will produce an estimate of the number of households in Lane County with each unique combination of these five characteristics. This data should be formatted as file *PUMS80.dbf* with one record per household in the Eugene-Springfield MSA, and the following fields:

Income Group
Age Group of Head
Household Size
Children Present
Housing Type
Mobility Status
Household Sample Weight

- ♦ The STF3 tables should be aggregated into the same class boundaries as the PUMS variables (e.g. Children=0 or 1). These data may be formatted as file STF3.dbf with one record per block group in the Eugene-Springfield MSA, and the following fields:

TractBlockGroup
Households by Income
Households by Age
Households by Size
Households by Presence of Children
Households by Housing Type
Households by Mobility Status (moved last 5 years)

5 Urban Growth Boundary

The 1980 Urban Growth Boundary, if different than the 1994 UGB, will be needed for integration with the parcel database as described in section 2. Our expectation is that this boundary has not changed since 1980, and the 1994 boundary may be used in the GIS overlay operations. This needs to be verified.

6 Land Use Plan

A metropolitan land use plan is needed for 1980, and the land use plan designations for each parcel should be reflected in the parcel database. We expect that the land use plan designations are already in the parcel database, but this needs to be verified.

7 Development Cost Estimates

Average construction costs are needed per unit of housing by type (single-family, residential 2-4 unit, and multi-family), and per square foot of non-residential floorspace by type (industrial, warehouse, retail, office, and special purpose). Also needed are demolition cost estimates by type (per unit and sqft for residential and non-residential). Finally, 'soft' development costs attributable to local policies, such as development impact fees, may be specified in a similar way. Simple estimates of these three costs per ALU are needed. If the soft costs are estimated to vary significantly by zone, then a zonal additive adjustment may be provided (the soft cost may be negative to reflect a subsidized development).

8 Travel Model Output

A 1980 travel network and functioning travel model is needed to produce base year travel characteristics used by the model. In particular, the logsum values from the mode choice model are needed, as is an auto am peak highway time. These should be based on the current (1994) zone structure. The matrix names to use would be mf1 for the mode choice logsum for the one car auto ownership group, and mf2 for the am peak highway travel time. The file format will be a binary emme2ban.

LCOG has already prepared these data for 1980. We do not anticipate that further processing of these data are necessary.