

8-23-1990

GIS Address-Matching and Transportation Analysis

James D. Orrell
Portland State University

Follow this and additional works at: https://pdxscholar.library.pdx.edu/open_access_etds



Part of the [Geographic Information Sciences Commons](#), and the [Transportation Engineering Commons](#)

Let us know how access to this document benefits you.

Recommended Citation

Orrell, James D., "GIS Address-Matching and Transportation Analysis" (1990). *Dissertations and Theses*. Paper 4133.

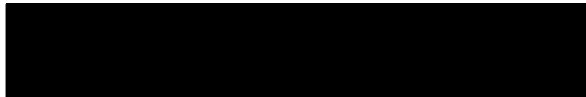
<https://doi.org/10.15760/etd.6016>

This Thesis is brought to you for free and open access. It has been accepted for inclusion in Dissertations and Theses by an authorized administrator of PDXScholar. Please contact us if we can make this document more accessible: pdxscholar@pdx.edu.

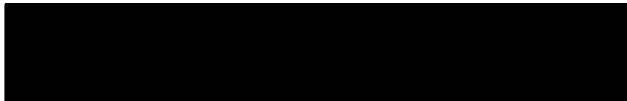
AN ABSTRACT OF THE THESIS OF James D. Orrell for the Master of Science in Geography presented August 23, 1990.

Title: GIS Address-Matching and Transportation Analysis.

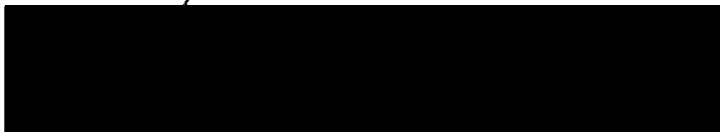
APPROVED BY THE MEMBERS OF THE THESIS COMMITTEE:



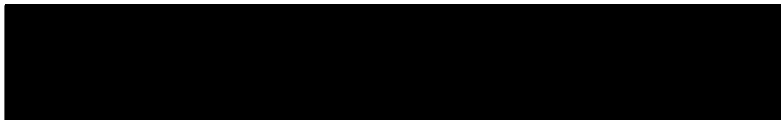
D. Richard Lycan, Chair



Joseph Poracsky



Thomas M. Poulsen



Kenneth J. Dueker

Geographic Information System (GIS) address-matching combined with other GIS processing offers new analytical opportunities in the area of transportation planning and analysis. Address-matching, an automated method for generating geographically-referenced (geocoded) point locations on a map from common tabular databases, can facilitate transportation analysis by

providing a planning tool based on individual rather than aggregated spatial distributions more common to transportation issues.

This thesis presents a case study in which GIS address-matching was applied to a transportation problem at Oregon Health Sciences University (OHSU) in Portland, Oregon. The transportation problem at OHSU is common to major employment centers in urban areas: too few parking spaces for the number of employees and patrons who work and seek services at the location.

Transportation planners at Tri-Met, Portland's transit authority were interested in the spatial distribution of bus services in relation to the residences of OHSU employees. Address-matching and other GIS processing provided detailed information about employee locations within varying buffer zones around the bus routes.

The research problem of this thesis was whether or not, in the context of the OHSU problem, the detailed information provided additional insight and analytical possibilities in comparison to the use of aggregated data. Two approaches were used to evaluate this question. The first approach compared estimates of employees within bus-route buffer zones derived from address-matching with estimates derived from employee density data aggregated by zip code areas.

The results from this comparison revealed limited statistical difference between the two methods. The estimates of the number of employees within the route buffer zones were statistically different at a one-quarter mile buffer

distance for a large study area but were not statistically different at a 500 ft buffer distance nor for a small-area analysis around a single route.

The second approach used in the critique of the utility of address-matching was based on testing the employee point distribution for spatial patterns relative to the bus-route network. This was undertaken in an attempt to demonstrate the additional modeling capabilities available with point data. Descriptive statistics were calculated, and hypotheses using multiple linear regression and analysis of variance were tested, in an attempt to reveal groupings of employees in relation to the route network.

The results from this part of the research were inconclusive in demonstrating any patterns in employee distribution relative to the network. However, data quality problems in the OHSU employee database prevented firm conclusions from being made about the overall utility of pursuing an address-matching approach for this type of problem.

In the summary section of this thesis further research dealing with the type of transportation problem characterized by the OHSU situation is recommended. Several enhancements and GIS alternatives to the approaches used in this study are recommended to further evaluate the utility of address-matching for transportation analysis.

GIS ADDRESS-MATCHING AND TRANSPORTATION ANALYSIS

by

JAMES D. ORRELL

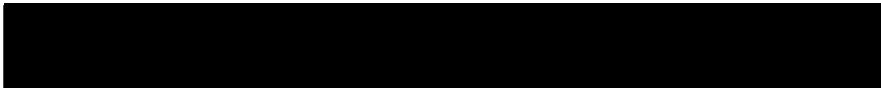
**A thesis submitted in partial fulfillment of the
requirements for the degree of**

**MASTER OF SCIENCE
in
GEOGRAPHY**

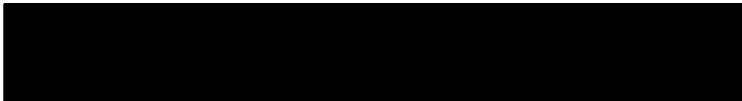
**Portland State University
1990**

TO THE OFFICE OF GRADUATE STUDIES:

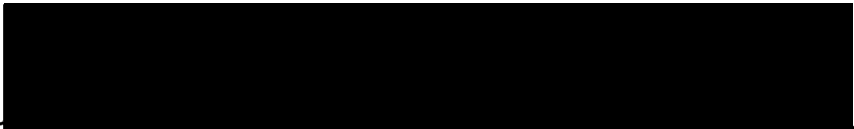
The members of the Committee approve the thesis of James D. Orrell presented August 23, 1990.



D. Richard Lycan, Chair



Joseph Poracsky

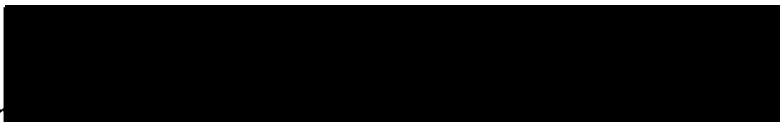


Thomas M. Poulsen

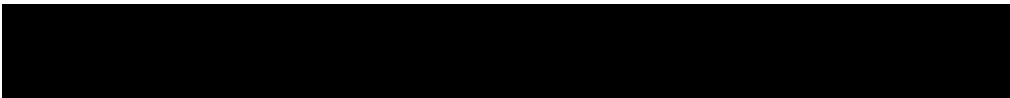


Kenneth J. Dueker

APPROVED:



Thomas M. Poulsen, Chair, Department of Geography



C. William Savery, Vice Provost For Graduate Studies and Research

ACKNOWLEDGEMENTS

I would like to thank the members of the committee for the extra effort made on my behalf in facilitating the completion of this thesis during the summer term. I would also like to thank Carolyn Perry, the department secretary, for her assistance in coordinating this and many other "details" of my graduate program.

Special thanks are extended to Bill Rabiega for his advice on statistical methods and to Ric Vrana for his many helpful comments and conversations during the long days of August.

Finally, I would like to thank Dick Lycan and Ken Dueker for providing research opportunities and intellectual guidance at the beginning of my career as a geographer.

TABLE OF CONTENTS

	PAGE
ACKNOWLEDGEMENTS	iii
LIST OF TABLES	vi
LIST OF FIGURES	ix
 CHAPTER	
I INTRODUCTION	1
Background	1
GIS Concepts	
GIS Address-Matching	
Related Literature	6
Transportation Analysis and Planning	8
Thesis Organization	9
II DESCRIPTION OF THE PROBLEM	11
Background: OHSU Transportation Problem	11
Research Framework	17
III POTENTIAL TRANSIT PATRONAGE: POINT AND DENSITY MODELS	18
Study Area	18

Methods 20

Point-Based Estimates Of Potential Bus Patronage
Density-Based Estimates Of Potential Patrongage
Chi Square Statistic

Analysis And Results 26

Original Study Area
Route 8 Study Area
Observations: Chi Square Comparisons

IV POINT DATA EXPLORATORY STATISTICS AND
TRANSPORTATION ANALYSIS 45

Data 45

Analysis And Results 46

Descriptive Statistics
Multiple Linear Regression
Two-way Analysis of Variance (ANOVA)
Observations: Exploratory Point-Based Statistics

V SUMMARY 65

Conclusions 67

Further Research 70

Bus Patronage Estimating Models

REFERENCES CONSULTED 74

LIST OF TABLES

TABLE		PAGE
I	Chi Square Contingency Table and Statistic Point and Density Employee Estimates 500 Ft Buffer Zone: All Employees Original Study Area	31
II	Chi Square Contingency Table and Statistic Point and Density Employee Estimates 1/4-Mile Buffer Zone: All Employees Original Study Area	32
III	Chi Square Contingency Table and Statistic Point and Density Employee Estimates 500 Ft Buffer Zone: Employees With Low Parking Seniority Original Study Area	33
IV	Chi Square Contingency Table and Statistic Point and Density Employee Estimates 1/4-Mile Buffer Zone: Employees With Low Parking Seniority Original Study Area	34
V	Chi Square Contingency Table and Statistic Point and Density Employee Estimates 500 Ft Buffer Zone: All Employees Route 8 Study Area	38

VI	Chi Square Contingency Table and Statistic Point and Density Employee Estimates 1/4-Mile Buffer Zone: All Employees Route 8 Study Area	39
VII	Chi Square Contingency Table and Statistic Point and Density Employee Estimates 500 Ft Buffer Zone: Employees With Low Parking Seniority Route 8 Study Area	40
VIII	Chi Square Contingency Table and Statistic Point and Density Employee Estimates 1/4-Mile Buffer Zone: Employees With Low Parking Seniority Route 8 Study Area	41
IX	Summary of Chi Square Comparisons	43
X	Mean Distance of Employees To Nearest Bus Route Original Study Area Employees: Complete Data Cases Only	48
XI	Mean Age, Seniority, Parking Seniority Original Study Area Employees: Complete Data Cases Only	49
XII	Multiple Linear Regression: Employee Parking Seniority Against All Other Variables Original Study Area Employees: Complete Data Cases Only	54

XIII	Multiple Linear Regression: Employee Distance From Nearest Bus Route Against All Other Variables Original Study Area Employees: Complete Data Cases Only	57
XIV	Two-Way Analysis of Variance: Parking Seniority And Across-Route And Along Route Buffer Segements . .	61

LIST OF FIGURES

FIGURE		PAGE
1.	A Generalized Diagram of the Address-Matching Process . . .	4
2.	OHSU facilities on Marquam Hill As Viewed From Southeast Portland	13
3.	Address-Matched Locations of OHSU Employees Relative to 500 Ft Bus-Route Buffer Zones For Part of the Portland Metropolitan Area	15
4.	OHSU Employees Within A Quarter-Mile Buffer Zone Along Four Major Commuter Corridors	16
5.	Study Area Bus Routes and Zip-Code Boundaries	19
6.	Route 8 Study Area Quarter-Mile Bus-Route Buffer Buffer Polygon Overlaid on Zip-Code Polygons . . .	22
7.	An Example of the Calculations involved in Estimating the Number of Employees Within a Bus-Route Buffer Based on a Uniform Density Approach	22
8.	General Model of a Two-Sample, Two-Category Chi Square Contingency Table and Test Statistic Formula	23
9.	Original Study Area; All Employees Within 500 Ft and 1/4-Mile Bus-Route Buffer Zones	29

10.	Original Study Area; Employees With Low Parking Seniority (Less Than One Year) Within 500 Ft and 1/4-Mile Bus-Route Buffer Zones	30
11.	Route 8 Study Area; All Employees Within 500 Ft and 1/4-Mile Bus-Route Buffer Zones	36
12.	Route 8 Study Area; Employees With Low Parking Seniority Within 500 Ft and 1/4-Mile Bus-Route Buffer Zones .	37
13.	Mean Parking Seniority Associated With Buffer Segments Created For Route 8	62

CHAPTER I

INTRODUCTION

Geographic Information System (GIS) address-matching as applied to transportation planning and analysis is explored in this thesis. Address-matching is an automated method for generating geographically-referenced (geocoded) point locations on a map from data sources containing address descriptions. This capability, combined with other GIS functions, provides opportunities for transportation analysis based on individual rather than aggregated spatial distributions.

This introduction provides background to GIS, and address-matching concepts, and introduces the transportation planning problem from which the use of address-matched data is evaluated. A review of related literature and overall thesis organization are also presented.

BACKGROUND

GIS Concepts

GIS is most readily described as a set of software tools which combine computerized mapping with traditional database software. GIS systems are designed around specialized computer data structures and analytical routines

which facilitate the display, manipulation and storage of maps and associated tabular information (Burrough, 1986).

Primary among the functions common to GIS systems are routines which allow for complex map overlays. GIS overlay procedures allow for creation of composite maps, with redefined spatial relationships, resulting from the geometric intersections of two or more original maps. This redefinition of spatial relationships makes possible queries of the composite map, and composite map attributes, not easily undertaken without a GIS.

Basic map objects representing points, lines, and areas (polygons) provide the framework for three common types of overlay analysis in GIS systems: 1) point-in-polygon, 2) line-in-polygon, and 3) polygon-in-polygon. In this thesis, point-in-polygon and polygon-in-polygon overlay processing were applied extensively to identify and tabulate the number of individuals falling within varying buffer zones around public bus routes.

GIS Address-Matching

One of the additional capabilities of some GIS systems is the ability to translate a text description of an address ("85 NW 10th") into an approximate geographic location along a street segment on a map. This process generally referred to as automated address-matching involves two address-based databases (Figure 1). One is a tabular database, usually maintained for traditional data management purposes, containing addresses and other

information related to a specific subject. In this thesis, the tabular database is an employee database.

The second address database represents street geography. It is displayed on the computer screen as a street map, but it is stored in a separate computer file which includes block-number range and street-name descriptions ("0-100","NW 10th"; "1-99","NW 10th") for both sides of every street segment.

The address-matching process matches subject data to corresponding street segments, and the proper side of the street, by finding the street segment with the same street name and inclusive (odd or even) block-number range. When a matching segment is located, the approximate x,y point location for the address is geocoded through an interpolation routine. The interpolation process places the point by calculating a proportional distance along the street segment based on the address number and associated block-range numbers.

The process proceeds for each address record in the tabular database, ultimately updating the attributes of the original database with x,y point locations for each of the subject records successfully matched. After the addresses are geocoded with x,y locations on the map, other information in the address database (e.g., type of employee, sex, age, etc.) can be mapped and manipulated by the GIS .

TABULAR DATA

<u>Address</u>	<u>Name</u>	<u>Age</u>
85 NW 10TH	SMITH	43

(A)

STREET GEOGRAPHY

<u>Block-Range</u>	<u>Street</u>
1-99	NW 10TH
2-100	NW 10TH

(B)

POINT INTERPOLATION

2		100 (Even Numbers)
	<u>NW 10TH</u>	
1	85	99 (Odd Numbers)

(C)

UPDATED TABULAR DATABASE

<u>Address</u>	<u>Name</u>	<u>Age</u>	<u>X,Y Map Coordinates</u>
85 NW 10TH	SMITH	43	1248869,7000645

(D)

Figure 1. A generalized diagram of the address-matching process. Address-based data from a tabular database (A) is matched to a street-segment in the database storing information about the digital street network (B). A point location is interpolated along the street segment based on the inclusive (odd or even) block-range numbers and the subject address number (C). Finally, the interpolated point-location coordinates are used to update the attributes of the original address database (D).

Records which cannot be matched are saved to a third database. Subsequent evaluation of the unmatched records allows for deciphering and correcting matching problems as necessary and appropriate to the analysis. Data quality problems (miscoded addresses) in either of the files will result in matching failure as will missing street segments in the street database.

The U.S. Bureau of the Census has developed a nationwide digital street network known as the TIGER file (Topologically Integrated Geographic Encoding and Referencing) which provides a convenient base from which to pursue address-matching. This digital representation of street geography created for use in preparing and tabulating the 1990 decennial census includes all of the street-segment attributes necessary for address-matching. It also contains information about the way in which the street segments are linked (topology) that is important for other network analysis functions available in GIS systems (Marx, 1990).

GIS address-matching can be pursued with any tabular database which contains basic address information. The capability to readily generate point information via address-matching allows for analysis of a wide range of questions dealing with locations of point features across geographic space. As discussed by Marx (1990) and Schwartz (1989), it is expected that the interface provided by TIGER-facilitated geocoding to diverse address-based data will greatly increase the use of GIS address-matching and other spatial analysis based on TIGER products.

RELATED LITERATURE

Discussion of address-matching applications in GIS literature has focused on the usefulness of the technique as a point location mechanism for pin mapping (e.g. crime-locations), as a precursor to optimal path-finding between two or more geocoded points (e.g. dispatching) and for allocating resources (e.g. students) to one or more central locations such as in school planning issues (Lupien et al., 1987). Literature discussing the general utility of address-matching as a precursor for analysis of geographic distributions in the transportation setting is not as readily identified.

Traditional methods of transportation analysis employ data aggregated by traffic analysis zones, census tracts or other areal units (Hanson, 1986). Nyerges and Dueker (1988;10) discuss the use of address-matching of building permit, vital statistics, and employment information to update traffic zone data for transportation planning applications. The direct use of such geocoded point data prior to aggregation by traffic zones is not discussed.

Hunt et. al (1986) describe an approach to bus-route demand modeling using route buffer zones combined with census-tract level demographic information. Overlay of route buffers on census tracts allowed for estimates of potential riders falling within a quarter-mile buffer of a given route. A major area identified for future work in this study is evaluation of model sensitivity to the assumption of homogeneous population distributions across areal units. This question was considered in Chapter III where the assumption of

homogeneous population distribution across areal units is compared to point data.

Sosslau and McDonnell (1984) describe bus-routing analysis techniques using transportation-related census data. Journey-to-work information for Central Business District (CBD) Zones and census block groups are suggested as generally appropriate aggregation levels for routing questions.

Other researchers have recommended that transportation analysis based on disaggregated data is desirable and needs further investigation to advance theory and understanding about transportation processes (Hanson, 1981; Hecht, 1974). Hecht analyzed 440 residential locations for employees of 10 industrial firms in the Worcester, Massachusetts CBD seeking a relationship between straight-line distance from the work place and several socioeconomic variables. The methodology of this study, and the problem in general, is mirrored in parts of this thesis.

The analysis of point-patterns, an area of interest to geographers and other spatial analysts, is related to this thesis by the point data resulting from address-matching. However, the literature in this area does not bring much light to the use, or limitations, of address-matching as a method for generating point-data sets. Most empirical studies in this area generally describe the use of manual geocoding of point locations from a base map prior to analytical processing (Barff and Hewitt, 1989; Haining, 1982; Getis, 1984).

Other GIS and professional geography literature have raised questions about the utility, and limitations, of GIS applications and methodology (Muehrcke, 1990). In a review of the status of GIS, Waters (1989) discusses GIS in terms of why it is used and what use is made of its products. He compares the current interest in GIS with a critical examination of the development and wide-scale application of factor-analysis techniques to geographic research during the 1970's (Williams, 1971) . Waters concludes that GIS applications can be justified by their ability to manage and store information but that synthesis and extended analysis, among other areas in the field of GIS, need to be further demonstrated if GIS is going to significantly aid in the advancement of geographic understanding.

Waters acknowledges that analytical capabilities of GIS are starting to emerge with much research in the GIS community currently directed at this issue. This thesis is an example of research seeking to contribute to the use of GIS for extended modeling and analysis.

TRANSPORTATION ANALYSIS AND PLANNING

In Chapter II a detailed description of a transportation planning problem which made use of GIS address-matching is described. In general, the problem involved analyzing the spatial distribution of commuters to Oregon Health Sciences University (OHSU) in relation to the public bus route network in Portland, Oregon.

Address-matching, combined with other GIS processing, allowed transit planners to evaluate the number of potential riders within varying buffer zones around existing and alternative bus routes. The application of address-matching to the problem, provided detailed spatial information about employee locations which was useful to transit planners.

The research focus of this thesis is on the degree to which this detail, which comes at the expense of additional computer processing, provided greater insight and analytical possibilities than would have been possible with aggregated data. The research hypothesis was that the disaggregated point data provides useful and statistically significant information for transportation planning not available when using aggregated data.

THESIS ORGANIZATION

The OHSU transportation problem described in Chapter II provides the departure point for this thesis. The research framework, as applied in the analyses of Chapters III and IV, is also presented in Chapter II.

Chapter III compares point and density models for analyzing potential transit patronage as the first critique of address-matching. In Chapter IV, statistical tests based on the point data are used to further evaluate the OHSU problem and to demonstrate the applicability of point data in exploring hypotheses not easily pursued with aggregated data.

Chapter V summarizes the research and recommends areas that would benefit from further research in the area of GIS address-matching and transportation analysis.

CHAPTER II

DESCRIPTION OF THE PROBLEM

The introduction identified transportation-related applications of point data generated from GIS address-matching as the focus of this thesis. In this chapter the Oregon Health Sciences University (OHSU) transportation problem is described in detail. This background is followed by an introduction to the research framework used in the analyses of Chapters III and IV.

BACKGROUND: OHSU TRANSPORTATION PROBLEM

The point data set used in this thesis was generated from address-matching the locations of OHSU employees, faculty and students for part of the Portland, Oregon Metropolitan area. The purpose of this effort was to evaluate the accessibility of the OHSU community to transit services and to demonstrate how GIS can be applied in the context of a transportation planning problem (Lycan and Orrell, 1989).

OHSU is a major employer in the metropolitan area with over 5000 employees and approximately 2500 students (Hannum, 1989). The transportation and parking problems inherent with a commuter base of this size, are compounded by the site and situation of OHSU. The facility is perched on Marquam Hill at the southwest edge of downtown Portland (Figure 2), and is

surrounded by residential neighborhoods with limited vehicular access by two-lane roads.

Tri-Met, Portland's mass transit agency, was interested in investigating the spatial distribution of OHSU employees in relation to bus routes. Did bus services mesh well with employee locations as a whole and for subgroups deemed more likely to use mass transit? Could direct-service vans along major commuter corridors potentially be justified based on the number of employees located along these corridors?

The evaluation of accessibility to transit services was based on proximity of OHSU employees to bus lines using buffer zones of 500 feet around all of the bus routes covering the study area (Figure 3). The buffer zone distance of 500 feet (two-city blocks) was used as an acceptable walking distance for most potential riders. In the case of the direct service alternatives, a quarter-mile buffer zone was also evaluated (Figure 4).

The results from the analysis showed that approximately 40% of the employees successfully address-matched were within 500 ft of a bus route. Based on these findings Tri-Met planners decided not to alter the regular bus routes in an attempt to provide greater accessibility to OHSU employees.



Figure 2. OHSU facilities on Marquam Hill as viewed from southeast Portland.

Tabulation of employees along the alternative corridors, as interpreted by Tri-Met, indicated potential for direct-service vans to OHSU. Follow up research in the form of an intent-to-use survey for those employees within the alternative buffer zones was recommended by Tri-Met to further evaluate potential for this service. Although this was not pursued by Tri-Met, the address-matching results would have facilitated the survey mailing.

The original OHSU study was limited to cross-tabulating the numbers of employees falling within the buffer zones based on a number of categories associated with the employee database. It did not attempt, beyond cross-tabulation, to statistically evaluate the characteristics of the employees within the buffer zones, nor in general, how the information compared to that which could have been generated without address-matching.

Transit planners at Tri Met had manually attempted to evaluate the locations of OHSU employees, aggregated by zip code areas, prior to application of the GIS address-matching approach. The use of address-matching bypassed an alternative of using the zip-code area totals as the basis for estimating the number of employees within the buffer zones.

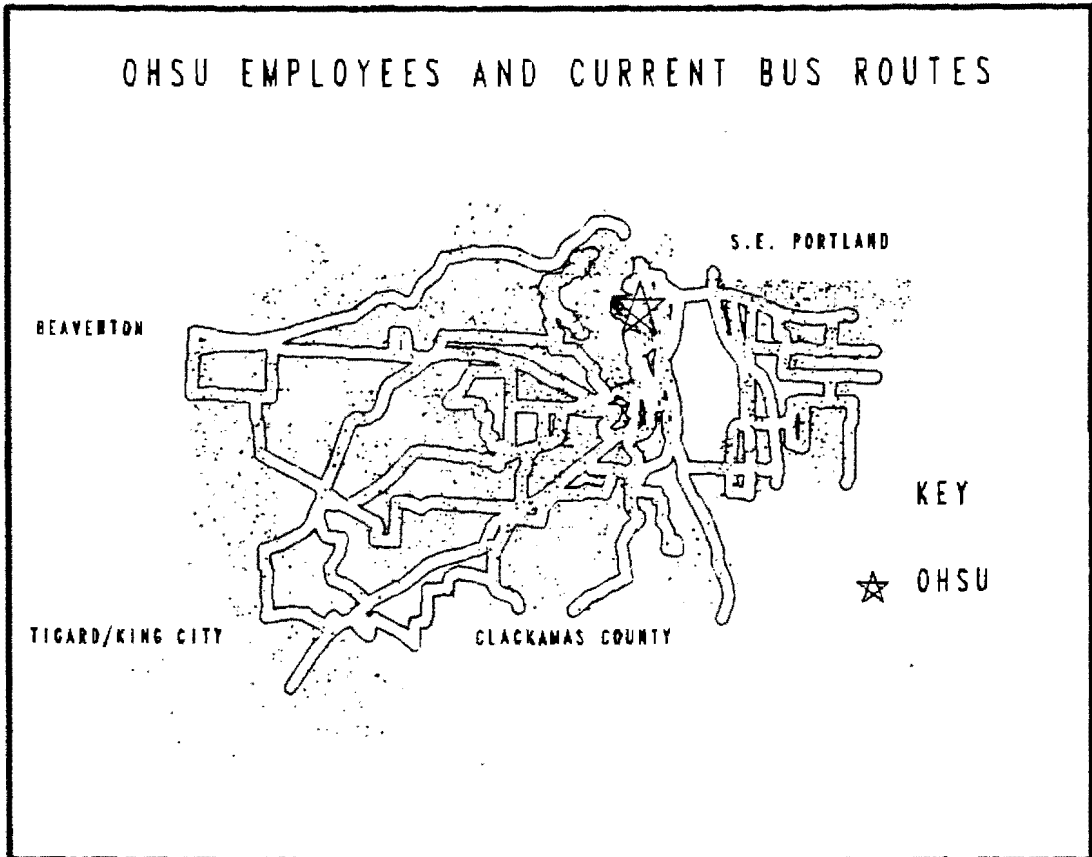


Figure 3. Address-matched locations of OHSU employees relative to 500 ft bus-route buffer zones for part of the Portland Metropolitan Area.

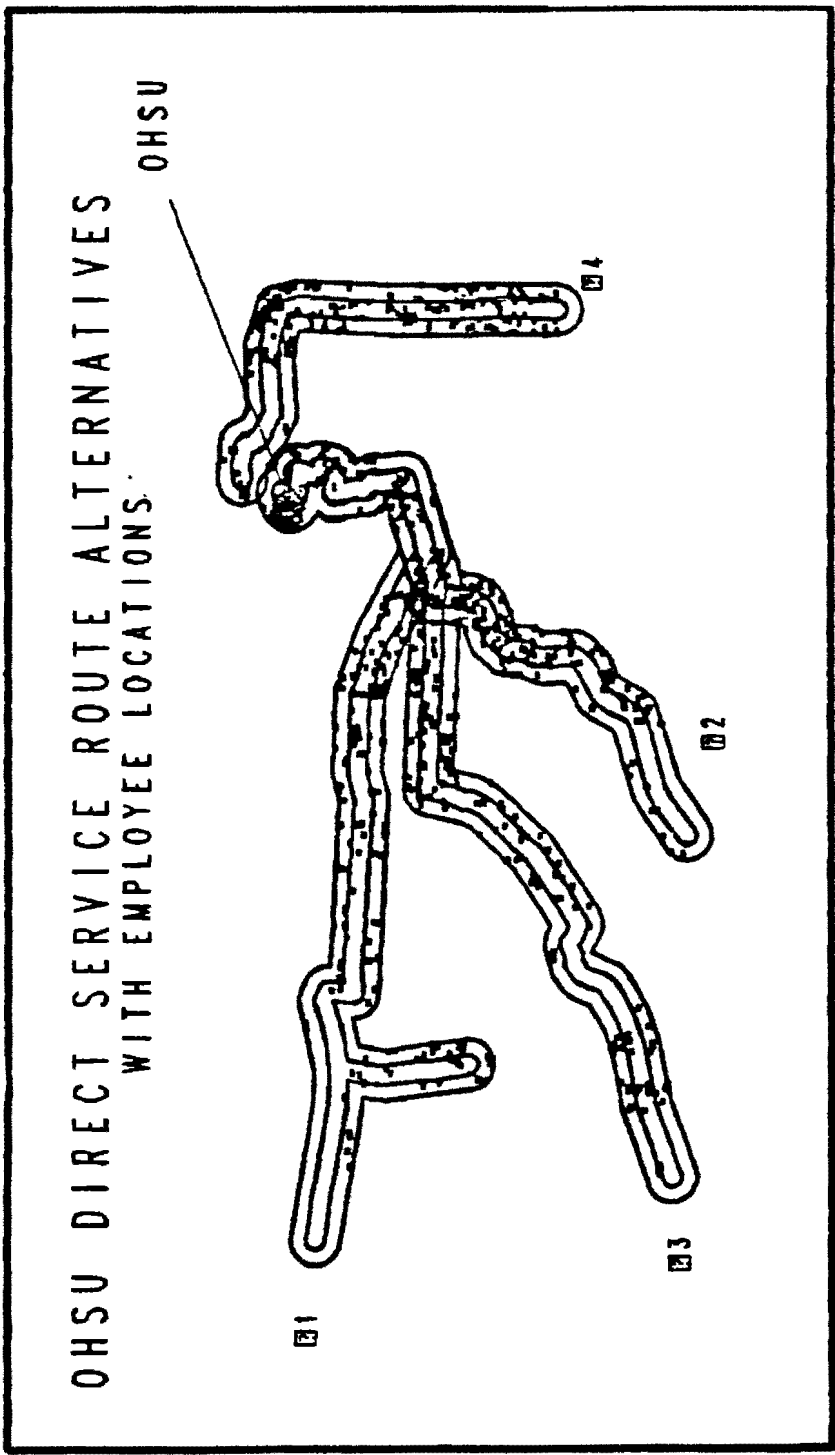


Figure 4. OHSU employees within a quarter-mile buffer zone along four major commuter corridors.

RESEARCH FRAMEWORK

The problem explored in this thesis thus became a question of the benefits and limitations of address-matching. Two approaches were conceived to evaluate this question. The first approach (Chapter III) centered on comparison of buffer-zone frequency counts estimated from the address-matched point data and from employee zip-code-area densities. The use of employee densities by zip code area was a logical alternative to use with the OHSU data in the absence of address-matching software.

The second approach (Chapter IV) sought ways in which the point data could be successfully used with statistical analysis to potentially reveal subtle map patterns not observable with aggregated data. An underlying perception of OHSU planners, prior to address-matching, was that employees were somehow grouped in large clusters of "unmined" transit-patronage potential. As the original buffer analysis did not confirm this, further statistical analysis was necessary to attempt to unlock any hidden patterns in the data.

One of the main advantages of using disaggregated data is the ability to pursue multi-variate hypotheses based on knowledge of individual locations. In an exploratory transportation planning problem such as described here, can advantage be made of the point data to further the planning process?

CHAPTER III

POTENTIAL TRANSIT PATRONAGE: POINT AND DENSITY MODELS

This chapter describes the comparison of point and density models for estimating the number of employees in bus-route buffer zones. The study area, GIS methods, statistical tests, and results for this comparison are all presented in this chapter.

STUDY AREA

The study area described throughout this thesis includes two sections (Figure 5). The first section (Original Study Area) covers the area used with the original OHSU analysis. The second section (Route 8 Study Area) includes part of northeast Portland in proximity to Bus Route 8, the only route which directly serves OHSU from that sector of the city. The addition of the Route 8 Study Area served to extend the analysis beyond the original OHSU study. It also allowed for comparison of the methods at two different scales: one for large area analysis common to preliminary planning efforts characterized by the original study; and one for a small area analysis around a single route.

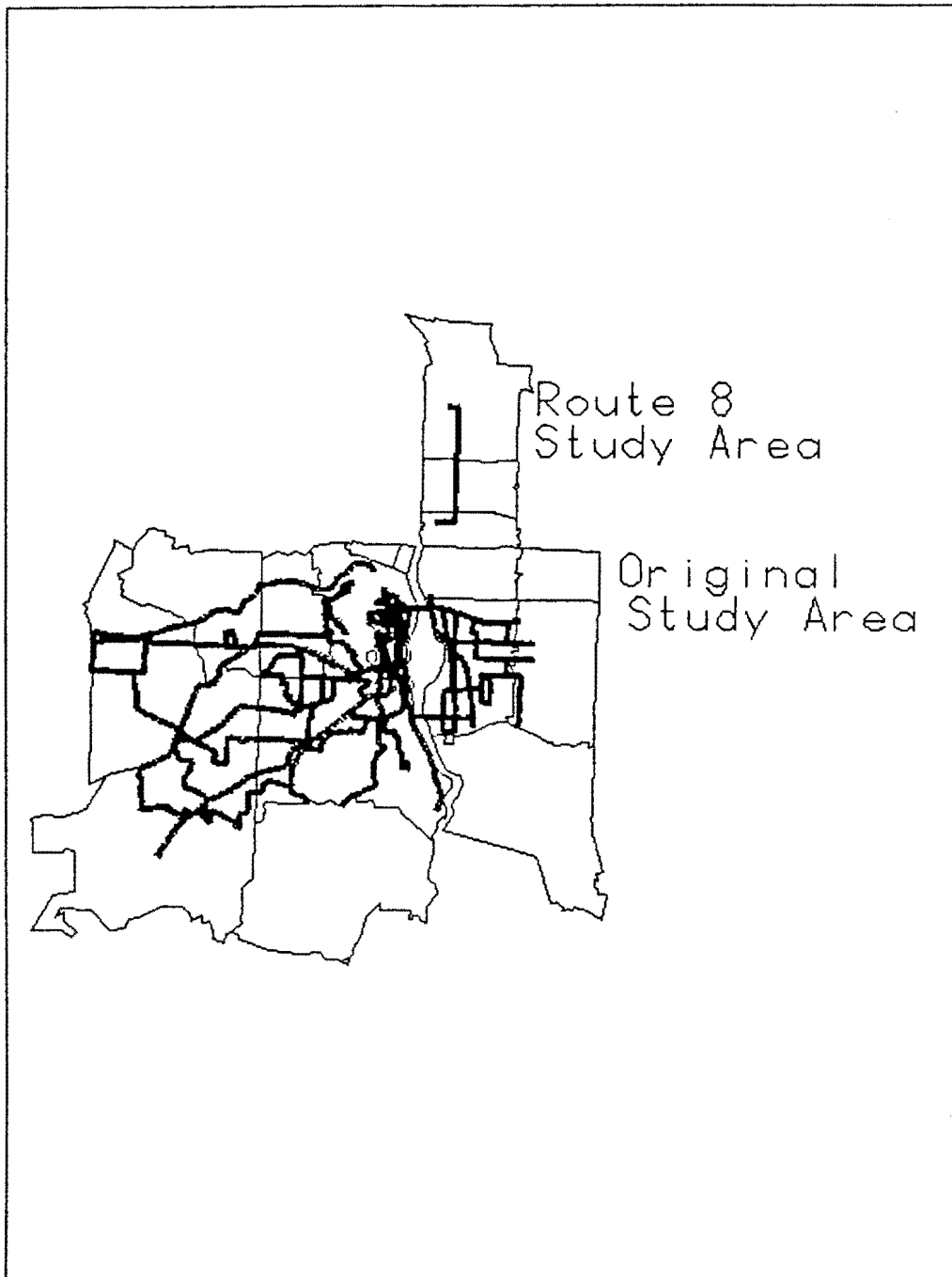


Figure 5. Study area bus routes and zip-code boundaries. The isolated route at the top of the map is the Route 8 Study Area. The remaining route network is the Original Study Area.

METHODS

Point-Based Estimates Of Potential Bus Patronage

Calculation of the total employees within the various bus-route buffer zones using the address-matched point data involved overlay (point-in-polygon) of the buffer zones with the employee point locations. Each overlay resulted in the "tagging" of all points inside and outside the buffer zone with unique codes. Totals of all the points inside a given buffer were thus tabulated based on the coding scheme resulting from the overlay process.

Density Based Estimates of Potential Bus Patronage

The estimates of the number of employees within the buffer zones using a uniform density approach were also based on overlay processing. However, rather than counting points inside the buffer zone, calculation of the areal extent of each zip code within the bus-route buffer was used to prorate an employee estimate based on the density of employees in the zip code area.

Prior to overlay processing, employee densities for each zip code area were calculated based on the assumption of homogenous distributions of employees across the separate zip code areas. This was accomplished by dividing the total number of employees within the zip code by the corresponding zip code area.

The overlay process between the bus-route buffer polygons and zip code polygons (polygon-in-polygon) resulted in new polygons wherever the two

intersected. An example can be seen in Figure 6 which shows a quarter-mile buffer zone around Route 8 intersecting three zip code areas. In this example, the overlay process resulted in the creation of three new polygons which were used to estimate the number of potential patrons in proximity to the route.

Multiplication of each partial zip code area inside of the buffer zone, by its corresponding employee density, provided an estimate of the number of employees from each zip code area inside the buffer zone. Summation of all zip code areas inside the buffer zone provided a total estimate for the buffer. Figure 7 gives an example of the steps involved in deriving these estimates for the Route 8 example seen in Figure 6.

Chi Square Statistic

The chi square test allows for comparison of two-sample data sets which are measured on a nominal basis (Ebdon, 1977). In testing this thesis, chi square was used as the basis for comparing the nominal frequencies of OHSU employees calculated as inside or outside bus-route buffer zones using the two different methods described above.

The general form of the two-sample, two-category chi square contingency table and test-statistic calculation are indicated in Figure 8.

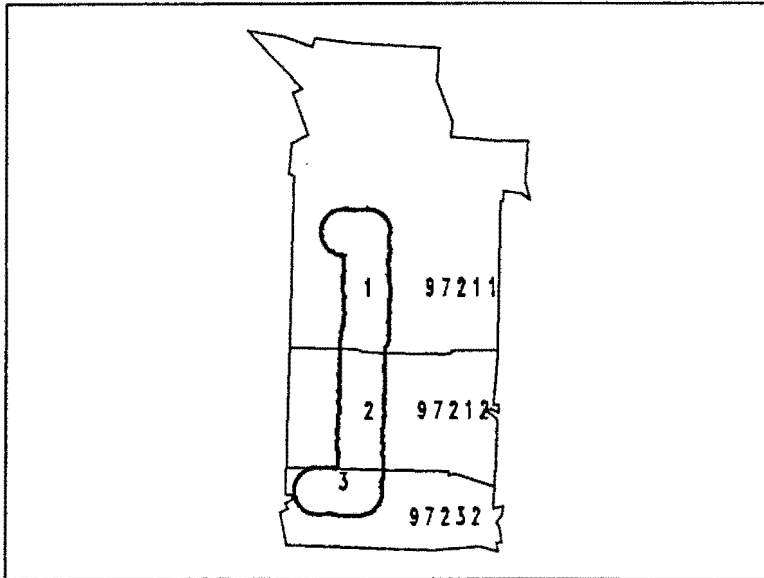


Figure 6. Route 8 study area quarter-mile bus-route buffer polygon overlaid on zip-code polygons. The result of the overlay is creation of three new polygons where the buffer crosses the zip-code boundaries. These polygons numbered 1, 2, and 3 were used to calculate the number of employees within the buffer zone based on the area and employee density of each polygon.

<u>Zip Code</u>	<u>Employee Density (#/Mile²)</u>	<u>Area Inside Buffer (Miles²)</u>	<u>Estimated Employees Inside Buffer</u>
97211	26.5	0.85	23
97212	97.2	0.62	60
97232	58.7	0.43	26
		Total:	<hr/> 109

Figure 7. An example of the calculations involved in estimating the number of employees within a bus-route buffer based on a uniform density approach.

	Inside Buffer	Outside Buffer	Total
Address-Matched Employees	A	B	(A+B)
Zip Code Area Estimate	C	D	(C+D)
	<u>(A+C)</u>	<u>(B+D)</u>	<u>N</u>

Chi Square Statistic:

$$X^2 = \frac{N(|AD - BC| - N/2)^2}{(A+B)(C+D)(A+C)(B+D)}$$

Figure 8. General model of a two-sample, two-category chi square contingency table and test statistic formula. [Source: Ebdon, 1977;63-64]

The chi square statistic (X^2) is used in combination with the overall degrees of freedom for the model to look up the critical value of the statistic at the selected significance level (Ebdon,1977). As stated by Ebdon, the degrees of freedom is one for the above model and is given by multiplying the number of rows in the table minus one, times the number of columns in the table minus one ($[2-1]*[2-1]=1$).

The significance level chosen for use with this test was .01. The minimum chi square statistic for one degree of freedom at the .01 significance level is 6.64. Confirmation of hypotheses at the .01 significance level, by chi square values greater than or equal to 6.64, implies that the probability of chance causing the observed difference between the two measurements is one in one-hundred.

Generally, formal statistical testing employs two statistical hypotheses: a null hypothesis stating that the expected relationship between two samples does not exist; and an alternative hypothesis based on rejection of the null and confirmation of the expected relationship (Silk, 1979). This convention is followed in this thesis.

The chi square test was applied to buffer distances of 500 feet and one-quarter mile for the Original Study Area and for the Route 8 Study Area in northeast Portland. For these buffer distances statistical comparisons were made for the employee sample as a whole and for only those employees with

the least parking seniority.¹ The purpose of this stratification was to evaluate any sensitivity of the two methods to the size of the study area or the size of the population.

A final consideration in using the chi square test in comparing the point and area techniques for estimating potential bus-route patronage was identifying the total number of observations for each approach. This number includes both the cases inside and outside the buffer zone for the study area.

In the case of the address-matched point data this was easily identified by the total number of points address-matched. However, calculating the total number of employees for the density-based methods was more involved. Zip code boundaries do not in all cases neatly coincide with street geography and as a result an additional overlay between the zip code zones and the study area street-map boundary was necessary to insure that the same areal extent for each method was compared.

As with the overlay of the bus-route buffer zones on the zip code areas described above, an estimate of the total number of employees within the zip code area coincident with the address-matched study area was calculated based on the area and employee densities for each zip code zone falling within the street-map boundary. The total number of employees estimated from this

¹Parking seniority at OHSU is indicative of access to a parking permit. Employees with low parking seniority (less than 1 year) were thought to be the most likely users of transit services and were thus targeted in the analysis.

adjustment varied slightly from the total points address-matched as can be seen in the contingency tables which follow.

The discrepancy between the totals was attributed to the address-matching process which matches to street addresses independent of the zip code descriptions. In the address-matching process, records were matched to a street address without regard to the zip code. Thus, records with miscoded zip codes were matched correctly to their actual street location even when the zip code was incorrect for the record. On the other hand, employee zip code totals and densities were based on cross-tabulations of the original database without regard to the street address of each record.

The problem of "zip code" data quality was accepted as part of the data set for the purpose of comparing address-matching methods with the density-based methods. This type of data quality problem would have most likely been unnoticed by an analysis that did not employ address-matching.

ANALYSIS AND RESULTS

The following results document the chi square comparisons of the two subject techniques for estimating numbers of employees inside bus-route buffer zones. Chi square analysis only assesses whether a statistical difference exists between the frequency counts. The meaning of any statistical difference must be interpreted.

For the purposes of this thesis, the interpretation of chi square results assumed that any statistically-significant difference between frequency counts implies a positive finding for the utility of address-matching for this application. Any results not showing statistical significance were interpreted as negating the utility of address-matching methods for the OHSU data set.

The following statistical hypotheses apply for all of the chi square tests.

Null hypothesis: Estimation of total observations inside bus-route buffer zones utilizing address-matched point data does not differ from results obtainable from areal/density estimation procedures. The statistical criteria being a .01 confidence interval, based on one degree of freedom.

Alternative Hypothesis: Address-matched point data do show a significant difference in the number of observations estimated within a bus-route buffer zone as compared to zone-based estimation procedures.

The first results reported are for the Original Study Area and are followed by results for Route 8. One map showing point locations within the two buffer-zone distances tested (500 ft and 1/4-Mile) is presented with each study area scenario (All Employees and Employees With Low Parking Seniority). Because

the zip code estimates involved density calculations done with a tabular database as indicated in Figure 7 similar maps are not shown for each of the density estimates.

Original Study Area

500 ft Buffer Zone: All Employees. The null hypothesis cannot be rejected based on the results seen in Figure 8 and the X^2 value of .18 seen in Table I. The estimate of buffer membership using address-matched point data offers no statistical advantage compared to density estimates for this case.

1/4 Mile Buffer Zone: All Employees. The null hypothesis can be rejected based on the X^2 value of 8.59 seen in Table II. The difference between address-matched estimates of buffer membership and the density estimate of buffer membership is significant at the .01 level.

500 ft Buffer Zone: Low Parking Seniority Employees. The null hypothesis cannot be rejected based on X^2 value of 0.002 in Table III. The estimate of buffer membership using address-matched point data offers no statistical advantage compared to density estimates for this case.

1/4 Mile Buffer Zone: Low Parking Seniority Employees. The null hypothesis can be rejected based on the X^2 value of 20.54 seen in Table IV. The difference between the point estimate of buffer membership and the density estimate of buffer membership is significant at the .01 level.

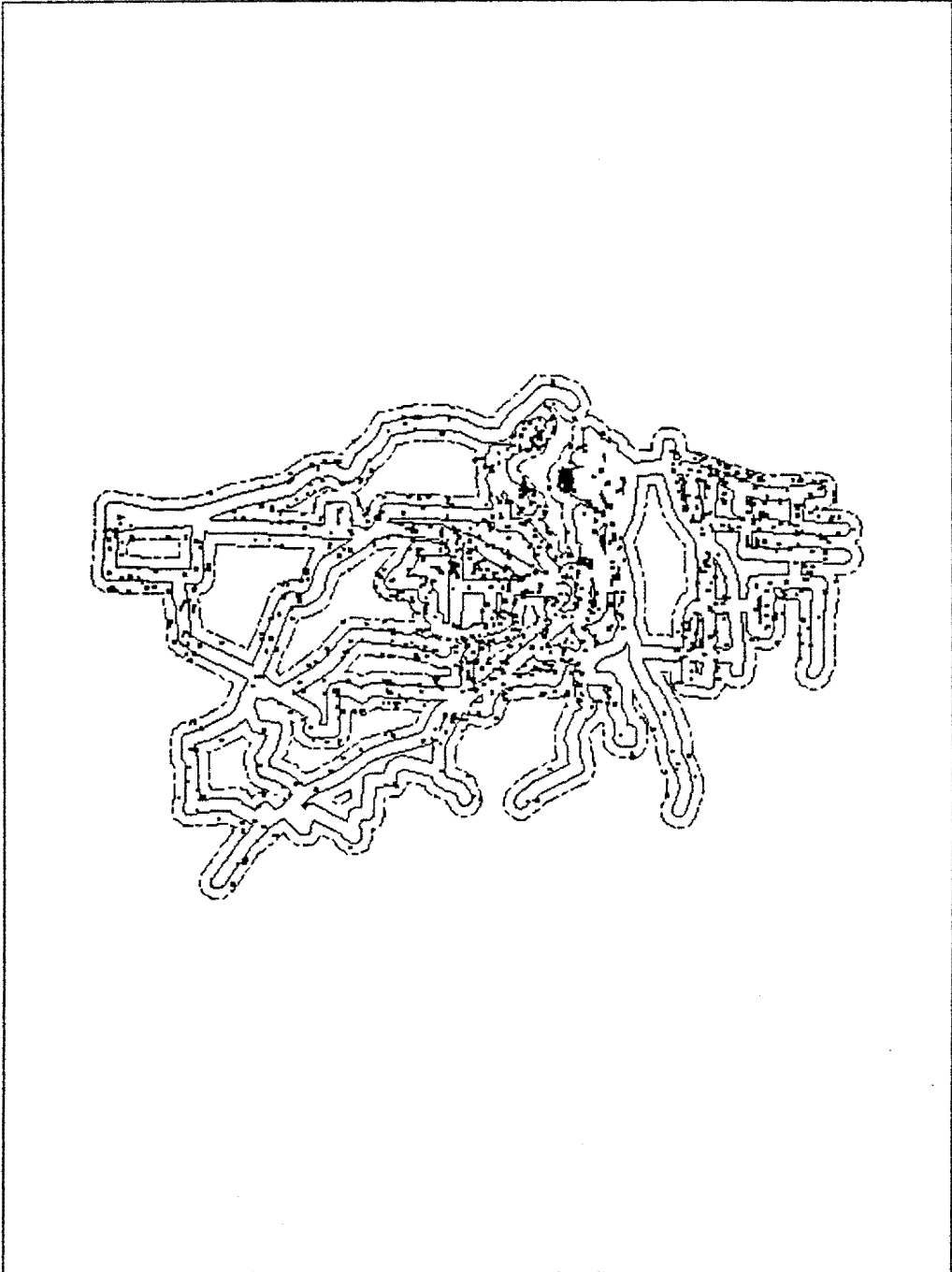


Figure 9. Original study area; all employees within 500 ft and 1/4 mile bus-route buffer zones.

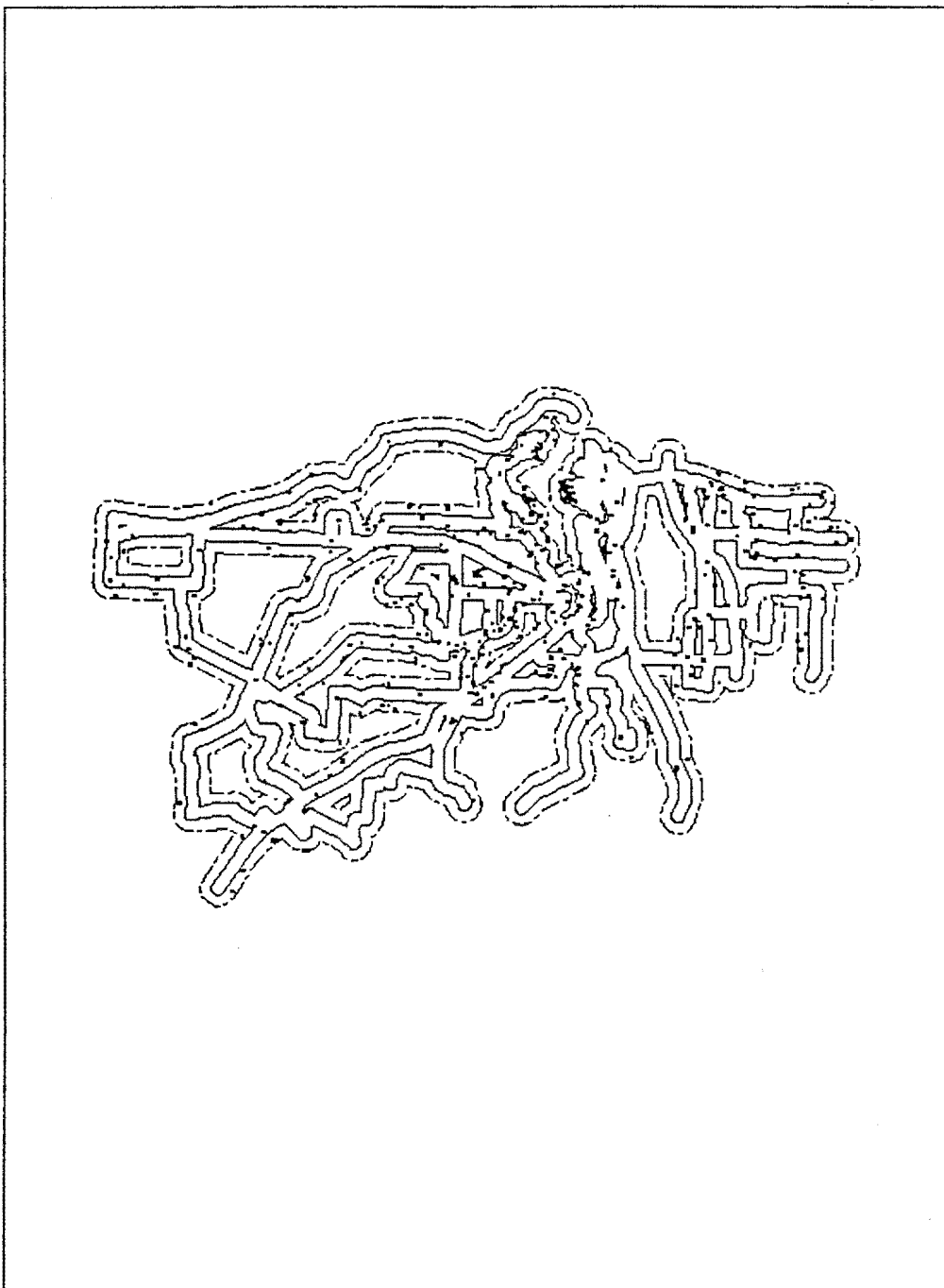


Figure 10. Original Study Area; All Employees With Low Parking Seniority. 500 ft and 1/4 mile bus-route buffer zones.

TABLE I

CHI SQUARE CONTINGENCY TABLE AND STATISTIC
POINT AND DENSITY EMPLOYEE ESTIMATES
500 FT BUFFER ZONE: ALL EMPLOYEES
ORIGINAL STUDY AREA

	<u>Employees</u>		
	<u>Inside Buffer</u>	<u>Outside Buffer</u>	<u>Total</u>
Address-Matched Data	923	1218	2141
Zip-Code Zone Data	<u>905</u>	<u>1161</u>	<u>2066</u>
	1828	2379	4207

Chi Square Statistic:

$X^2 = 0.18$

Critical Value at .01 Significance Level:

$X^2 \geq 6.64$

TABLE II
 CHI SQUARE CONTINGENCY TABLE AND STATISTIC
 POINT AND DENSITY EMPLOYEE ESTIMATES
 1/4-MILE BUFFER ZONE: ALL EMPLOYEES
 ORIGINAL STUDY AREA

	<u>Employees</u>		<u>Total</u>
	<u>Inside Buffer</u>	<u>Outside Buffer</u>	
Address-Matched Data	1654	484	2141
Zip-Code Zone Data	<u>1673</u>	<u>393</u>	<u>2066</u>
	3327	880	4207

Chi Square Statistic:

$X^2 = 8.59$

Critical Value at .01 Significance Level:

$X^2 \geq 6.64$

TABLE III

CHI SQUARE CONTINGENCY TABLE AND STATISTIC
POINT AND DENSITY EMPLOYEE ESTIMATES
500 FT BUFFER ZONE: EMPLOYEES WITH
LOW PARKING SENIORITY
ORIGINAL STUDY AREA

	<u>Employees</u>		
	<u>Inside Buffer</u>	<u>Outside Buffer</u>	<u>Total</u>
Address-Matched Data	283	381	664
Zip-Code Zone Data	<u>294</u> 577	<u>376</u> 757	<u>670</u> 1334
Chi Square Statistic:			
$X^2 = 0.002$			
Critical Value at .01 Significance Level:			
$X^2 \geq 6.64$			

TABLE IV

CHI SQUARE CONTINGENCY TABLE AND STATISTIC
 POINT AND DENSITY EMPLOYEE ESTIMATES
 1/4-MILE BUFFER ZONE: EMPLOYEES WITH
 LOW PARKING SENIORITY
 ORIGINAL STUDY AREA

	<u>Employees</u>		
	<u>Inside Buffer</u>	<u>Outside Buffer</u>	<u>Total</u>
Address-Matched Data	515	149	664
Zip-Code Zone Data	<u>584</u>	<u>86</u>	<u>670</u>
	1099	235	1334

Chi Square Statistic:

$X^2 = 20.54$

Critical Value at .01 Significance Level:

$X^2 \geq 6.64$

Route 8 Study Area

500 ft Buffer Zone: All Employees. The null hypothesis cannot be rejected based on the results seen in Figure 11 and the insignificant X^2 value of 0.96 in Table V. The estimate of buffer membership using address-matched point data offers no statistical advantage compared to zonal data in this case.

1/4 Mile Buffer Zone: All Employees. The null hypothesis cannot be rejected based on the X^2 value of 0.91 in Table VI. The estimate of buffer membership using address-matched point data offers no statistical advantage compared to zonal data in this case.

500 ft Buffer Zone: Low Parking Seniority Employees. The null hypothesis cannot be rejected based on the insignificant X^2 value of 2.26 in Table VII. The estimate of buffer membership using address-matched point data offers no statistical advantage compared to zonal data in this case.

1/4 Mile Buffer Zone: Low Parking Seniority Employees. The null hypothesis cannot be rejected based on the results the insignificant X^2 value of 1.4 in Table VIII. The estimate of buffer membership using address-matched point data offers no statistical advantage compared to zonal data in this case.

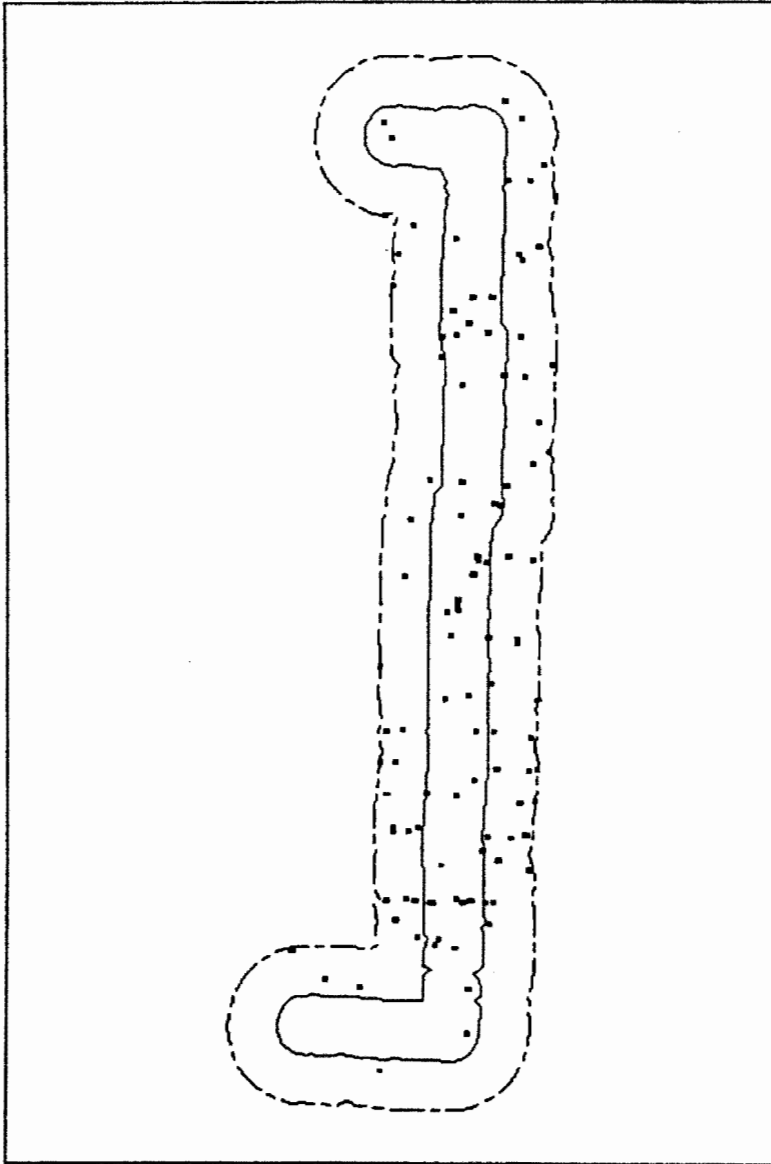


Figure 11. Route 8 Study Area; all employees within 500 ft and 1-4 mile bus-route buffer zones.

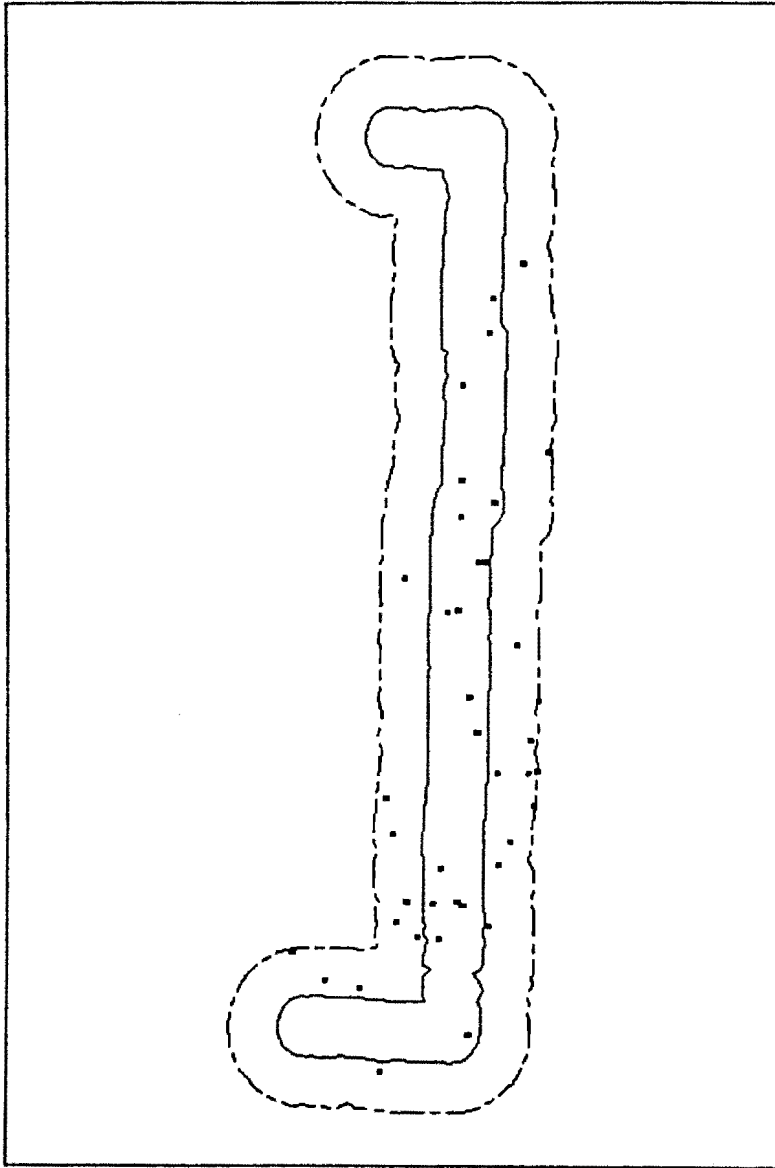


Figure 12. Route 8 Study Area; employees with low parking seniority within 500 ft and 1/4-mile bus-route buffer zones.

TABLE V

CHI SQUARE CONTINGENCY TABLE AND STATISTIC
POINT AND DENSITY EMPLOYEE ESTIMATES
500 FT BUFFER ZONE: ALL EMPLOYEES
ROUTE 8 STUDY AREA

	<u>Employees</u>		
	<u>Inside Buffer</u>	<u>Outside Buffer</u>	<u>Total</u>
Address-Matched Data	48	434	482
Zip-Code Zone Data	<u>40</u>	<u>462</u>	<u>502</u>
	88	896	984

Chi Square Statistic:

$X^2 = 0.96$

Critical Value at .01 Significance Level:

$X^2 \geq 6.64$

TABLE VI
 CHI SQUARE CONTINGENCY TABLE AND STATISTIC
 POINT AND DENSITY EMPLOYEE ESTIMATES
 1/4-MILE BUFFER ZONE: ALL EMPLOYEES
 ROUTE 8 STUDY AREA

	<u>Employees</u>		
	<u>Inside Buffer</u>	<u>Outside Buffer</u>	<u>Total</u>
Address-Matched Data	118	364	482
Zip-Code Zone Data	<u>109</u> 227	<u>393</u> 757	<u>502</u> 984

Chi Square Statistic:
 $X^2 = 0.91$

Critical Value at .01 Significance Level:
 $X^2 \geq 6.64$

TABLE VII

CHI SQUARE CONTINGENCY TABLE AND STATISTIC
POINT AND DENSITY EMPLOYEE ESTIMATES
500 FT BUFFER ZONE: EMPLOYEES WITH
LOW PARKING SENIORITY
ROUTE 8 STUDY AREA

	<u>Employees</u>		<u>Total</u>
	<u>Inside Buffer</u>	<u>Outside Buffer</u>	
Address-Matched Data	19	162	181
Zip-Code Zone Data	<u>12</u> 31	<u>169</u> 331	<u>181</u> 362

Chi Square Statistic:

$X^2 = 2.26$

Critical Value at .01 Significance Level:

$X^2 \geq 6.64$

TABLE VIII

CHI SQUARE CONTINGENCY TABLE AND STATISTIC
POINT AND DENSITY EMPLOYEE ESTIMATES
1/4-MILE BUFFER ZONE: EMPLOYEES WITH
LOW PARKING SENIORITY
ROUTE 8 STUDY AREA

	<u>Employees</u>		
	<u>Inside Buffer</u>	<u>Outside Buffer</u>	<u>Total</u>
Address-Matched Data	40	141	181
Zip-Code Zone Data	<u>32</u> 72	<u>149</u> 290	<u>181</u> 362

Chi Square Statistic:

$$X^2 = 1.4$$

Critical Value at .01 Significance Level:

$$X^2 \geq 6.64$$

Observations: Chi Square Comparisons

The chi square results summarized in Table VIV suggest that there is no statistical difference between the point and density buffer-zone estimates at the 500 ft buffer distance for both the original study area and the Route 8 study area. Statistical significance can be demonstrated at the quarter-mile buffer distance for the larger original study area, but not for the smaller Route 8 study area.

The results are mirrored for the sample subsets based on employees with low parking seniority. Again, the chi square statistic is significant at the quarter-mile buffer distance for the larger study area, but not for any of the Route 8 study area scenarios.

One interpretation of these findings is that the point-based model is sensitive to scale, as neither the smaller buffer distance nor the smaller study area provided estimates statistically different than those for the density approach.

This apparent scale sensitivity could be indicative of employee clustering observable only over larger areas or larger buffer distances. Getis and Franklin (1987) have demonstrated that the heterogeneity of a point sample can vary at different scales of analysis. This hypothesis and issue of scale-sensitivity was not pursued as part of this thesis.

TABLE IX
SUMMARY OF CHI SQUARE COMPARISONS

	<u>X² Values Original Study Area</u>	<u>X² Values Route 8 Study Area</u>
<u>500 ft Buffer</u>		
All Employees	0.18	0.96
Low Parking Seniority	0.002	2.26
 <u>1/4 Mile Buffer</u>		
All Employees	8.59*	0.91
Low Parking Seniority	20.54*	1.40

Critical Value at .01 Significance Level:

$$X^2 \geq 6.64$$

*Statistically significant at the .01 level.

It is also possible that the relatively low numbers of observations for the smaller areas and narrower buffer zone makes differentiation between the two methods impossible. Concern over the data quality issue discussed in terms of the miscoded zip code information which resulted in different total (inside and outside) buffer zone membership used in calculating the chi square values must also be raised.

Regardless of the causes behind the results, the blind assumption that address-matched data will provide statistically different results compared to a density approach must be questioned by the research results derived from the data utilized in this study.

CHAPTER IV

POINT DATA EXPLORATORY STATISTICS AND TRANSPORTATION ANALYSIS

The results reported in this chapter document the application of several research hypotheses based strictly on the OHSU point data. Descriptive statistics, multiple linear regression and two-way analysis of variance were used to demonstrate additional ways in which point data, generated from address-matching, might be applied to the analysis of transit-related spatial distributions. The chapter is organized around each statistical test as in the previous chapter.

DATA

The OHSU employee database included the following variables: Address; Zip Code; Sex; Age; Seniority (In years, based on date hired); Classification (State of Oregon Job Classification); Type_of_Employee (Classified, Faculty & Management, Students, Interns, and Volunteers); Work Period (Day, Night, Swing, or Varies); Work Time (full or part); Parking Seniority (In years, based on the date of parking permit assignment); and Car Pool Number (indicating participation in a car pool)

In addition, a variable named "Distance" was calculated based on the employee address-matched location and the nearest bus route to the employee.

None of the available variables directly provided information about the use of transit services by OHSU employees. However, low parking seniority was considered indicative of a higher-likelihood to use transit.

ANALYSIS AND RESULTS

Descriptive Statistics

Measurement of mean values for key variables and buffer zone groups were calculated from the point data in the original study area. No formal hypotheses were associated with this effort. Similar calculations for the Route 8 study area were not undertaken because of the low number of observations.

Due to a large number of missing observations in the OHSU database, particularly involving two key variables (sex and age) shown by previous research (Pucher et al., 1981; Teal, 1987) to be important contributors in profiling transit users, the data set was reduced for the cases with missing data on these variables. The net number of cases after this reduction was 1042 as compared to the total data set of 2141. The resulting set was complete in regard to all variables.

At the outset, this extreme reduction in cases introduced a high degree of uncertainty in interpreting the results. It is likely that the results were biased based on this reduction.

Table X and Table XI show the descriptive statistics grouped by all employees, the five major employee categories, and four buffer zones. The mean distance of employees to bus lines does not vary much across the employee groups. The low number of students in the reduced data set negates any significance that might be placed on the relatively low mean value of 708 feet on the distance variable for students. It also supports the supposition that some bias was introduced into the analysis by the reduction of cases discussed above. The proportion of students should be much higher than the number in the reduced data set.

Overall, the mean value of 944 feet (approximately four city blocks) supports the findings of the original study in which the route structure was found to provide adequate accessibility to transit services. The high standard deviation values around the overall employee mean, and for the employee groupings, indicates that there are a wide range of values dispersed about the mean.

The age, seniority, and parking seniority variables were not subject to unusual standard deviations and as a result they were not reported. The mean values on these variables confirm general expectations about the employee group characteristics. All of the means on these variables trend with the employee groupings as one would expect for a medical school and hospital setting.

TABLE X
 MEAN DISTANCE OF EMPLOYEES TO NEAREST BUS ROUTE
 ORIGINAL STUDY AREA EMPLOYEES:
 COMPLETE DATA CASES ONLY

<u>Description</u>	<u>No. Cases</u>	<u>Mean Distance (Feet)</u>	<u>Standard Deviation (Feet)</u>
All Employees	1042	944	1071
Faculty & Management	342	912	1024
Classified	516	941	1066
Interns	160	1012	1164
Students	14	708	1205
Volunteers	7	1469	1287
0 - 100 ft Buffer	105	47	29
100 - 500 ft Buffer	341	293	115
500 - 1/4-Mile Buffer	364	826	223
> 1/4-Mile Buffer	232	2489	1290

TABLE XI
 MEAN AGE, SENIORITY, PARKING SENIORITY
 ORIGINAL STUDY AREA EMPLOYEES:
 COMPLETE DATA CASES ONLY

<u>Description</u>	<u>Mean Age (Years)</u>	<u>Mean Seniority (Years)</u>	<u>Mean Parking Seniority (Years)</u>
All Employees	40	4.4	5.9
Faculty & Management	44	3.6	7.7
Classified	39	5.4	5.8
Interns	34	3.2	2.7
Students	30	2.8	2.7
Volunteers	48	0.7	7.4
0 - 100 ft Buffer	39	4.0	6.2
100 - 500 ft Buffer	39	4.7	5.8
500 - 1/4-Mile Buffer	40	4.3	6.0
>1/4-Mile Buffer	41	4.1	6.0

Multiple Linear Regression

Multiple linear regression was used with the 1042 records in the data subset with complete information. This statistical method is applied when seeking to demonstrate that one or more independent variables (X) have a direct relationship with a dependent variable (Y). The general form of the regression equation being as follows:

$$Y = a + bX_n$$

where,

Y = dependent variable

a = y-axis intercept

b = regression coefficients

X_n = independent variables available with the data set.

An R^2 value, indicating the combined ability of the independent variables to predict the dependent variable, is most often the criteria evaluated for the overall significance of the model. Interpretation of R^2 values is based on a scale of 0 to 1. The R^2 value of 1 indicates a perfect relationship between the dependent and independent variables. Decreasing significance is associated with R^2 values less than 1 with 0 indicating no significance.

The regression applications pursued in this thesis were exploratory in nature and were not expected to result in a high overall R^2 value. The primary goal was to demonstrate statistical significance between one or more

independent variables and the dependent variable as indicated by standardized regression coefficients (b).

The application of multiple linear regression for exploring variable relationships based on standardized regression coefficients has been discussed by Lewis-Beck (1989) among others. Two different hypotheses using multiple-linear regression were tested and are described below.

Parking Seniority And Other Employee Variables. The first regression test sought to confirm the apparent relationship between parking seniority and the other employee variables seen in the descriptive statistics. As discussed in Chapter III, the parking seniority variable was thought to represent the best available indicator of employee likelihood to use transit services. The following regression equation was used to evaluate this apparent relationship.

$$Y_{\text{park}} = a + bX_1 + bX_2 + bX_3 + bX_4 + bX_5$$

where,

Y_{park} = assigned parking seniority of the employee.

a = Y intercept

b = regression coefficients

X_1 = distance to nearest bus route

X_2 = Age

X_3 = Seniority

X_4 = Sex (Female=1, Male=2)

X_5 = Type of Employee (Faculty=1, Classified=2, Intern=3, Student=4, Volunteer=5)

Statistical significance was predicated on finding one or more significantly related independent variables in the prediction of parking seniority. A significance level of .05 was chosen with the following tests. This less demanding significance level was selected due to the exploratory nature of the hypotheses tested.

Null Hypothesis: There is no statistical significance between employee parking seniority and one or more of the characteristics of the employee as defined by the independent variables. The statistical significance criteria used in evaluating the standardized regression coefficients was a .05 confidence interval.

Alternative Hypothesis: Statistically significant (.05) relationships do exist between parking seniority and one or more of the characteristics of the employee as defined by the independent variables.

The results reported in Table XII confirmed, as was expected, that age and employee classification were statistically significant in predicting the parking seniority of the employee. Employee age showing the strongest contribution to the regression equation with a standardized regression coefficient of (.374). The significance level for this variable being (0.000) or less than 1 chance in 1000 of being caused by chance.

Employee classification (Type_of_Employee) also shows statistical significance (.009) significance in predicting parking seniority. The interpretation of this result must be more qualitative in that the values for employee classification (1 thru 5) were assigned based on general income levels associated with the groups. Faculty and Management in general assumed to be highest in salary base, classified employees second, interns third, students fourth and volunteers fifth.

The regression coefficient for Type_of_Employee is reported as a negative number implying an inverse relationship with parking seniority. This is a result of the coding scheme used in ranking the employee groups. When the variable is interpreted in light of the coding scheme, parking seniority is predicted as greatest for faculty and management employees, followed by classified, interns, students and volunteers. Overall, this regression application confirmed the general overview of variable relationships spelled out by the descriptive statistics.

TABLE XII
 MULTIPLE LINEAR REGRESSION: EMPLOYEE PARKING
 SENIORITY AGAINST ALL OTHER VARIABLES
 ORIGINAL STUDY AREA EMPLOYEES:
 COMPLETE DATA CASES ONLY

<u>Variable</u>	<u>Standardized Regression Coefficient</u>	<u>Significance</u>
Distance to Bus Line	-0.033	0.252
Age	0.374	0.000*
Seniority	0.010	0.729
Sex	-0.004	0.880
Type of Employee	-0.078	0.009*

Overall $R^2 = 0.164$

*Statistically significant at the .05 level.

Distance To Nearest Bus Line And Other Variables. The second multiple regression model evaluated the ability to predict employee distance from the nearest bus line based on the other available employee variables. This test was conducted in an attempt to delineate any relationship or pattern of employee distance from bus lines and other characteristics of the employee.

A significant relationship between one or more of the independent variables, and the prediction of distance to a bus line, were again evaluated by standardized regression coefficients. The following regression equation was used to evaluate this scenario.

$$Y_{\text{distance}} = a + bX_1 + bX_2 + bX_3 + bX_4 + bX_5$$

where,

Y_{distance} = distance of employee to nearest bus route.

a = intercept

b = regression coefficients

X_1 = Parking Seniority

X_2 = Age

X_3 = Seniority

X_4 = Sex (Female=1, Male=2)

X_5 = Type of Employee (Faculty=1, Classified=2, Intern=3, Student=4, Volunteer=5)

Null Hypothesis: There is no statistical significance between employee distance to the nearest bus line and one or more of the characteristics of the employee as defined by the independent variables. The statistical significance criteria for evaluating the standardized regression coefficients was a .05 confidence interval.

Alternative Hypothesis: A statistically significant (.05) relationships does exist between employee distance to the nearest bus line and one or more of the characteristics of the employee as defined by the independent variables.

Table XIII reports the standardized regression coefficients and overall R^2 value for this regression equation. No statistical significance at the .05 level can be supported for any of the variables and thus the null hypothesis cannot be rejected.

The results again confirm the overall findings exhibited by the descriptive statistics which showed that employee distance to bus routes, on the whole, were not closely associated with the other variables and that it cannot be predicted based on the data set.

In completing the regression analyses over 20 cases were reported as being outliers compared to the normal distribution of variable values. This fact, combined with the high-degree of missing data must be considered very limiting in the overall interpretation of the results.

TABLE XIII

MULTIPLE LINEAR REGRESSION: EMPLOYEE DISTANCE FROM
 NEAREST BUS ROUTE AGAINST ALL OTHER VARIABLES
 ORIGINAL STUDY AREA EMPLOYEES:
 COMPLETE DATA CASES ONLY

<u>Variable</u>	<u>Standardized Regression Coefficient</u>	<u>Significance</u>
Parking Seniority	-0.039	0.252
Age	0.053	0.134
Seniority	0.014	0.649
Sex	-0.013	0.679
Type of Employee	0.040	0.226

Overall $R^2 = 0.004$

Two-Way Analysis of Variance (ANOVA)

Qualitative review of a land use map for the study area, combined with observations made along the routes, indicated that multi-family residential land use was generally higher along major arterials commonly traversed by buses. Alternatively, as you move away from the arterials multi-family land use tends to decrease.

Assuming that low-income employees are more likely to be found in multi-family residences, and that they are also more likely to use transit (Pucher et al.,1981; Teal), a final statistical test was undertaken to attempt the classification of employees relative to the bus-route network.

Classification of employees based on parking seniority was sought based on location in across-route and along-route buffer segments. The research premise for this classification was that, overall, it was more likely to see greater differences in employee distributions moving out from the bus route than along the arterial strips characterized by higher multi-family residential land use.

The method applied to this attempted classification was a two-way analysis of variance. This statistical test requires a criterion (dependent) variable Y measured on an interval scale and treatment (independent) variables X which must be categorical (Silk, 1981). In this research the criterion variable (Y) was parking seniority and the treatments were across-route buffer position (X_1) specified by buffer zone membership (500 ft or 500-1/4 mile) and along-

route location (X_2) as specified by membership in one of three or four buffer cross sections.

The model for this test is described as follows:

$Y = X_1 + X_2 + (X_1 * X_2)$ where X_1 and X_2 are the categorical treatment conditions as described above and $(X_1 * X_2)$ an interaction effect between the treatment conditions.

Interpretation of ANOVA results is based on the statistical significance for each treatment condition and for the interaction effect. If either or both of the treatment conditions are found to be significant it is indicative that the classification is predicting the criterion variable as a traditional linear model similar to multiple linear regression. If the interaction effect is significant then a curvilinear relationship between the criterion and treatments is specified. The treatments in the latter case, are in effect interacting in some combination in their prediction of the criterion variable.¹ ANOVA is extremely sensitive to classification cells with missing cases. Thus, every possible classification position (buffer segment) must have at least two cases from which to calculate a mean for the criterion variable. As applied here, at least two cases must be found in every buffer segment.

Segmented buffer zones around the four alternative corridors in the original study and the Route 8 study area were evaluated in separate statistical

¹The application of ANOVA described in this section was based on Data Analysis lecture notes and suggestions from Professor William Rabiega.

runs. The research hypothesis, as stated above, was based on the assumption that across-route buffer position should show significant groupings of employees based on their parking seniority.

The results reported in Table XIV show that employee locations as grouped by across-route buffer position for the five bus routes cannot be predicted based on parking seniority.

For one case, Route 8, statistical significance is seen for along-route segment position. In this case the results suggest that parking seniority is related to position in the separate along-route segments. Although this was not the relationship hypothesized, general trends in residential land-use for this high-density multi-family area can be qualitatively associated with the segment positions.

Figure 12 shows a map of the Route 8 study area with associated mean parking seniority values for the employees located within each segment. The map shows a general trend of higher parking seniority in segment 3 as compared to the other two segments. This roughly corresponds with land-use in the study area in that multi-family land use is much more concentrated in segments 1 and 2. However, as this type of relationship could not be demonstrated for any of the other routes tested, and since significant along-route clustering was not the focus of this analysis, any significance associated with the apparent pattern must be considered anecdotal.

TABLE XIV

TWO-WAY ANALYSIS OF VARIANCE: PARKING SENIORITY AND
ACROSS-ROUTE AND ALONG-ROUTE BUFFER SEGMENTS

<u>Route</u>	<u>F-Ratio</u>	<u>Significance</u>
Route 8		
Across Buffer	0.762	0.385
Along Buffer	3.627	0.030*
Interaction Effect	2.562	0.082
Route 1		
Across Buffer	1.425	0.233
Along Buffer	0.981	0.402
Interaction Effect	0.572	0.633
Route 2		
Across Buffer	0.057	0.811
Along Buffer	0.794	0.498
Interaction Effect	1.33	0.264
Route 3		
Across Buffer	0.040	0.842
Along Buffer	0.345	0.793
Interaction Effect	0.449	0.718
Route 4		
Across Buffer	0.361	0.549
Along Buffer	0.187	0.829
Interaction Effect	1.098	0.337

*Statistically significant at the .05 level.

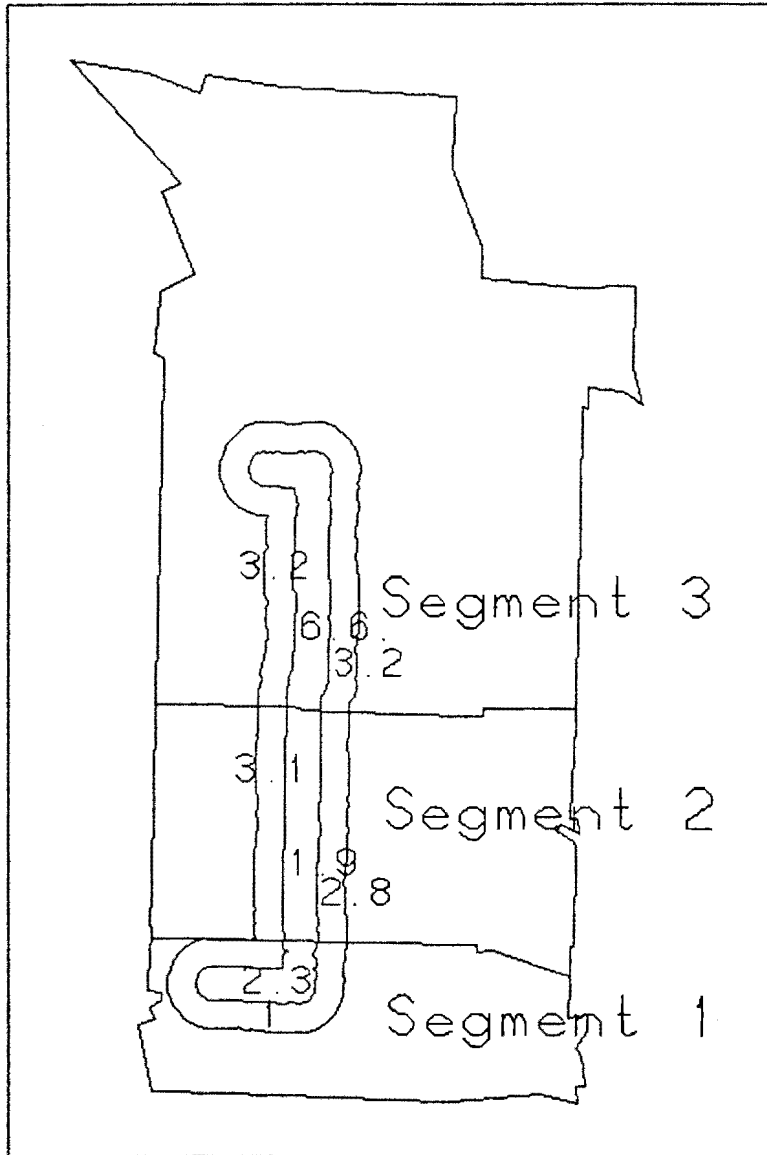


Figure 13. Mean parking seniority associated with buffer segments created for Route 8. The mean values associated with the three horizontal zones labelled 1, 2, 3 were shown to be statistically significant in classifying employees based on parking seniority. Qualitative assessment of land-use for the area reveals an apparent association between multi-family land-use and lower parking seniority.

Observations: Exploratory Point-Based Statistics

The techniques and hypotheses tested in this chapter were of limited success in demonstrating significant patterns of employee distributions relative to transit services.

The lack of findings can be attributed to data quality concerns, research design and ultimately to the possibility that no practically discernable spatial patterns exist for the given problem. Data quality concerns, evidenced first by the high-degree of missing attributes in the data set and then by the large number of extreme values in the parking seniority field during regression, cannot be discounted.

Questions over research design must also be considered. Much emphasis was placed on parking seniority as an indicator of a greater likelihood to need transit services on the part of the OHSU community. The use of this variable as a strong indicator of the likelihood of an employee to use transit must be questioned.

Some criticism over its use might be negated by the apparent relationship between parking seniority and employee groupings as demonstrated by the descriptive statistics and significant regression results, which showed a relationship between employee categories and parking seniority. Furthermore, parking seniority did offer an interval class variable required by most of the statistics, in a data set primarily characterized by categorical descriptions.

Further delineation of income classes for employees based on the State of Oregon Job Classifications available in the database may have added to the analysis. This information, would have allowed for classification of employees by income levels associated with their classification. This would have provided an alternative, and perhaps more meaningful, dependent variable from which to pursue the multi-variate statistical techniques demonstrated in this chapter. However, in light of the data quality problems in the data set this was not pursued.

Since for the most part the research was inconclusive, the overall research hypothesis which predicted an ability to demonstrate significant statistical applications based on the OHSU point data must be negated.

CHAPTER V

SUMMARY

This thesis sought to demonstrate the degree to which the additional spatial detail available from point data was beneficial to transportation planning. GIS address-matching, a software tool designed to generate point locations from common tabular databases which include address information, is expected to make point data more readily available for transportation planning. As the use of data generated from automated address-matching has not been as commonly applied to such problems, there has been limited research directed at critical examination of the benefits and limitations of its use.

A transportation problem at Oregon Health Sciences University (OHSU) in Portland, Oregon provided the empirical setting for evaluating the use of data generated from address-matching. The site and situation of OHSU and a growing commuter base have created increasing transportation-related problems at OHSU.

The spatial distribution of OHSU employees in relation to existing and alternative bus routes was of interest to transit planners for evaluating employee accessibility to transit services to OHSU. GIS address-matching of the OHSU employee database and subsequent GIS overlay processing

provided detailed information about employee locations within varying buffer zones around bus routes.

The research question of this thesis was whether or not, in the context of the OHSU problem and in regard to the additional computer processing required for address-matching, point information provided additional insight and analytical possibilities in comparison to the use of aggregated data.

In Chapter III a comparison was made of point and density techniques for estimating bus-patronage potential along bus routes for the OHSU data set. This comparison revealed limited statistical difference between the two methods. Estimated bus-patronage potential was statistically different at a one-quarter mile buffer distance for a large study area, and for a subset of the population in this study area deemed of higher potential to need transit service. However, at a 500 ft buffer distance, and for all scenarios in a small-area analysis around a single route, statistically significant differences between the two methods could not be demonstrated.

From a practical perspective, the absolute numbers estimated by the two methods were within such a close range of each other, even when the statistical comparisons were significant, the presumed superiority of point data for estimating potential bus patrons in the OHSU example was questioned.

In Chapter IV the OHSU point data was used to test for spatial patterns of employee distributions relative to the bus-route network in an attempt to demonstrate additional modeling capabilities available with point data.

Descriptive statistics were calculated, and hypotheses using multiple linear regression and analysis of variance were tested, in an attempt to reveal employee groupings in relation to the route network.

Results from this part of the research were for the most part inconclusive in demonstrating any pattern of employee distribution relative to the route network. Data quality problems in the OHSU employee database combined with the lack of information about employee mode-choices limited the interpretation of results.

CONCLUSIONS

The additional spatial detail available via address-matching, and the broad range of data sources it makes accessible, represents a great deal of potential for geographic analysis in general. In the case of the OHSU transportation problem, address-matching provided detailed information about employee residences in relation to transit services. This information could have been used for further efforts in planning and marketing of transit services to OHSU.

The point detail also allowed for hypothesis testing about potential spatial distributions that was not possible with aggregated data. The tests performed were limited in scope, and were limited by data availability and quality, which in further research could be extended for more thorough and conclusive testing.

This modeling capability presents new opportunities for expanding geographic understanding about transportation related issues.

A disadvantage of adopting point data as the unit of analysis is the issue of technology "costs". Address-matching requires additional software and computer processing of data not required by more common alternatives. However, as address-matching and GIS technology become more commonly applied, the issue of technology "costs" will not be as important of a criteria in considering its use.

A further disadvantage to address-matching centers on the issue of data quality. Miscoded address information in the subject database combined with missing street segments in the TIGER file can result in incomplete address-matching. The problem of miscoded address descriptions will most likely be a concern for any application of address-matching. Additional processing of the data can minimize this problem, but the degree to which this is necessary should be considered prior to implementing an address-matching approach.

Data quality associated with the TIGER file street map (e.g. missing and cartographically-distorted street segments) is somewhat more problematic. Cooperative efforts among city and regional governments to enhance and maintain the TIGER file will hopefully decrease address-matching problems associated with the TIGER representation of street geography.

Other concerns can be raised over generalization in point placement from the address-matching process. Uniform offset distances from streets

segments specified as part of the processing, and the potential for distorted along-street positioning are two examples of potential generalization problems.

The street segment offset distance for point placement is uniformly specified for all points during the address-matching process. This uniformity distorts the real world variation in the distance of building locations from streets. It also introduces a potential source of error when combined with polygon overlay as seen with the buffer zones used in this thesis. Different offset distances could result in different counts of points inside the buffer polygons, particularly along the streets at the periphery of the buffer zone.

Additional generalization involving along-street positioning, can occur when block numbering systems and assignment of numbers to buildings are not representative of building position along the block as interpolated by the address-matching process. This problem can result in artificial clustering of points along a block, particularly at the two ends of a segment, when in reality points should be more evenly spaced.¹ This type of artificial clustering resulting from the address-matching process again could introduce error in an overlay analysis involving point counts.

¹This problem was noted by Professor Richard Lycan after address-matching all of the tax parcel data for a neighborhood in Northeast Portland. Address number assignments for the land parcels in many cases were grouped closely to the numbers at each end of the block.

FURTHER RESEARCH

Further research is needed to develop and evaluate applications of GIS and address-matching to transportation related questions. As the OHSU situation in many respects is representative of a common problem involving transportation to major employment centers, further evaluation of the techniques demonstrated in this thesis is warranted. Extension of these methods using multiple employee databases from businesses in central business districts or outlying business parks could more clearly define the benefits and limitations of using point data for this type of transportation analysis.

The application of address-matching to this problem would benefit from additional information about mode-choice by the employees. Knowledge of the means by which the employees are currently commuting would provide a starting point for profiling potential service users. It would also allow for additional pattern analysis of employee distributions in relation to services to determine if spatial accessibility was a factor in an individual's choice of transportation.

Identification and profiling of potential transit users could also benefit from GIS integration of land-use information with the bus-route network and employee locations. Spatial queries for areas of higher concentrations of multi-family housing could identify areas on which to test the extent to which residential land-use is correlated with individual transportation mode choices.

Incorporation of transit travel-time and route-transfer requirements to a central employment destination would also enhance the modeling capabilities for this type of GIS application. Identification of areas with high numbers of potential transit users subject to unacceptable time costs for using transit could focus efforts of planners on providing more direct services.

Bus Patronage Estimating Models

A number of comparisons between point-based and zone-based bus-patronage models could be tested for model sensitivity to the unit of areal aggregation. Comparisons could be made between address-matched point data and data aggregated by traffic analysis zones (TAZ's), or census tracts, following the same route-buffering and overlay approach used in this thesis. Integration of land-use information within the areal units could also provide a more focused method of assigning the population within the areas as compared to the assumption equal distribution across the areal units as followed in this thesis.

For cases like OHSU where the employee data is not coded for areal units such as TAZ's or census tracts, address-matching and GIS overlay using these other areal boundaries would provide a means by which to aggregate the data for the zones. If comparison of patronage estimates based on the original point data, and the point data aggregated by areal unit, did not reveal significant differences, address-matching would still be useful as a method to aggregate data. This methodology would allow for integration of point data with

other socioeconomic data available with the zonal units (Dueker and Vrana, 1990).

Bus-stop locations along routes could also be pursued as the basis for estimating potential transit patronage. Radial buffers around bus stops would provide a refined means by which to identify potential transit users along a route. Rather than buffering the entire route, radial buffers from bus-stop locations could be used to identify only those residences within a specified distance of a stop.

Address-matching could facilitate this approach by locating bus-stop point locations in the street network based on address descriptions maintained for bus routes in transit scheduling programs (Orrell, 1990). This would preclude the need to digitize bus routes as part of the analysis. It would also open some interesting questions about how changes in bus-stop locations might alter accessibility to services for a specified user group.

Aggregation of point data to the street segments along which they are originally matched would also make possible additional bus-patronage modeling. This approach would allow for weighting street segments according to the number of employees along each segment in the network. Analysis of potential route patronage using network demand could then be employed to evaluate existing routes or an alternative route through the network structure (Shaw, 1989). Total demand for services along a given route, or within a network-defined walking distance from a route, would provide an alternative to

estimating potential patronage. This would also eliminate any potential problems associated with the generalization of point locations resulting from address-matching as discussed above.

These and other methods for modeling potential bus-patronage all could be pursued in the GIS environment. The modeling opportunities available with GIS and address-matching for transportation analysis and planning are many, but further work is needed to assess its benefits and limitations.

REFERENCES CONSULTED

- Barff, Richard A. and Dawn E. Hewitt. 1989. "Second-Order Analysis of Bivariate Point Patterns." Professional Geographer 41(2):183-89.
- Burrough, P.A. 1986. Principles of Geographic Information Systems. Oxford: Clarendon Press.
- Dueker, Kenneth J. and Ric Vrana. 1990. "Geographic Information Systems in Urban Public Transportation." Portland, Oregon, Portland State University: Center for Urban Studies.
- Ebdon, David. 1977. Statistics in Geography A Practical Approach. Oxford: Basil Blackwell.
- ESRI. 1990. Arc/Info User Documentation. Environmental Systems Resources, Redlands, CA.
- Haining, R. 1982. "Describing and Modeling Rural Settlement Maps." Annals of the Association of American Geographers 72:211-23.
- Hanum, K. 1989. "Homestead Residents still leery of OHSU." Oregonian, June 1, 1989, sec. M W-MP,p.1.
- Hanson, Susan. 1986. "Dimensions of the Urban Transportation Problem." In The Geography of Urban Transportation 3-23. Edited by Susan Hanson. New York: Guilford Press.
- Hanson, S. and P. Hanson. 1981. "The Travel-Activity Patterns of Urban Residents: Dimensions and Relationships to Sociodemographic Characteristics". Economic Geography 57:332-347.
- Hecht, A. 1974. "The Journey-To-Work Distance In Relation To The Socioeconomic Characteristics of Workers." Canadian Geographer 18(4):367-378.

- Hunt, David T., Stephen E. Still, J. Douglas Carroll, and Alan O. Kruse. 1986. "A Geo-Demo-Graphic Model For Bus Service Planning and Marketing." Paper presented at the 65th Annual Meeting of the Transportation Research Board.
- Getis, Arthur. 1984. "Interaction Modeling Using Second-Order Analysis." Environment and Planning A 16:173-183.
- Getis, Arthur and Janet Franklin. 1987. "Second-Order Neighborhood Analysis of Mapped Point Patterns." Ecology 68(3):473-477.
- Lewis-Beck, Michael. 1989. Applied Regression An Introduction. Sage University Paper series on Quantitative Applications in the Social Sciences,07-022. Beverly Hills and London: Sage Publications.
- Lupien, Anthony E., William H. Moreland and Jack Dangermond. 1987. "Network Analysis in Geographic Information Systems." Photogrammetric Engineering and Remote Sensing 53(10):1417-1421.
- Lycan, Richard and Jim Orrell. 1989. "An Analysis of Bus Ridership Potential To Oregon Health Sciences University Using A Geographic Information System Approach." Portland, Oregon: Center for Urban Studies, Portland State University.
- Marx, Robert W. 1990. "The TIGER System: Yesterday, Today and Tomorrow." Cartography and Geographic Information Systems 17(1):89-97.
- Nyerges, T. L. and K. J. Dueker. 1988. "Geographic Information Systems in Transportation, Technical Report." U.S. Department of Transportation, Federal Highway Administration Planning Division.
- Nyerges, Timothy L. 1989. "GIS Support for Micro-Marco Spatial Modeling." Auto-Carto 9 Proceedings 567-579. Falls Church, Va.: ACSM-ASP
- Orrell, Jim. 1990. "GIS and Mapping Activities in Transit Organizations: Using the TIGER File As a Common Geographic Base." Portland, Oregon: Center for Urban Studies, Portland State University.
- Muehrcke, Phillip C. 1990. "Cartography and Geographic Information Systems." Cartography and Geographic Information Systems 17(1):7-15.

- Pucher, John, Chris Hendrickson and Sue McNeil. 1981. "Socioeconomic Characteristics of Transit Riders: Some Recent Evidence." Traffic Quarterly 35(3):461-483.
- Schwartz, Joe. 1989. "The Census Means Business." American Demographics 11:18-23.
- Shaw, S-L. 1989. "GIS As A Decision Support Tool in Transportation Analysis." paper presented at the Annual Meeting of the Association of American Geographers.
- Silk, John. 1979. Statistical Concepts in Geography. London: Allen & Unwin.
- Silk, John. 1981. Analysis of Variance. Concepts and Techniques in Modern Geography Series 32. Norwich: University of East Anglia.
- Sosslau, Arthur B. and James J. McDonnell. 1984. "Uses of Census Data For Transportation Analysis." Transportation Research Record 981:59-70.
- Teal, Robert F. 1987. "Carpooling: Who, How and Why." Transportation Research - A 21A(3):203-214.
- U.S. Department of Commerce Bureau of the Census. 1990. Tiger/Line Precensus Files, 1990 Technical Documentation. Washington: The Bureau.
- Warren, William D. 1988. "Impacts of Land Use on Mass Transit Development: A Comparison of Canberra and Springfield." Transportation Quarterly 42(2):223-242.
- Waters, Nigel M. 1989. "Big Bytes, Micro Bytes, Tid Bytes and Nibbles Do You Sincerely Want To Be a GIS Analyst?" The Operational Geographer 7(4):30-35.
- Williams, K. 1971. "Do You Sincerely Want to be a Factor Analyst?" Area 3:228-230.