

## ABSTRACT

Title of Dissertation: BAYESIAN APPROACHES TO LEARNING FROM DATA HOW TO UNTANGLE THE TRAVEL BEHAVIOR AND LAND USE RELATIONSHIPS  
Dissertation Directed By: Marco Giovanni Scuderi, 2005  
Dr. Kelly Clifton, Department of Urban Studies and Planning

The body of research on land use and travel behavior relationships reaches widely different conclusions with results varying even when considering a single author. The hypothesis of this research is that these differences arise, in part, from the fact that the vast majority of these studies do not address all the theoretical travel behavior tenets and are therefore ad-hoc in nature. An inductive approach to the study of the relationships between land use and travel behavior, prior to carrying out traditional deductive studies, can help improve the outcomes by providing an opportunity to identify and test such relationships.

With data sourced from the 2001 National Household Travel Survey Add-On, supplemented with local land use data, this study uses heuristic search algorithms to evaluate relationships hidden in the data without these being framed, a priori, by specific statistical constructs. Bayesian scoring is used to evaluate and compare the results from actual data collected for the Baltimore Metropolitan Area with the set of predominant conceptual frameworks linking travel behavior and land use obtained from the literature.

Results show that socioeconomic factors and land use characteristics act in a nested fashion, one in which socioeconomic factors do not influence travel behavior independently of land use characteristics. The land use travel behavior connection is

specifically strong only for particular combinations of socioeconomic characteristics and a land use mix which includes both moderate residential densities and a significant amount of commercial opportunities.

The study also finds that the heuristic search approach to derive relationships between land use and travel behavior does work, that this technique needs to be fine tuned for the proper use of spatially explicit data, and that although the research outputs are an unbiased representation of the land use travel behavior relationships, they need proper interpretation, especially in light of persisting theoretical questions still driving this research field.

The study concludes that an inductive approach to the analysis of the relationships between land use and travel behavior provides valuable knowledge of the data that can be used to better formulate deductive studies, so that the two methodologies are complementary to each other.

BAYESIAN APPROACHES TO LEARNING FROM DATA HOW TO  
UNTANGLE THE TRAVEL BEHAVIOR AND LAND USE RELATIONSHIPS

By

Marco Giovanni Scuderi.

Dissertation submitted to the Faculty of the Graduate School of the  
University of Maryland, College Park, in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
2005

Advisory Committee:  
Dr. Kelly Clifton, Chair  
Dr. Charles Christian  
Dr. Jochen Albrecht  
Dr. Gerrit Knaap  
Dr. Qing Shen

© Copyright by  
Marco Giovanni Scuderi  
2005

## PREFACE

With our ever-changing society and increased demand for goods and services, urban landscapes are changing rapidly and nowhere is this more evident than in land use change. Transportation choices are affected by these changes and in turn there is a need to know and understand how transportation conditions and availability influence land use.

Despite twenty plus years of studies on this subject, there is still no unified theory and there remain conflicting results and proposed frameworks about the possibility of recursive effects between the two domains. The research community finds itself divided. Some researchers advocate more in-depth modeling attempts, for example, to maximize the effects of traffic reductions that can be obtained by proper land use practices. Others argue that the relationship between land use and transportation cannot be modeled and call for alternative approaches, such as the study of collateral effects that transportation has on the environment. In the search for more compelling evidence on this matter, studies on urban growth, sprawl, land use and travel behavior have been published in great numbers. In fact since the urbanization of North America and most of Western Europe, particularly from the end of the 1800s through the 1900s, interest has intensified and has today expanded to embrace the understanding of urban form, the role of transportation and telecommunication technology, human behavior and social dynamics. Cities are in a

state of continuous change, and the movement of people and goods through different land uses can be expected to continue. These movements result in a relationship of challenging complexity that links travel behavior and land use.

## ACKNOWLEDGMENTS

Writing a dissertation is like crewing on a sailing vessel with a skipper who makes decisions, a boatswain, who is key for the execution of those decisions, and shipmates who carry out the various tasks at hand.

My role in this dissertation is that of the shipmate who has convinced the skipper and the boatswain to chart a particular course, and it is up to them to make sure the passage is safe. I was lucky to have many skippers in all of my committee members, Dr. Kelly Clifton, Dr. Charles Christian, Dr. Jochen Albrecht, Dr. Qing Shen and Dr. Gerrit Knaap, the most disciplined boatswain in my wife Alex, the unconditional love and support of my parents Saro and Pina, and many other colleagues who shared with me the enthusiasm of reaching port, or handing-in this manuscript.

To all of you, thank you for all your patience listening to me, thank you for your patience correcting my mistakes, thank you for your patience in reading about a very obscure subject. Hopefully our work will change that last fact...

## TABLE OF CONTENT

Chapter 1: Introduction .....	1
1.1 Research Question .....	2
1.2 Research Motivation .....	3
1.3 Conceptual Framework .....	5
Chapter 2: Literature Review .....	10
2.1 Travel Behavior Theory .....	13
2.2 Methodological Approaches to the Study of the Land Use Travel Behavior Relationships .....	20
2.3 Evidence on the Land Use Travel Behavior Connection .....	33
2.4 Concluding Remarks for the Literature Review in Travel Behavior and Land Use .....	42
2.5 Background Notes on Bayesian Belief Networks and Notable Research in the Field .....	44
Chapter 3: Research Design .....	50
3.1 Bayesian Belief Networks .....	53
3.1.1 Structure Learning .....	54
3.2 Methodology .....	61
3.2.1 Study Area and Data .....	62
3.2.2 Model Runs .....	69
3.2.3 Measurements .....	71
Chapter 4: Findings .....	75
4.1 Cross-Tabulations and Chi Square Statistics .....	75
4.2 Bayesian Belief Networks for Travel Mode Choice with Individual Trips and Tract Level Aggregated Land Use Measures – Model 1 .....	79
4.2.1 Model 1, Decision Tree .....	83
4.2.2 Binary Mode Choice Modeling: Privately Owned Vehicles, with Individual Trips and Tract Level Aggregated Land Use Measures ..	89
4.2.3 Binary Mode Choice Modeling: Walking, with Individual Trips and Tract Level Aggregated Land Use Measures .....	92
4.2.4 Binary Mode Choice Modeling: Transit, with Individual Trips and Tract Level Aggregated Land Use Measures .....	95
4.3 Building a Conceptual Model .....	99
4.3.1 How to Interpret Dags and Decision Trees .....	102
4.3.2 Cause and Effect or Feedback Loops? .....	105
4.4 Bayesian Belief Networks for Both Travel Choices and Land Use Characteristics Aggregated at the Tract Level – Model 2 .....	107
4.5 Bayesian Belief Networks for Both Travel Choices and Land Use Characteristics Aggregated at the Zip Level – Model 3 .....	110
4.5.1 Threshold Analysis .....	112
4.6 Model 1 Stability Tests .....	117



4.7 Internal & External Validity .....	119
4.7.1 Graph Representation Validity .....	123
4.8 Synthesis .....	123
Chapter 5: Discussion .....	125
5.1 The Heuristic Perspective on Behavior and Land Use .....	126
5.2 Major Study Limitations and its Overall Applicability .....	137
Chapter 6: Conclusion .....	141
6.1 Possible Benefits from a Heuristic Search Approach .....	141
Appendix A: Bayesian Statistics .....	146
Appendix B: Local Dags for the Parent Set of Travel Mode Choice .....	146
Bibliography .....	155

## GLOSSARY OF ACRONYMS

BBN: Bayesian Belief Network  
 CPD: Conditional Probability Distribution. See also CPT  
 CPT: Conditional Probability Table  
 DAG: Directed Acyclic Graph  
 GA: Genetic Algorithm  
 MPO: Metropolitan Planning Organization  
 NHTS: National Household Travel Survey  
 VO: Variable Ordering

## LIST OF TABLES

Table 2.1:	Authors, methods and findings on land use travel behavior relationships .....	12
Table 3.1:	Number of directed acyclic dags as a function of the number of nodes .....	55
Table 3.2:	Variables and data classification .....	71
Table 4.1:	Frequency distribution for trip distance .....	76
Table 4.2:	Cross tabulation of trip distance by trip mode .....	76
Table 4.3:	Frequency distribution for mode choice .....	77
Table 4.4:	Cross tabulation of percentages of low residential densities with mode choice .....	77
Table 4.5:	Cross tabulation of percentages of low residential densities with mode choice .....	78
Table 4.6:	Chi-Square and Cramer’s V tests for income vs. travel mode choice .....	79
Table 4.7:	Sample conditional probability table instances for land use and income variables .....	109
Table 4.8:	Sample conditional probability table for land use variables only ....	111
Table 4.9:	Thresholds for mode choice likelihood and land use mix .....	113
Table 4.10:	Spatial query legend .....	116
Table 4.11:	Variable substitution summary for Model 1 .....	119
Table 4.12:	T-tests for the difference of number of drivers and cars in households between the local data and the national NHTS sample .....	121
Table 4.13:	Selected percentage differences for variables in both the local NHTS Add-On and the national NHTS data .....	122
Table 5.1:	Sensitivity of specific travel mode choice to all variables .....	133

## LIST OF FIGURES

Figure 1.1:	Deductive and inductive research approaches .....	7
Figure 2.1:	Badoe and Miller’s complex representation of the travel behavior land use connection .....	18
Figure 2.2:	Travel - Land Use elasticity .....	22
Figure 3.1:	Wegener’s “model of urban models”, adapted and simplified form	51
Figure 4.1:	Simplified BBN for all available transportation mode choices considered in “modechoice” .....	82
Figure 4.2a:	Decision tree for Model 1 .....	88
Figure 4.2b:	Sample probabilities query for the decision tree of Model 1 .....	88
Figure 4.3a:	Binary BBN for privately owned vehicles .....	90
Figure 4.3b:	Decision tree for privately owned vehicles .....	91
Figure 4.4a:	Binary BBN for walking .....	93
Figure 4.4b:	Decision tree for walking .....	94
Figure 4.5a:	Binary BBN for transit .....	99
Figure 4.5b:	Decision tree for transit .....	98
Figure 4.6a:	BBN for commercial space .....	101
Figure 4.7:	Building a conceptual model for income, household vehicle count and travel mode choice .....	104
Figure 4.8:	Decision tree for trip distance .....	106
Figure 4.9:	Model 3: five digit zip areas, three parents of “modechoice” .....	112
Figure 4.10:	Model 3: five digit zip areas .....	112
Figure 5.1:	BBN for low density residential .....	132
Figure A.1:	A graphical model for the wet grass example .....	147
Figure A.2:	Serial, diverging, and converging connections respectively .....	148
Figure A.3:	A directed acyclic graph. The probabilities to specify are shown ...	152
Figure A.4:	Convergent Bayesian Belief Network with numerically complete sample .....	153
Figure B.1:	BBN for race .....	155
Figure B.2:	BBN for medium density residential .....	157
Figure B.3:	BBN for accessibility to transit for residential areas .....	158
Figure B.4:	BBN for trip distance .....	159
Figure B.5:	BBN for income .....	160
Figure B.6:	BBN for vehicle count .....	161
Figure B.7:	BBN for road density .....	162
Figure B.8:	BBN for education .....	163
Figure B.9:	BBN for number of children in households .....	164
Figure B.10:	BBN for age class .....	165
Figure B.11:	BBN for driver status .....	166

## LIST OF MAPS

- Map 3.1: Study area: Baltimore Metropolitan Area
- Map 3.2a: Discrete trip origins
- Map 3.2b: Aggregated trip origins, by tract
- Map 4.1: Sample Spatial Queries with probabilities of a trip being made by car, foot or transit

## Chapter 1: Introduction

The aim of this study is to provide an analytically derived representation of the relationships among land uses, transportation and travel behavior, one that is less prone to undue subjectivity than previous research. It provides the opportunity to expose the information contained in dependent measurements without overlaying specific conceptual frameworks and constraints inherent in other statistical approaches, such as the need to specify the role of variables and their distributions. In doing so the study presents a method capable of providing all researchers with a set of consistent and independently derived relationships that can be used as the basis for discussion when trying to bridge differences in the study of travel behavior and land use characteristics. The approach contributes a new and original method to the analysis of complex spatial and behavioral systems such as human interactions within the urban environment, and it presents an opportunity to expand our theoretical knowledge in this field.

To fulfill the objectives of this research, the analysis centers on the uncertainties around the nature of these relationships, which manifest themselves in a number of issues. For example, it is still difficult to determine the social and economic implications of infrastructure investments even when they are limited in size and scope. A new interchange can lead to increased congestion or worse. When expensive public transit projects do not result in the forecasted numbers of travelers, taxpayers object to the unjustified use of public funds. Some Transit-Oriented Developments (TOD) are successful while others are not. Transportation forecasting is made more uncertain by not taking into account changing travel demands, which result from a variety of land uses. All these problems could benefit from a more

specific and quantitative characterization of the relationship between land use and the associated individual travel behavior.

### **1.1 Research Question**

The fundamental research question that follows is: what are the nature, magnitude and direction of the relationships between land use and travel behavior? Currently the research community is divided in its responses. Some argue that even if such linkages exist, they are of marginal interest and not influential enough to condition our daily lives or form the basis of sound transportation policy. The existing body of theoretical knowledge on travel behavior focuses on what people do over space, and how people use transport. It asks broad questions in relation to activity patterns, and treats time and travel costs as utility measures. However the majority of researchers are working with much narrower research questions that link specific factors together, and often artificially isolate the conditions that surround our studies, just to make them manageable. So in essence, this piecemeal approach has not yet reached the much hoped-for consensus on the issue at hand, with the effect that at one extreme end of the research findings some argue that there is no detectable connection between travel behavior and land use, while others argue that land use and the human response to transportation needs are so intimately related that even the desire for a cup of coffee can affect total traffic patterns and roadside gridlocks (McGuckin 2004).

Such a variety of answers to the same question is cause for concern. The literature review that follows provides an overview of the many authors and their studies concerned with this particular problem. One point I am particularly interested in is, not so much what are the differences among these various findings but rather, why are we still dealing with such wide discrepancies after so many years of research.

The problem faced in trying to understand the relationship linking land use and travel behavior is that it is extremely difficult to distinguish cause and effect between the two domains. Numerous researchers have made endeavors in such studies and the most frequent conclusion is that transportation does in fact affect urban form (Cervero 1989; Cervero and Seskin 1995) and that, in turn, urban form affects travel choices. Such a conclusion was derived by means of deductive scientific methods whereby researchers review existing theoretical concepts pertaining to a given problem, organize a methodological solution for the study of such problems, assemble relevant data, carry out the analysis based on the chosen methodology, obtain and discuss the results and, finally, feed the findings back into the body of existing theoretical knowledge. In following this process, two problems arise. One stems from the fact that most researchers source their theoretical frameworks from the same body of knowledge without questioning its validity. Secondly, the research methodologies, statistical models, comparative analyses of sorts, are always constructed in response to the assumed theory, which is hardly ever fully integrated in any given study.

If the fundamental theoretical beliefs are incorrect, misguided or incomplete, a deductive approach will not yield possible alternative findings. At the same time deductive reasoning also fails when asked to “prove something” because there is no unique way to ask the same question.

## **1.2 Research Motivation**

The motivation behind this research is that if so many of these ad hoc studies are inconclusive, a more inductive approach to the analysis of the relationships between land use and travel behavior is needed. An inductive approach could provide insights that would either confirm our established conceptual models, highlight some possible

erroneous tenets, or provide a different and much needed view of the factors at play. More importantly the inductive approach is assumed to be capable of handling all aspects of the theory of travel behavior, assuming appropriate data exists, no matter the complexity of the problem. The inductively derived relationships could then be used to either confirm our current theoretical frameworks, if they are indeed similar to those established in the literature, or provide the basis for alternative explanations. Should the findings differ from the most accepted frameworks, the researchers should critically discuss the resulting relationships, as they provide the clue to possible alternative explanations.

More specifically this research is guided by the author's interest in the following questions:

- Are travel mode choices more strongly influenced by socioeconomic factors than by land use characteristics, or do they interact with each other?
- What types of land use affect mode choice? In particular what is the role of density and commercial land uses?
- What is the sensitivity of different transportation mode choices to different land use variables? Are there some unique land uses that have an influence only on specific travel modes?
- What is the appropriate spatial scale at which to aggregate land use measures to make sure their effects are appropriately analyzed?

These tenets are often at the center of discussion among the research community.

In an inductive environment, these questions can be asked without the confinements dictated by specific statistical constructs or analytical methods.



Inductive reasoning breaks away from the deductive reasoning process and allows the analyst to query the data directly for possible relationships within them, giving the research analyst greater confidence about the theoretical framework to be employed, one less fragmented and more universal than what we are currently working with.

The one significant obstacle to such a study is that there are many ways to inductively derive relationships among constructs and variables, probably more than with standard deductive methods. In fact some authors (Clifton and Handy 2001:12) argue that such an approach is qualitative in nature. The challenge is to proceed in such a way as to ensure the utmost scientific rigor of the inductive approach used for this study.

To provide objectivity this study makes use of heuristic computer algorithms to “learn” the appropriate structure to model statistical relationships among the variables used to capture land use and travel behavior. The method of choice is based on Bayesian Belief Networks (BBN) that, like causal networks, encode relationships among variables and provide a graphical and quantitative representation for them. Furthermore, the analysis will cover a large number of urban environments within a metropolitan region, in line with the most recent and rigorous studies, an approach that can potentially cover a varied continuum of urban built environments.

### **1.3 Conceptual Framework**

This research is undertaken as a complementary line of inquiry to current deductive methodologies, which tend to take a “top down” approach, by employing an inductive or “bottom up” approach. Deductive reasoning requires an a priori hypothesis of theoretical constructs, and its models and data collection are based upon such

reasoning.

Instead, the contribution of the inductive reasoning approach is that it allows for the investigation of statistical relationships between various variables without imposing any constraints or overarching theoretical construct. It provides a foundation to explore different hypotheses leading to the emergence of alternative theories, which in turn can improve future empirical investigations using deductive methods. One limitation of this inductive approach can be found in the need for larger, more comprehensive datasets.

This concern is particularly felt in travel behavior research since travel data are often collected with deductive theoretical frameworks in mind, so that the use of secondary data might not provide an exhaustive pool of measurements that the method is actually able to use. Furthermore, all data driven exercises, such as the one in this study, tend to be applicable to the specific area the data is collected from, and therefore are harder to generalize from than other types of studies.

The diagram in Figure 1.1 presents a schema of the traditional research wheel with deductive reasoning shown proceeding from the top to the bottom of the diagram. The review of existing theories in the body of literature, the formulation of research hypotheses, often based on empirical observations, and the actual analysis of data are all part of the traditional line of inquiry used by researchers. However the approach employed in this study can be thought of as beginning from the bottom, at the point where “data collection” enters the process. Here the extensive overview of the current theoretical knowledge linking land use properties with travel behavior and transportation characteristics already provides many relevant factors, but there is no limit to how many variables and factors can be assessed by the methodology used in

this study.

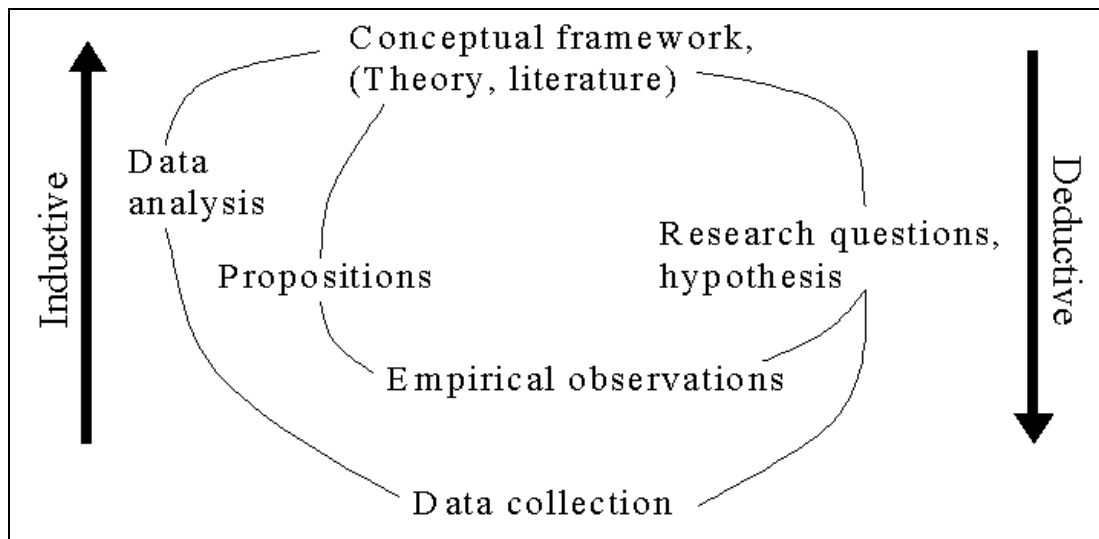


Figure 1.1: Deductive and inductive research approaches. Source: Rudenstam and Newton 2001

In fact, this is typical of inductive methods where researchers are free to search for any number of confounding and moderating variables, no matter how implausible they could be. Simply put, with Bayesian heuristics if a variable is found not to relate to any other variable it will be labeled as such. In fact the problem this study focuses on is that even if a variable, e.g. density, is known to have a particular effect on travel behavior, e.g. higher number of walking trips, it is difficult to establish whether this effect is direct, indirect, or just a strong correlation which hides other possible causal relationships. Also, highly correlated variables do not necessarily imply causation, yet they are often used in current statistical models in terms of independent and dependent variables. The methodology employed addresses this weakness in that it goes beyond mere correlation in finding causal relationships embedded in the data.

Some issues that are still debated and that could benefit from an alternative research approach, relate to whether we should hold VMT and trip frequencies constant, or whether we should imply a cause and effect relationship between job accessibility and trip mode, or any other pair of variables, prior to deciding the design of our statistical model. Deductive studies in these subjects seem to conclude properly that either of the assumptions above are correct. Perhaps this is due to the inability to study concurrently all aspects of travel behavior while controlling for all the factors that influence it, forcing researchers to look at more specific concerns and forcing us to make assumptions that might lead to fallacious conclusions. The literature review that follows in Chapter 2 aims to give the reader an overview of the diverging results obtained so far, and why there is a need for a different methodological approach to research into land use and travel behavior.

Also included in the literature review chapter are some applications of the Bayesian approach, especially as a tool to analyze survey data. The following Chapter 3 introduces the reader to the details of the overall methodology, the data used, how they have been assembled, the types of results that can be generated and how to interpret them. It introduces Baltimore as a typical North American city with a variety of transportation choices and as having a rather variable landscape.

Chapter 4 presents the results from the heuristic search algorithm used to generate Bayesian Belief Networks, expands on the notion of modeling travel mode choice as a binary choice among specific options, highlights the challenge of identifying feedback loops and explains how to build conceptual models of the relationships between land use and travel behavior. This chapter also includes a

robustness analysis for the substitution effects among variables, and an internal and external validity assessment for the variables used in the study.

Finally, Chapter 5 discusses the implications of the research findings in relation to current theories of the relationship between land use and travel mode choice.

## Chapter 2: Literature Review

Research in land use, transportation and transportation mode choice is extensive and covers several disciplines, from social studies to planning, geography, economics and civil engineering. Studies on the relationships between land use and transportation extend from central place theory in urban economics to the anecdotal evidence of real estate advertisements promoting properties near transportation facilities (Polzin 1999:135).

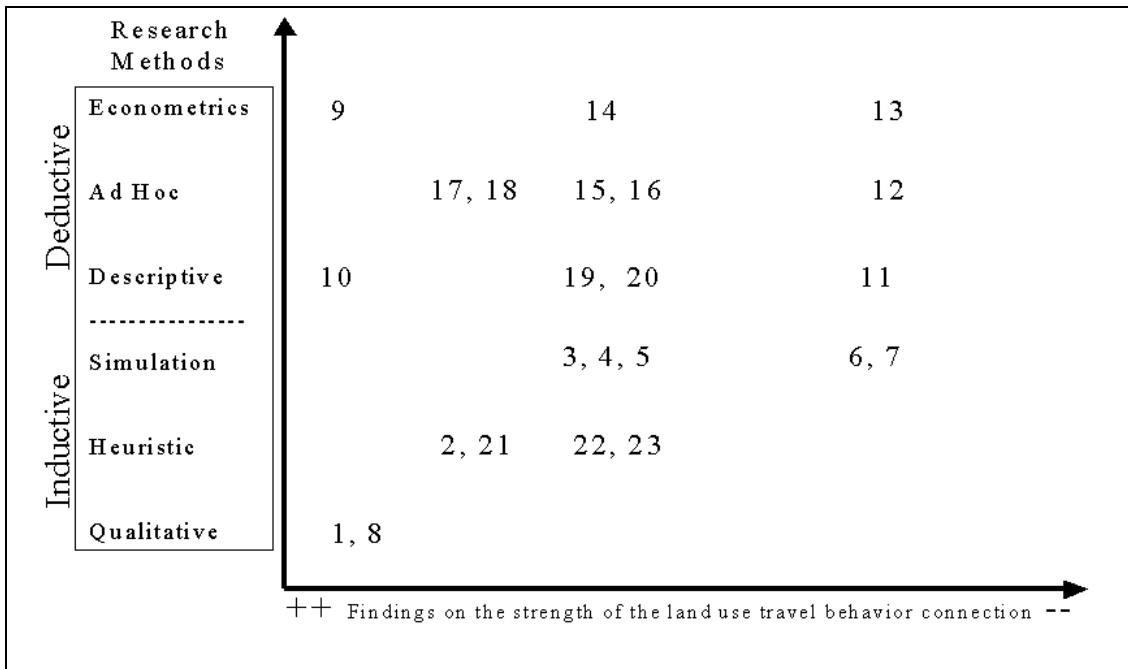
For the purposes of this research the author has reviewed the relevant literature from the planning and transportation domain that spans from the mid 1980s to the present day (2005). This particular time frame is not arbitrary but was selected to include the findings most relevant to a subject that has attracted numerous scientists, engineers and planners since the realization of increased traffic congestion on North American roads in the late 1960s.

The social and environmental impact associated with the growth in travel demand has raised several questions about the sustainability of such trends over the long run. In order to address these concerns researchers have therefore tried to understand what are the root causes of this phenomenon and what are the driving factors guiding human decision processes which ultimately have resulted in current congestion levels and land use patterns. There has been a recent emphasis on encouraging the actual use and implementation of those policies exploiting the possible relationship between urban form, land use patterns and travel behavior. However, as Badoe and Miller (2000:236) warn, what policies to implement and how effective they might be in addressing the concerns mentioned, are yet to be determined.

Such indetermination is due as much to the complexity of the problem as it is to data limitations and methodological weaknesses. This chapter therefore first presents the basic theoretical understanding of what affects travel behavior, and it then reviews the available empirical evidence, not just in relation to the findings, but also with emphasis on what methodologies were used to derive such conclusions.

Table 2.1 summarizes the current preferred research methods, some of the most influential authors who use such methods, and their current position on the strength and magnitude of the relationships between land use and travel behavior. They all share the same available theory of travel behavior presented in Section 2.1 and methods presented in Section 2.1, however not all the authors have found the same evidence that there is a connection between urban form, land use and travel behavior, as introduced in Section 2.3.

Section 2.4 synthesizes the discussion and introduces an alternative methodological solution to the analysis of travel behavior land use relationships, namely Bayesian heuristics, which are then covered in Section 2.5.



- 1) Clifton, K. (2001). Qualitative Methods in Travel Behavior Research.
- 2) Scuderi and Clifton (2006). Bayesian Approaches to Land Use Travel Behavior Relationships.
- 3) Holtzclaw, J. (1994). Using Residential Patterns and Transit to Decrease Auto Dependence and Costs.
- 4) Rabiaga, W. A. et Al. (1994). Shopping Travel Efficiency of Traditional, Neotraditional, and Cul-de-sac Neighborhoods with Clustered and Strip Commercial
- 5) McNally, M. G. (1995). How Neighborhood Design Affects Travel.
- 6) McNally, M. G. et Al. (1993) Comparative Assessment of Travel Characteristics for Neotraditional Designs.
- 7) Rodriguez, D. A., and Joonwon, J. (2004). The relationship between non-motorized mode choice and the local physical environment.
- 8) Waddell, P. (2001). Towards a Behavioural Integration of Land Use and Transportation Modelling. 9<sup>th</sup> International Association for Travel Behaviour Research Conference.
- 9) Greenwald, M. J. (2003). The Road Less Traveled – New Urbanist Inducements to Travel Mode Substitution for Nonwork Trips.
- 10) Cervero, R. (2002). Built environments and mode choice: toward a normative framework.
- 11) Giuliano, G. (1995). The Weakening Transportation-Land Use Connection.
- 12) Gordon, P., and Richardson, H. W. (1997). Are Compact Cities a Desirable Planning Goal?
- 13) Crane, R. (2000). The Impacts of Urban Form on Travel: An Interpretive Review.
- 14) Chatman D. G (2003). How workplace land use affects personal commercial travel and commute mode choice.
- 15) Krizek, K. J. (2003). Residential Relocation and Changes in Urban Travel: Does Neighborhood-Scale Urban Form Matter?
- 16) Kitamura, R., et Al. (1997). Micro-Analysis of Land Use and Travel in Five Neighborhoods in the San Francisco Bay Area.
- 17) Ewing, R., and Cervero, R. (2001). Travel and the Built Environment: A Synthesis.
- 18) Kulkarni, A. A. (1996). The Influence of Land Use and Network Structure on Travel Behavior.
- 19) Taaffe, E. J, and Gauthier, H. L., O'Kelly, M. E. (1996). Geography of Transportation.
- 20) Polzin, S. E. (1999). Transportation/Land-Use Relationship: Public Transit's Impact on Land Use.
- 21) Torres, F. J, and Huber, M. (2003). Learning a Causal Model from Household Survey Data Using a Bayesian Belief Network.
- 22) Arentze, T.A., et Al. (2004). Capturing the Role of Awareness and Information Search Processes on Choice Set Formation in Models of Activity-Travel Behavior.
- 23) Janssens, D., et Al. (2003). Identifying Behavioral Principles Underlying Activity Patterns by means of Bayesian Networks

Table 2.1: Authors, methods and findings on land use travel behavior relationships



## 2.1 Theory of Travel Behavior

The beginning of a cognitive work specifically aimed at the derivation of an organized theory of travel behavior started with Liepmann in 1945, in England.

Liepmann analyzed 1930s data on work travel and was the first to find out about the importance of time, cost and purposes in affecting mode choice. Transportation mode choice received more attention during the 1970s, and the International Association for Travel Behaviour Research emerged with the specific scope of studying mode choice as a behavioral phenomenon.

Mitchell and Rapkin (1954) were the first researchers to identify and explain the possible connection between land use and travel behavior after the authors were able to forecast travel demand as a function of the distribution of population and employment. However it was not until later that Fried et al (1977) published a complete treaty on household travel behavior theory in urban areas. They found that the distribution and location of opportunities for activity has a critical influence on household daily activity patterns and travel demands. Their work evolved from the observation that households need to adapt to these external constraints and try to reduce the imbalance between their current or expected needs and what are the available opportunities at a given location. Household adjustments form a hierarchy from small scale-travel, or activity adjustments, to major changes in residential or work location, and they are related in that higher order adaptations take place only when lower order changes do not sufficiently reduce the imbalance to households needs below some threshold. Fried et al found that the major determinants of activity and travel patterns, as well as of the actual process of adaptation, were: social class position, ethnicity, life cycle status, and residential location. These influence activities

and attitudes, resources, social norms, and perceptions of opportunities which all affect behavior which, in turn, modifies the perception of all these intervening factors in a continuum cycle of adaptation.

The development of such a microtheory of adaptation and change of travel behavior allowed researchers to:

- Model travel behavior based on theoretical principles
- Establish basic systematic hypotheses for empirical research
- Provide the basis for behaviorally informed policy development
- Formulate behavioral criteria relevant for policy assessment
- Establish a framework within which further and more detailed analysis of travel issues could be developed.

Not surprisingly at the same time that this theoretical framework was being developed Hagerstrand introduced his time and space path analysis, often called the time-space prism, which allowed researchers to focus more on the practicalities and constraints of travel behavior and less on the theoretical principles per se.

This emphasis on more practical approaches has led, in more recent times, to the substitution of Fried's undetermined travel behavior thresholds with discrete mode choice frameworks for the study of travel behavior and land use characteristics as well as the implementation of models based on utility functions. In economics, utility is a measure of the happiness or satisfaction gained from a good or service, so that each travel choice made offers a certain value to the individual who seeks to maximize his/her utility. This approach however suffers from the absence of an objective measure of utility when comparing the utility gained from consumption of a particular good by one individual as opposed to another individual. For example,

maximizing utility often means minimizing travel time, but there are also many other unknown factors that could outweigh time. This theoretical approach also implies that humankind is rational, that is, people maximize their utility wherever possible, which is not always the case.

Even so the the concept of utility and individual preferences is the foundation of the neoclassical econometric approach used by many researchers such as Boarnet (2001a), Crane (2000) and Greenwald (2003). The basic theory is that people exercise their preference or choice to live and locate their businesses based, in part, on their proximity to work, other potential destinations and markets for their products and labor in general; this decision in turn affects the resulting travel patterns. An example of how economic theory is applied to the analysis of travel behavior is found in a study by Greenwald (2003), which uses the microeconomic assumption that if someone consumes two or more different goods at the rate that gives them the most satisfaction, the willingness to substitute between those goods is the ratio of the prices between them (2003:43). By the same token, if the goods in question require two different travel times for a given activity one can analyze different urban forms against the change in the ratio of travel times among different transportation means.

However, the simplicity of this top down approach has to contend with the variety of behavioral responses individuals show within environmental conditions. So the econometric theoretical principles, according to Waddell (2001), should be expanded to incorporate not just basic household choices of vehicle ownership, residential and job location, but also their interdependence over time, detailed household decision making processes, housing and labour market conditions, lifestyle choices, living costs, market processes, and household expectations. Both Crane

(2000) and Waddell (2001) highlight that most research in the field has focused only on particular aspects of the overall theoretical principles so that some studies focus on residential densities and travel response, others highlight accessibility constraints, others focus on microscale aspects of urban design and travel choices. Case in point, a recent article by Chatman (2003) points out that until now no studies have ever been carried out specifically on commute mode choice as a function of workplace land use mixes, and Handy (2005a) is only now beginning to look into what portion of travel is made by necessity and what portion is really undertaken as a discretionary choice.

The fact that there are a number of ad hoc studies focusing only on specific interactions between land use characteristics, the environment and travel behavior, could explain why some of the effects predicted by travel behavior theory are not always observed in practice. For example, researchers would expect that dense residential areas with well mixed land uses have a higher frequency of walking trips and fewer auto trips than low density residential areas with exclusionary land use arrangements. This is not always the case, with some studies reporting higher auto trip frequencies (Crane 1996) even in the former case. Basic theory of transportation and land use also suggests that light rail systems will, with time, create economic development and lead to higher residential densities but this is not always the case, as Handy (2005b), and Cervero (1984) have found. These discrepancies help explain the unabated interest by the academic community in continuing to study travel behavior as both a social and an environmental response.

Badoe and Miller (2000) have synthesized a significant number of studies on the travel behavior and land use connection in conceptual frameworks such as the one of Figure 2.1, where travel behavior is shown to be the result of a complex set of

interactions among the various factors discussed in the theory. They assert that land use and physical design (i.e. residential density, employment density and neighborhood design) provide a context for human behavior, which, in this case, includes location decisions (residence, job locations), auto ownership decisions and, ultimately, activity/travel decisions. Therefore, they conclude, increased residential density does not directly “cause” reductions in auto use. Rather, under the right circumstances, it may attract a resident population with particular socioeconomic characteristics and desired activity patterns who will make auto ownership and travel decisions that will result in increased transit/walk usage, reduced auto usage, etc., relative to what they might do in other urban form contexts. They also find that there are many supply and demand or “feedback” interactions within the system of Figure 2.1 (i.e. travel decisions affect road congestion levels, which, in turn, affect travel decisions). Also, it appears that the residential densities combined with attributes of the resident population affect the level of transit service provided, which, in turn, affects the attractiveness of the residential area for people of different walks of life. They also warn “ignoring these complex interactions and analyzing the system in a partial, overly simplified way almost inevitably leads to misleading or even erroneous results” (Badoe Miller 2000:252-253).

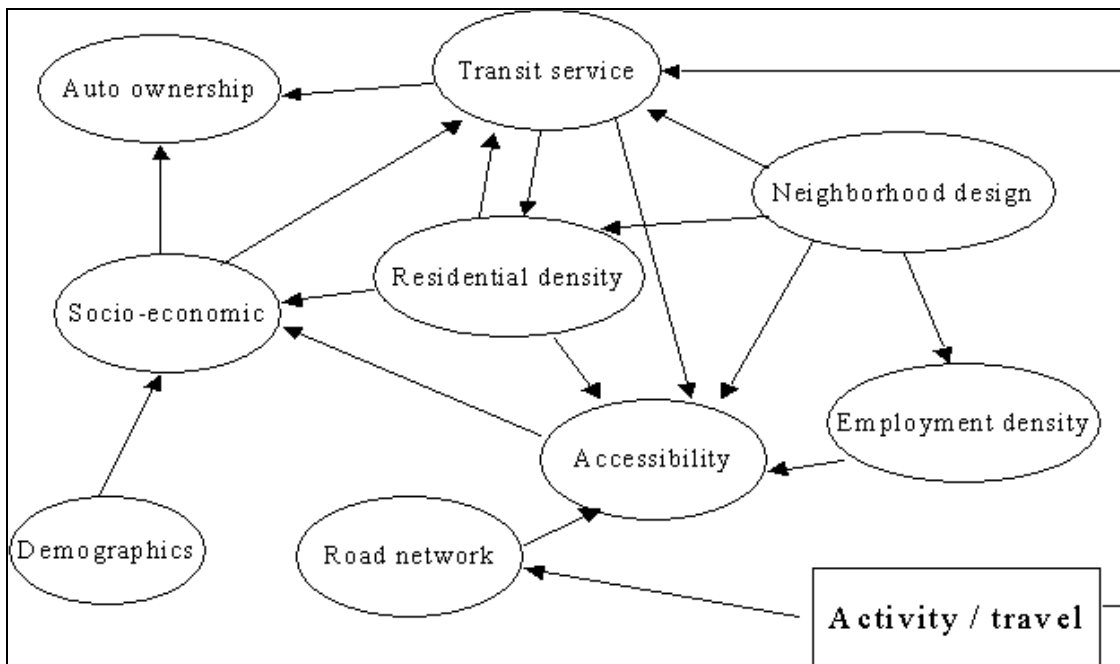


Figure 2.1: Badoe and Miller's complex representation of the travel behavior land use connection

The content of Figure 2.1 is certainly well thought out but its interpretation and, more importantly, its implementation are a matter of choices that each researcher has to face. For example, the National Transit Institute (NTI) defines the relationship between transportation and land use as one of “synergy and interaction” where they “mutually influence each other, under the right conditions in powerful and meaningful ways” (NTI 2000c:8.3). This definition *fits* into the conceptual model above quite well but the real challenge is that nobody can specify what these powerful and meaningful ways are. Furthermore, the NTI says that this relationship is “subtle and complex, ongoing and dynamic” and “it is difficult to trace through relationships in terms of well-defined causal sequences” (NTI 2000c:8.3), just as feedback loops take place over time in Badoe and Miller's conceptual framework.

Taaffe (1996) also, describes the same relationships as a central axiom in urban transport analysis. In his words, travel is considered a derived demand and as

such it is dependent on the frequency and distribution of activities. These in turn are dependent on the distribution of land use and economic activities in a given zone, but all the underlying relationships are assumed. Activities are a very important component of Figure 2.1 but again, there is no unique way to specify how they are dependent on land use and economic activities. In fact Boarnet and Greenwald (2000), in Portland, find a different relationship between land use and transportation, one where land use primarily affects transportation costs rather than trip frequency and distribution, and this too could be easily applied to the model of Figure 2.1.

In another study Boarnet and Crane (2001a, 2001b) try to address how urban design influences travel behavior based on the use of the land made in new planned communities. The authors examine hypothetical studies and descriptive studies which they consider useful to the process of understanding *what is going on*, but there is no intent to explain travel behavior. Even the econometric model they propose, advanced to the point of including travel behavior, unfortunately fails to identify how land use policy can address transportation problems, although the two authors successfully manage to argue their hypothesis with observations based on empirical evidence.

As suggested by Polzin (1999), large gambles derive from “the absence of strong or consistent findings regarding transit's impacts on land use”, which “leaves decision makers at risk of making multiyear, multi-billion dollar investments with only the hope, or weak evidence, that they will produce desired land-use responses”; he advocates searching for a “high degree of confidence in the relationship, even without fully articulated theoretical underpinnings” (p:138) and that “we need to strive to fully and frankly understand the transportation/land-use relationship and then correctly portray this knowledge to decision makers and the public” (p:150).

Even as recently as 2005 Handy (2005b) reiterates that, at least for what concerns the application of smart growth principles, our ability to predict the effects of such policies remains limited. She derives her conclusion from a review of the available evidence, theoretical principles and some of the most comprehensive studies from the past decade. Not only is the focus for these studies quite different but their methodologies differ profoundly as well therefore, before reviewing the most salient literature findings, the next section introduces the most common methodological approaches to the study of travel behavior and land use.

## **2.2 Methodological Approaches to the Study of the Land Use Travel Behavior Relationships**

From the various studies on the subject, there are a number of contrasting views about what links land use and transportation due in part to the application of specific research methodologies to a given theoretical framework. So far this rather extensive body of research has not resulted in a unified and comprehensive meta-theory, nor has it returned a consensus about the meaning of the vast evidence assembled from these efforts. In general, transportation and land use analysis, at both the site-specific and regional level, require some understanding of current and future land use, and how this relates to overall travel demand and patterns. This cause and effect relationship can take place among many variables, can change depending on what instances of this variable are under consideration and also as a function of the geographic scale of inquiry. In all cases we need to know “where people are coming from and where they are going to”, which translates into the need to understand land use patterns and the “causal relationships between land use characteristics and travel” (NTI 2000b:6.3-6.4).



In general the complexity of these methods varies from simply asking for experts' opinions to the use of complex mathematical equations. Once statistical methods are employed the various methodologies can be broadly classified into descriptive studies, multivariate techniques (usually in the form of regression analysis), hypothetical studies and simulation studies. What follows is but a small sample of researchers' applications of these methods to the study of the land use, transportation and travel behavior relationships, organized following the classification developed by Crane in his 2000 critical review of the impacts of urban form on travel (Crane 2000).

### **Basic Non-Statistical and Statistical Approaches**

The very simple approach of comparative analysis has been touted as providing elected officials with results that are easier to comprehend than those derived by studies based on statistical approaches. This method, albeit largely criticized for relying excessively on aggregate data and for using generalized land use characteristics, aims at testing the impact of land use environments on travel demand by looking at matched-pair locations with similar and different land use and urban form. Comparisons are also made of travel patterns for work and non-work related trips in percentages and net vehicle miles traveled (NTI 2000a:9.15-9.18). The results are based on concrete cases but this is probably the only aspect for which comparative analysis can be recommended because these types of studies cannot deal with the direction of causality. "Do roads generate traffic, or does traffic generate roads? Common sense suggests that causality runs in both. If so analyses that treat road supply as cause and VMT as the result can yield misleading results. This problem is known as *simultaneity bias*." (Hansen 1995:17).

National data for major metropolitan areas also indicate that the tendency towards short walk and bicycle trips, for commuters traveling under 1.5 miles, peaks where mixed land uses are more prominent, with 2 to 4 percentage points likely to be added to transit modal share in moderate to high density residential locations (NTI2000a:9.19).

Analytically, the elasticity in land use and travel demand to which Cervero alluded can be measured as the percent change in travel demand over the percent change in land use; as the two dimensional graph of Figure 2.2 shows, an increase in density corresponds to a decrease in trips per capita, and higher reduction rates can be obtained if travel pricing for privately owned vehicles is properly set to reflect all the externalities attributed to them (Hansen 1995:17-18).

$$\text{Travel-Land Use Elasticity} = \frac{\% \Delta \text{Travel Demand}}{\% \Delta \text{Land Use Measure}}$$

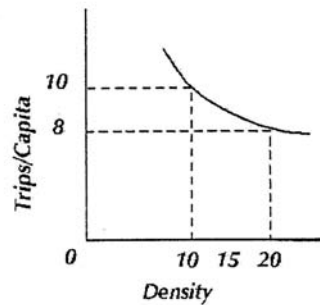


Figure 2.2: Travel - land use elasticity (National Transit Institute 2000)

This represents the percent change in measure of travel demand (e.g. trip rate) for a one percent increase in measure of land use (e.g. density).

## **Descriptive Studies**

Descriptive work provides at most a simple account of travel experiences, individual or averaged, and this simplicity is both an advantage and a disadvantage. For example two neighborhoods might exhibit different travel patterns but such information is rarely sufficient to explain why those patterns are different. At the same time descriptive studies are an extremely important part of the process of understanding what is going on in that they provide a picture of observed behavior and may reveal insights regarding travel patterns in different settings. However their inability to explain behavior limits their utility to direct policy decisions.

The many findings in coordinating transportation and land use derived by multivariate studies can be summarized into three schools of thought: advocates believe that built environments, and how they affect accessibility, strongly influence travel behavior - trip rates, modes of travel, and trip distances. Other authors acknowledge the existence of such relationships but find them to be difficult to isolate and understand, while skeptics support the thesis that land use planning and management is not worth the effort. The latter position is that of Downs and Giuliano, who point out that empirical studies fail to explain how commuters have much longer commutes than many theories predict (Downs 1992, Giuliano 1995a:4). While many factors may help explain these discrepancies, such as the job-housing imbalance, low transportation costs and changing lifestyles, the fact remains that it takes a long time for land use intervention to be reflected in transportation choices and patterns. Quoting research by Downs, it appears that the “greatest percentage of reduction in trip length occurs when moving from low to medium density, while

going from medium to high density yields only a small additional reduction” and that “average densities increase much faster than average trip lengths decrease, meaning that very large increases in density are required to realize significant travel savings” (Giuliano 1995a:6).

As for transportation planning being used to influence land use patterns, Giuliano notes how “although downtown San Francisco benefited from the Bay Area Rapid Transit (BART) system, downtown Oakland and many other local stations along BART lines did not. Despite large subsidies for downtown redevelopment and other supportive policies, BART has had little effect in downtown Oakland. Downtown San Francisco enjoys many amenities that make it a vibrant urban center” (Giuliano 1996:12). Also public transit and rail transit, in particular, are an expensive way to influence development. It has no effects on the underlying forces of decentralization because it “does not influence exclusionary land use policies, local fiscal policy, or the price of owning and using an automobile” (Giuliano 1996:12). The root cause of the problem can be identified in the fact that “because urban areas in the United States are already so accessible, settlement patterns are so well established, and maintenance of privacy is so important transportation plays an ever-decreasing role in the locational decisions of households and businesses” so that “the land use transportation connection is too weak to matter as public policy objective” (Cervero and Landis 1996b:9).

There are a number of researchers who, in light of the difficulties in dealing with such uncertainties, have focused their efforts on what could be considered corollary effects that influence transportation and land use; among the most relevant proponents of such studies are Skinner (1996) and Landis (2001).

Skinner asks if it is possible to determine whether transportation investments influence land use and urban form or alternatively, if these investments are affected by the existing urban landscapes, land development markets, and economic activities. He does not find a specific conclusion and reiterates that “the timeless debate about interaction between transportation and land use continues today” (Skinner 1996:6). This debate is increasingly “less about transportation or land use per se and more about how the combination of the two affects environmental quality, economic growth, and social equity. When viewed in this context, the interaction becomes more important and considerably more complex” (Skinner 1996:6). Given that this *interaction* becomes more complex it is hard to see how it is possible to understand corollary aspects of the land use transportation relationship such as environmental quality, economic growth, and social equity, without first fully understanding the basic relationships. Again, this complexity and inability to understand it provide no guiding principles for policy making.

On a similar note, Polzin suggests studying other aspects of the relationship such as crime, soil conditions and other environmental constraints because the relationship itself cannot be resolved. Again, how can we quantify the effects of these attributes if the relationship is unresolved?

Descriptive studies hardly support the argument that our transportation investments should be supportive of "sustainable communities" and "livable cities", that they should be "neighborhood friendly" and, most important, help stop "sprawl" or that these goals can be achieved by means of “tools” such as "transit friendly design," "transit-oriented developments," "neo-traditional developments," and, of course, "balanced transportation systems"; this is to say everything and very little at

the same time because observations and anecdotal evidence are obviously in disagreement (Polzin 1999:135).

### **Multivariate Analysis**

Multivariate statistical studies examine observed, rather than hypothetical, behavior. In addition, these attempts to explain rather than merely describe what is going on have a more solid theoretical foundation. However these studies still leave unresolved the issue of how to explain the many reasons people have for choosing to travel as they do.

More advanced multivariate analytical techniques, such as ANOVA and linear regression, are used by Cervero, Ewing, Handy, Landis, Gordon and Richardson, among many others.

Gordon and Richardson, in their paper “Are Compact Cities a Desirable Planning Goal?” argue that the reality of American cities is one of “increasingly spread-out metropolitan development” (Gordon and Richardson 1997:95) because open space and agricultural land is not in short supply so that most people decide to live in low-density residential neighborhoods (Gordon and Richardson 1997:96); in these conditions transit is not effective and, contrary to common belief, the consequences of suburbanization and its typical land use are *benign* to transportation. “Suburbanization has been the dominant and successful mechanism for reducing congestion. It has lifted road and highway demand to less congested routes and away from core areas. All of the available recent data from national surveys on self-reported trip lengths and/or durations corroborate this view” (Gordon and Richardson 1997:98). With reference to 1990 NPTS files, Gordon shows that average commuting

times for central city residents in urbanized areas were 18.2 minutes (one-way, all modes), and that urbanized area residents living outside central area cities commuted 20.8 minutes, indicating minimal differences in travel times due to different land uses (Gordon 1997:99). This means that if the “aim is to reduce the environmental damage generated by automobiles, the effective remedy is to directly price and regulate car use, not land use” (Giuliano 1996:12).

However, while it is easy to propose higher and more effective gasoline and road pricing, this is not easily attainable; a 1995 study by the Transit Cooperative Research Program (TCRP report 95) illustrates the difficulties and limited results of road value pricing as the closest option to what is argued by Giuliano; it totally ignores increased gas prices but includes alternative land use plans.

The opposite conclusion on the effects of urban conditions on travel was recently reached by Krizek and Levinson (2003:277), who found that a reduction in VMT can indeed be expected as a result of increased accessibility to goods and services but that “it does not appear that changes in urban form trigger changes in overall mode split”; their longitudinal research followed four hundred households that relocated to neo-traditional neighborhoods in the Puget Sound area. This study extensively captured the physical environment at a large disaggregated level so that it is unlikely that significant variance in accessibility alone could cause changes in transportation mode splits. In a concurrent paper by Krizek (2003c) on pretest, posttest strategies for researching urban form and travel behavior, the author finds that supporters of new urbanism communities should not expect new residents moving into their neighborhoods to change their travel behavior and patterns because

“many individual determinants of travel behavior are firmly embedded in cultural and attitudinal approaches to travel” (p:54).

The author of this research therefore agrees with the views of Cervero (1995a, 1996b), Ewing (1995, 1997), and Black (2003) as well as many other researchers, who argue that land use still matters because it forms a closed-loop system of transport facilities and level of usage that operates in a steady state equilibrium which, when faced with changes to the individual components, will try to settle down to a new equilibrium, even though, as Blunden (1984) puts it, this state of equilibrium may be far from what society regards as optimum (Blunden and Black 1984:1-2).

Cervero strongly argues for these notions and suggests “the land use-transportation connection systems strongly affect urban conditions, including land use patterns, urban densities, and housing prices. Although new transportation investments do not shape urban form by themselves, they still play an important role channeling growth and determining the spatial extent of metropolitan regions by acting in combination with policies such as supportive zoning and government-assisted land assembly. In turn, the characteristics of built environments, such as the size and diversity of neighborhoods and the siting of jobs and housing, significantly influence travel demand” (Cervero and Landis 1996b:9) Cervero refers to this condition as an “elastical” relationship (Cervero and Landis 1996b:11).

In behavioral research logit models have also been used to study mode choices for both work and non-work trips. Cervero, Radisch and Kockleman (Cervero and Radisch 1996, Cervero 1994, Kockleman 1997) studied selected neighbors in San Francisco and other Californian metropolitan areas, and found confirmation that



walk/bike and transit shares are greater where retail uses complement office uses, and that rail transit commute share is greater for higher density residential settings.

Others, Boarnet and Crane (2001a), Boarnet and Greenwald (2000), rely heavily on logit and ordered probit models too. In several instances their studies estimated ordered probit models of trip generation assuming first that land uses fully capture trip costs, and then estimated the same models assuming that trip times have a separate and independent role from land use measures. In general when land use variables have an impact on non-work auto trip generation, it is through the effect on trip prices (speed and distance). When there is no link between land use and trip prices (possibly because land use has been incompletely measured), the model gives no evidence of a link between land use and trip generation.

Despite the above indeterminations, Polzin concludes that “as we place greater and greater importance on the land-use impacts of transportation, specifically transit investments, it is critically important that we fully understand the transportation/land-use relationship”. But “to date, there is no strong, shared perspective of what that relationship is” (Polzin 1999:137), and this is one position the author of this research shares. In fact here is one last contradiction: while Cervero and Radisch (1996) found that nontraditional environments do not significantly alter overall transportation mode shares and instead might increase the number of short trips to be made within and around one such neighborhood, Krizek (2003a, 2003b) and Krizek and Waddell (2002) found that once travel behavior is explicitly modeled there is stronger evidence that a significant shift in transportation mode share does indeed take place.

## **Econometrics**

Waddell (2001) expands on the notion of discretely modeling travel behavior to account for household vehicle ownership, job and residential location, daily activity patterns and various interdependences that take place among these factors over time. In assessing the methodological requirements to carry out such research he claims that traditional multivariate approaches are inadequate for the task. Rather, econometric and simulation approaches are the only feasible alternatives.

When the research aim is specifically to analyze locational and travel behavior, it is common to rely more on economic theory and other utility maximization techniques. In this framework the utility of residential choice is influenced by the degree of accessibility to desired goods and services modeled as a user's total benefit. Such a user is assumed to be a rational actor assessing the advantages and disadvantages of the travel opportunities he/she is offered, the housing location with respect to attractions and opportunities at other destinations, and the cost of transportation in terms of both money and travel time. However discrete choice utility-maximizing models, originally developed for trip and tour data, study travel behavior mostly by extending the pool of choices available to trip makers; this, as claimed by several scholars, does not reflect the true behavioral mechanisms underlying travel decisions, for example when people reason more in terms of "if-then" structures (Janssens 2004).

These types of analyses are further complicated by the intervention of other factors such as lifestyle choices, costs of housing and labor market trends. Waddell (2001:8) reports on a number of different methodological approaches that have been used to confront these research challenges. The most common are several variations

of random utility models which consider the household choice of residential location, job location, vehicle ownership, and daily activity patterns as discrete multi-dimensional choice problems (Waddell 2001:13).

The fundamental assumption in random utility models is that they provide the user with a choice of alternatives. The decision about which alternative is effectively pursued is left to the user so that the results tend to model human behavior rather than rely solely on statistics. Two problems arise however: one, the user will always act as a rational person and invariably select the choice with the maximum utility return, which is not a behavior that is always observed empirically, and two, when too many alternatives arise, the differences in utilities might be so small that the results become unreliable. More limiting is the fact that random utility models might be based on partial or incorrect theoretical formulations that bind all the available choices together (Waddell 2001:16). Moreover, these types of analyses are also computationally complex and are often carried out in terms of simulations. The work by Boarnet and Crane (2001a) falls into this category.

### **Simulation and Modeling**

Simulations assume that the relationships to model are already well understood and are the methodological choice for many Metropolitan Planning Organizations and other institutions focused on developing models (i.e. LUTRAQ, POLIS, CATLAS, ITLUP, MEPLAN and TRANSIM) aimed at predicting future travel demands on roads and transit systems. Unfortunately, all of these models assume increased future travel demand as a function of increased population with land use and zoning residential densities changing accordingly, when in fact they are seldom, if ever, directly modeled. These models often ignore walking, bicycle or transit trips and

hardly take into account feedback effects from land use to transportation. When they do so, it is usually in the form of a pre-set threshold that, once reached, allows for a certain percentage of alternative travel mode choice to be included (NTI 2000b:6.5-6.17). Besides, an estimate of VMT for a particular corridor is not just the result of the road itself but is due to people's decisions to use cars for their transportation needs. It is about people making the decision to travel along a particular direction, a stochastic decision making process that no linear programming model can really explain.

At a smaller scale, studies by MacNally and Ryan (1993), Rabiega and Howe (1994), Stone Foster and Johnson (1992) simulate whether grid-like patterns directly cause a reduction in VMT. These are of limited application because, in general, the number of trips is held constant, and when destinations come closer together a reduction in VMT is obvious but not necessarily true as a function of other factors.

A side note should be made about improvements being carried out in behavioral simulation models. Recent reports indicate success in identifying and modifying behavior for the purpose of improving travel demand. Taylor and Ampt (2003) successfully implemented policies for change in travel behavior. Breda-Vazquez and Riberio-Ramos (2003) studied how land use affects human behavior. Schlich and Axhausen (2003) studied how to actually and properly measure travel behavioral changes. Of interest also is any further development in the research framework proposed by Waddell (2001), in relation to the integration of household choice and decision processes in the context of a larger behavioral framework for integrating land use and transportation models.

In this dissertation, however, travel behavior has a smaller role than what is proposed by Waddell, and it is observed as the human response over the varying conditions of land use and transportation. It does, however, follow on Waddell's suggestion that Bayesian Belief Networks and Artificial Intelligence algorithms, as described in Section 2.6, can provide an "avenue to hybridizing several promising aspects of the preceding frameworks" (Waddell 2001:19).

### **2.3 Evidence on the Land Use Travel Behavior Connection**

In the early 1980s two seminal studies by Cervero, "Suburban Gridlock" (Cervero 1986) and "American Suburban Centers" (Cervero 1989), as well as other publications in the late 1980s, identified the land use connection as an underlying factor contributing to congestion. Research in this area continued with recent works by Boarnet and Crane (2001a), Krizek (2003a, 2003b), Johnston (2003), and others on land use impacts of transportation, and the birth of "new urbanism", "smart growth" and "transit villages" (Cervero 2002). Interestingly though, interest in the land use connection with transportation has had two peaks; the first was in 1995 with a series of papers, among others, by the Committee for Study of Impacts of Highway Capacity Improvements on Air Quality and Energy Consumption (1995), Ewing (1995), Hansen (1995), McNally (1995), Cervero and Seskin (1995), Wegener (1995), Cervero (1995a), Giuliano (1995a and 1995b), which clearly delineated two currents of thought between those who believe and those who do not believe in feedback processes linking transportation to land use and vice versa. More recently, this large amount of research on the subject was comprehensively revisited by Ewing

and Cervero (2001), Badoe and Miller (2000) and Crane (2000), who concluded that most of the reviewed literature was essentially empirically based rather than theoretically sound. Their work is particularly useful in that they present, sometimes in tabular form, the findings of many authors who have found a strong connection between land use characteristics and travel patterns.

For the purposes of this study, the available evidence is organized so as to cover the guiding research questions presented in Chapter 1 about the importance of socioeconomic characteristics vs. land use and environmental variables, the role of density and other land use variables on transportation mode choice, the sensitivity of mode choice to these variables and the appropriate scale of aggregation at which land use information should be analyzed.

### **Socioeconomic and land use characteristics**

In the year 2000, both Crane (2000) and Badoe and Miller (2000) comprehensively reviewed all current models of our understanding of the relationships between urban form, land use and travel demand. In Crane's review, Cervero and Kockleman (1997) found that travel mode choice is mainly a result of socioeconomic characteristics at work. Specifically, Kulkarni (1996) found that income matters the most among socioeconomic characteristics. Both studies include land use mix and some characterization of trip frequencies and vehicle miles traveled. The opposite conclusion is reached by Handy (1996), who finds that the number of travel trips, by all modes, increases as density and proximity to commercial spaces increase. She concludes that land uses influence trip frequency more than socioeconomic characteristics.

Different factors affecting travel mode choice are found by Kitamura et al. (1997) who assert land use measures and socioeconomic characteristics to be less important than independent personal attitudes. Yet Holtzclaw (1994) finds that a doubling of residential density alone can explain a 25% drop in car use, while accessibility to transit can reduce household VMT up to 9%. Handy et al. (1998) tested specific factors that affect strolling as opposed to walking. The type of housing and scenery affects the latter more than the amount of shade or the number of people living in the neighborhood, which were found to facilitate strolling. As Polzin explains “in spite of the overwhelming knowledge of the historical importance of transportation to land development, a great deal of uncertainty remains regarding the nature of the relationship in the modern world. The literature is replete with contradictory evidence regarding both the nature and magnitude of this relationship, and we have yet to successfully model the relationship in a manner that enables us to reasonably forecast land development” (Polzin 1999:135).

It should not be surprising that “most planners trying to deal with the interaction between land use and transportation continue to face one of two problems: either they are confused about the relationship, or they think they are not” (Moore 1996:7). As Moore suggests this confusion is fully understandable given that the literature only “roughly sketches the causal path from some transportation investment or policy (for example, an employer trip-reduction ordinance) through changes in travel behavior (changes in the choice of destination, transportation mode, route, time of travel, and so on) to impacts on other issues which the public wants planners to take care of (for example, economic development, environmental quality, community, land use patterns, and costs). Without such models—without actual measurements of

the key relationships to buttress theories about causal relationships—“confusion should be expected” (Moore 1996:8). Yet at a time when those American cities that depend the most on automotive travel are growing more than other walkable cities, this lack of knowledge is a cause for concern and should not be ignored (Glasser and Shapiro 2003:19).

### **Land use variables and road network characteristics**

In 2002 Cervero published his work on a normative framework for the study of the built environment and mode choice. He reiterated the fact that “in most multinomial mode formulations of mode choice, the systematic components of utility expressions weigh generalized costs of getting between points A and B as well as characteristics of trip-makers” but “rarely do equations account for the influence of points A (origins) and points B (destinations) themselves in explaining mode choice” that is the “potential effects of densities, land-use mixtures, and urban designs at these locations” (Cervero 2002:266). These observations are based on the previous work by both Cervero and Kockelman (1997) which established that density, diversity and design are the three major factors affecting travel demand. All of them deal with how the land is organized and all have received significantly more attention since.

Density or compactness is hypothesized to affect travel demand by shortening trips, inducing more non-motorized trips and encouraging the use of transit and ridesharing. Messenger and Ewing (1996) find that variations in travel mode choice as a function of density can be better explained if considered alongside the effects of car ownership levels. There are some interaction effects that would suggest “it is often what comes with density - mixed land uses, restricted parking, better bus



services, sometimes lower incomes — that lowers VMT per capita” (NTI 2000a:9.5), but there is no consensus among researchers. What is agreed upon is the non-linearity between per capita VMT and density. It appears that going from low to moderate densities might have a more profound impact than changing to very high residential densities. Overall O’Regan and Quigley (1988) found that, *ceteris paribus*, urban form explains two thirds of the variance in the difference in the proportion of households without cars in Boston and Phoenix. At the same time the design associated with these more dense neighborhoods plays an important role in avoiding their perceived shortcomings, such as traffic, overcrowding, etc. Varying the architectural design treatments of rooflines, building types and heights, setbacks, and parking locations, substantially reduce perceived densities. Specific design details such as tree-lined streets and grid-like street patterns might encourage walking trips, although these may be identified as additional trips rather than substituting walking for driving.

Empirical research by Handy (1992:263), comparing local and regional accessibility, finds that density results in more walking and less driving, while Cervero and Radisch (1996) reached a conclusion that favors the substitution hypothesis.

Neighborhood diversity is a direct measure of a variety of uses, housing, working environments, travel options, as well as the social and cultural character of communities. Diversity itself greatly benefits from mixed land uses “such as retail activities near offices generating on-site capture, or child-care facilities near rail stops allowing trip consolidation“ so that density and land use mix complement each other (NTI 2000a:9.10).

Design can be thought of as the assembly of appropriate land uses, the architectural details that are associated with the physical infrastructure (i.e. roads, buildings etc.). Cervero and Duncan (2003) found that land use mix (retail businesses within residential areas) is the strongest urban design factor affecting residents' propensity to make walking trips. Roadway design in particular can affect travel behavior in several ways because a well-connected road network, that is one that provides multiple choices to move from point A to point B, provides better accessibility than a conventional hierarchical road network with a large proportion of dead-end streets (Handy et al 2003).

Also, increased connectivity can reduce vehicle travel by reducing travel distances between destinations, so walking and cycling are relatively direct (Dill 2005). Traffic modeling by Kulash et al (1990) found that a connected road network can potentially reduce neighborhood VMT by 57% compared with conventional designs, although neighborhood travel represents only 5-15% of total vehicle travel, so that total per capita vehicle mileage is likely to decline only 3-10%.

Besides the characteristics of density, diversity and design, another factor characterizing the physical landscape is the degree of accessibility of a given location. For example buildings tend to be located where their use and function are best maximized, which is often a function of the accessibility of a given site. The theory in this respect is very limited. We can expect differences in land use change based on the factors outlined above but the outcomes (in terms of location, types of uses, density, and value impacts) we would expect are nothing more than possible scenarios based on the assumption that:

- “Since access is nearly ubiquitous with car-based transportation, activities tend to be dispersed and segregated”.
- “Compared to rail, impacts on location, intensities, and land values tend to be more diffuse and less easily measurable, whatever clustering and agglomeration that occurs tends to be at major access points, like a rail system or freeway interchanges”.
- “For regional land uses, major roads have brought about more concentrated forms of decentralized growth—e.g., shopping malls instead of neighborhood mom-and-pop stores” (NTI 2000c)

Accessibility to, from, and within a neighborhood or region, until recently considered to have less of an impact compared with land use, has been studied by Handy (1992) and Ewing (1995), who found that “good regional accessibility cuts down on household vehicular travel to a far greater extent than does localized density or mixed use” (Ewing 1995:20). His study is based on aggregated measures of travel and he concludes that “accessibility of residences to a mix of land uses is the key to vehicular travel reduction” (Ewing 1995:21), with the effects of reducing work and non-work average trip lengths and increasing trip chaining. Handy explains how “for the purpose of testing the relationship between spatial structure and travel patterns, accessibility is a more effective measure of spatial structure than either population density or jobs/housing ratios” because it reflects both “the attractiveness of potential destinations and the cost of reaching them” (Handy 1992:255). More recently Clifton and Targa (forthcoming in 2006) found that different levels of accessibility do affect trip generation rates and in particular those made by non-motorized modes.

### **Sensitivity of mode choice to land use variables**

In a recent study Rodriguez (2005) notes how travel mode choice models have been limited in dealing with relationships between the local physical environment and non-motorized modes, even though theory and an increasing body of empirical evidence would suggest that these modes too are very sensitive to the characteristics of the built environment. Studies focused on regional motorized travel have a limited use for decision-making and could be prone to misinterpretation of the results with respect to non-motorized travel modes. More importantly, Rodriguez says “by focusing only on a subset of attributes of the modes examined, such as travel time and cost, utility functions in travel mode choice models can be mis-specified and the significance of estimates misstated” (Rodriguez 2005:169). The use of more comprehensive data from the natural and the built environment is called for in order to gain a better estimation of choice models.

Overall it is unlikely that all natural and built environment characteristics equally affect different transportation mode choices. For example, a steeply inclined road might discourage pedestrian and cycling travel, but it has no or limited effect with respect to the use of public transit or privately owned vehicles. One approach to properly analyzing the varying effects of such variables is to use hierarchical linear models or, more commonly, nested logit models that classify homogeneous travel mode choices against which to test various land use variables. The results show that some mode choices have different sensitivities to the same variables. For example, individuals might value travel timesavings differently between motorized and non-motorized travel modes as Rodriguez reports.

However it is difficult to find this level of sophistication in less recent studies, and meta-analyses of the existing large body of evidence have to contend with an uneven treatment of travel mode choice and socio-environmental variables, which results in a number of discrete findings related to specific travel mode choices. For example, Cervero (1996) does find that commuters are more likely to use transit if they have access to commercial spaces to shop along the way to and from public transit, but what is the implication of having commercial activities for pedestrian mode choice? Theory would also tend to suggest this is an important factor in the personal decision process that leads to walking and it is important to structure the analysis of travel mode choice so that the significance of a variable for specific modes is not undetected because of methodological shortfalls.

### **Scale of aggregation**

In the discussion above care was exercised to report the type and scale of aggregation for the variables used in the various studies. The analysis of trip diaries and the socioeconomic characteristics of individual trip makers are expected to be more accurate than similar analyses carried out based on aggregated or averaged measures. At the same time, land use measures, road characteristics and environmental variables in general are most often aggregated to the Census Block, Census Tract, Neighborhood or Traffic Analysis Zone level, thereby providing a different level of abstraction normally associated with a certain loss of variance in the data. While parcel level data are often considered to be the best unit of analysis for physical characteristics one has to wonder whether such a level of specificity detracts from the

ability to generalize about land use and environmental characteristics in other locations.

A second problem relates to how to match the one to many relationships of individual trip diaries versus aggregated land use measures. Authors such as Frank (2004), Krizek and Waddell (2002), and Song and Knaap (2004) have used aggregated measures of land use based on buffer zones centered around household locations or other individual level records. This approach creates a one to one relationship between individual trip makers and their surrounding physical characteristics, while still allowing a certain generalization of the land use parameters. Yet again, uncertainties exist about the size for such buffer zones; the argument is that it is not yet known whether the coffee shop two miles away from a household has an influence on the household decision making process or not, thereby confounding how large a land use buffer should be.

## **2.4 Concluding Remarks for the Literature Review in Travel Behavior and Land Use**

It is not unusual to reach different conclusions when using different methodological approaches but the authors discussed above differ in their interpretation even when they share the same statistical solutions. Most studies are also only partially covering the full set of theoretical relationships that our understanding of travel behavior and land use characteristics predicts. The decade long debate between Ewing, Gordon and Richardson has been battled out using the same data and with similar statistical techniques. How is this possible? This curious state of affairs is not unique to the study of transportation and land use but common to all complex problems when it is

difficult for a single researcher, or even a team of researchers, to investigate all possible facets of a complex problem.

For example, from the Travel Model Improvement Program (Texas Transportation Institute 2000), we learn that research is needed to improve the understanding of travel as human behavior, and that we should also identify the feasibility and effectiveness of policy actions including: transportation demand management, road and parking pricing, traffic and transit operations, transit fares, telecommuting, and land use controls. The sensitivities of different socioeconomic groups to travel costs and changes in scope and level of transit service should also be better determined, as well as the factors and conditions that affect the amount, timing, and mode of discretionary travel such as land use and development density, urban design, trip chaining, peak spreading, off peak travel, and recreational travel. Not to be excluded from our analysis should be the effects of technology advances (vehicle and highway automation, facsimile machines, and telecommunications) for helping to improve congestion and air quality. Among other factors likely to affect travel behavior that should be considered for inclusion in the travel models are catalog shopping, sidewalk and bikeway improvements, traffic calming, various economic and social factors, and activity patterns of travelers. What a challenge it is to consider all of the above as concurrent processes. Even more challenging is establishing the relationships among the many factors above, and the many more variables needed to capture them, as they play out with each other. Of greatest necessity is the presentation of insightful and informed technical judgment by experienced technicians to decision makers. Bayesian Belief Networks can rise to this task because of their ability to both find and assess an infinite number of relationships at

the same time and to present them in graphical form. It is up to us as researchers to develop the method further so that it becomes more accessible and easier to apply.

## **2.5 Background Notes in Bayesian Belief Networks and Notable**

### **Research in the Field**

The idea that, if uncertain about the accuracy of a given probability, we could use a second probability to provide a measure of confidence on the first can be attributed to the Reverend Thomas Bayes (1702-1761), who, effectively, quantitatively incorporated the inferential nature of frequency based probabilities with the degree of imprecision these probabilities carry with them. This allows researchers to think of probabilities with an explicit dimension of uncertainty. To do so involves the use of many more probabilities whose values are often multiplied and added together, a task that is made easier by the use of Bayesian Networks. These integrate probability calculus and graphical notation.

The causal information encoded in Bayesian networks facilitates the analysis of situations, their consequences, their interaction with observations, and hence helps with the synthesis of plans and strategies when dealing with uncertain conditions (Dean and Wellman 1991, Pearl 1994). Applied to the land use and travel behavior relationships the networks can not only make explicit what these relationships are but can also explicitly provide the strength and confidence such causal relations have.

In biomedical applications, Bayesian Belief Networks (BBN) are well-established tools for decision support systems and predictive modeling or the mining of causal hypotheses (Aliferis et al 2003:1). Bayesian Belief Networks are also increasingly recognized as important representations for modeling relationships at a



finer granularity than standard clustering or regression methods, and as having sound statistical foundations for handling noise, missing data and doing inference (Baldi and Hatfield, 2002).

Bayesian Belief Networks have also been used to analyze survey data. Five attempts in particular have been carried out to use BBN to create graphs of relationships among the many variables of travel surveys: a study by Ramoni and Sebastini “On the use of Bayesian Networks to Analyze Survey Data” (2001) published by Eurostat, a study from the University of Texas, by Torres and Huber (2003), on “Learning a Causal Model from Household Survey Data Using a Bayesian Belief Network”, the works by Janssens et al. (2003, 2004) from the Faculty of Applied Economic Science in Diepenbeek, Belgium, and the study by Arentze and Timmermans (2004) from Eindhoven University of Technology in the Netherlands.

The first research deals with the British Household Survey but not one of the 13 variables of the study relates to transportation. Households and their characteristics comprised the scope of the project.

The study by Torres and Huber adds the analysis of some variables related to trip generation processes so that the results of such research can be used for trip generation inference. This study is of particular interest because the methodology employed is partially deductive, in that the algorithm used required as an input a manually constructed structure of relationships with the possibility that this original network may significantly affect the final networks. In both studies, researchers generated causal networks, or more specifically directed acyclic graphs (dags), by analyzing the collected data and the body of evidence that constitute the variables of interest. Torres and Huber (2003) concluded that income and accessibility to jobs are

not related to travel mode choice, contrary to what many researchers believe to be true, given the wealth of studies linking income to availability of vehicles in households and to trip to work distances.

The two studies by Janssens et al. (2003, 2004) modeled the information of a Dutch travel diary with both traditional fixed decision rules and Bayesian Belief Networks, and concluded that the activity patterns of a separate test database were more accurately forecasted by means of BBNs than by the use of traditional decision rules and discrete model choice methods. Their findings are specific to the factors that condition households, including the resulting travel behavior but also with respect to what other socio-economic factors influence daily patterns of activities. These studies do not include land use information.

The study by Arentze and Timmermans (2004) is an attempt to expand on discrete mode choice modeling by means of Bayesian reasoning. Their focus is also on daily activity patterns and how activity centers are considered in the personal decision making process. This study does not specify land use types but characterizes locations by the presence or absence of activity centers, and uses a synthetic landscape of randomly generated data. As such the study is not representative of any actual behavior but it does explain the methodology in quite some detail.

Although applied to different purposes from the analysis of travel survey data, many other researchers have successfully studied and applied this technique to generate causal networks in their own research fields. Key to the use of such a technique was the introduction by Cooper and Herskovits (1992), in the early 1990s, of an algorithm capable of automatically generating dags, called the K2 algorithm. This was soon followed by more efficient algorithms to generate dags such as those

proposed by Heckerman, Geiger and Chickering (1995), and Friedman and Koller (2000) who tackled the problem of generating dags based on small datasets by means of Monte Carlo Simulation.

Directed acyclic graphs are often created for the purpose of making statistical inference. In the context of Bayesian Belief Networks, inference is the process of using causal graphs for the purpose of predicting a variable given the evidence derived from the data, and allows for the use of BBNs as a traditional modeling tool.

The appropriate use of this type of model is, however, made somewhat difficult by some controversy related to the creation of dags and some difficulties in dealing with feedback effects. It should be mentioned that feedback effects can indeed be modeled by means of potentially cyclic dependency networks, whereby the notion of causality is dropped and only the notion of relationship between nodes is retained. Dependency networks have recently been developed by a Microsoft research group (Heckerman 2000), which designed search algorithms to generate graph-like structures representing the relationship among the variables in the data as relationships between parents and children, or independent and dependent variables. Differences lie in the fact that links are not directed, and that the parents of each variable belong to a structure where the roles of parent and child are not fixed over time. However, the probability component of a dependency network like a Bayesian Network is still a set of conditional probability distributions, one for each node, given its parents (Heckerman et al. 2000:49). Conditional probability distribution can potentially provide information on the strength and magnitude for all the variables in the data.

Another cautionary note on the use of Bayesian Networks refers to the order with which variables are used in the input dataset. Variable Ordering (*VO*) affects the way BBNs are capable of best reflecting the dependence relations in a database of cases; many graphs can be built from the same data depending on the *VO* of their input. Genetic Algorithms (GA) can identify those solutions that best maximize the probability of the graph being correct. This is possible if the researcher is dealing with a research question that has an objective function which an artificial algorithm can maximize, but because this is not always the case the problem of variable ordering is often dealt with in a variety of different ways, mostly specific to the algorithms being used. In particular, VO affects the generation of BBNs by means of the K2 algorithm which is not used in this study. Larrañaga et al. (1996), Hsu (2002) and Guo et al (2002) have covered the Genetic Algorithm approach extensively but it appears that this approach is not quite ready to be fully implemented, not only because of the problem of which maximization function to use, but also because of the many adjustments both research groups have had to perform in order to obtain a plausible result.

Guo also warns about the unproven ability of GA-VO applications to converge to a fixed point under a learning and stochastic data generation process (Guo et al 2002:2).

Appendix A contains a more in depth discussion on the origin of the Bayesian theorem and its application. It also explains how causal networks can be derived by means of probability calculus and mathematical logic, as opposed to the more traditional way of asking experts about what the networks should include and what

relationships should be encoded. In the appendix is also a fully worked example of a simple BBN with its associated inference tables.

While Appendix A provides the necessary technical background to the use of Bayesian statistics, Chapter 3 includes the overall methodology used in the study. Together with a discussion on the data and the steps used to process them, the chapter also explains the heuristic approach employed to derive a representation of the land use and travel behavior relationships. The approach is specific to this study, both in the use of a particular computer algorithm and in its application to a specific problem domain, so it is included in the main body of the research.

## Chapter 3: Research Design

This chapter explains the overall research design, data, and methodology with particular emphasis on the latter, as well as how to properly interpret the results. The aim is to create a relatively simple chronological sequence of steps that produce an unbiased conceptual model of the relationship between land use and transportation, not derived by an interpretation of the existing literature, but as learned from actual measurements and data.

The biggest challenge in conceptualizing travel behavior theory is that it involves human judgments and as such the decision processes that are at the base of transportation mode choice are difficult to identify and model, often requiring expert knowledge by those who attempt to analyze them. For example, researchers are called upon to make a distinction between independent and dependent variables, choose what factors to include, choose particular types of data distributions, and make assumptions about the independence of the data variables.

As a result, a number of attempts have been made to address these issues but, given the current level of empirical evidence, it is always possible to produce ad hoc fitting conceptual models of the relationship between land use and transportation. In fact the number of possible frameworks to be derived is quite large, especially when an extended set of variables is included. One comprehensive review of how these models work is the framework (Figure 3.1) derived by Wegener (1995) where the author considers the urban environment as a dynamic system among transportation networks, travel demands, shifts in population, housing, land use, workplaces, employment, and the availability of public transport.

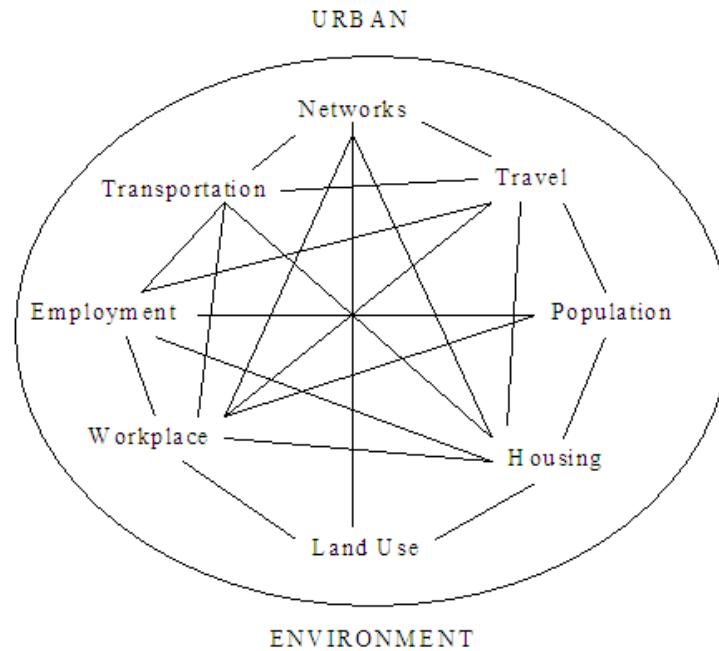


Figure 3.1: Wegener’s “model of urban models”, adapted and simplified form (1995)

Wegener looked at 13 models of urban functions and urban structures, but more have been proposed since the mid 1990s, most notably TRANSIMS (Federal Transit Administration 2003). This great number of different models can be explained by the lack of a precise understanding of the relationship between the dependent and independent variables, their causes and effects, which effectively results in the need to *try* all possible combinations. At the same time a lack of understanding of these very issues explains the variety of methods applied to the same problems by most researchers.

Before embarking on building and testing confirmatory models it is thus necessary to learn about the sometimes weak relationships among the variables at hand. This study aims to recreate the pattern of relationships proposed by Wegener, Badoe and Miller (2000) but with quantifiable causal relationships attached.

It is in such cases that advanced exploratory data techniques, also referred to as “geographical knowledge discovery” (Miller and Han 2001:4), allow the analyst to reduce the number of possible meaningful models that should be tested by first learning the relationships between these variables.

This study, using data from the Baltimore metropolitan area, proposes a new approach to the analysis of the interaction between land use and travel behavior, and specifically travel mode choice, one which does not require the design of statistical models prior to the analysis of the data. It is based on a process of inductive knowledge discovery which uses Bayesian Belief Networks (see Appendix A for a primer), a method that generates directed acyclic graphs representing potential causal dependencies among variables. The variables used in this study, accessibility indices, land uses, and socioeconomic diversity, have all been identified in past studies as factors in this relationship.

The author has identified two possible ways to organize the data: by assembling information based on location – using census tract boundaries or five digit zip boundaries to aggregate socioeconomic variables, travel mode choices and land use characteristics – or by analyzing the data at the individual level, which in this case refers to the use of single records for each possible trip in the data. Associated with single trip records, land use characteristics are always aggregated to the tract or zip code spatial level. Other types of data aggregations, not covered in this study, have been deemed possible with the data at hand, for example trip tour and household level aggregates.



### 3.1 Bayesian Belief Networks

Bayesian Belief Networks are computational objects able to represent joint probability distributions compactly, by means of directed acyclic graphs, which denote dependencies and independencies among variables as well as the conditional probability distributions of each variable, given its parents in the graph (Aliferis et al. 2003:1, Neapolitan 1990). The fundamental axiom of BBNs is the *Markov Condition* that allows for a concise factorization of the joint distribution by means of chains and captures the main characteristic of causation in macroscopic systems, namely that causation is *local* (Glymour and Cooper. 1999).

Let BN be a Bayesian Network over  $U = (A_1, A_2, \dots, A_n)$ . Then the joint probability distribution  $P(U)$  – or chain – is the product of all conditional probabilities specified in BN:

$$P(U) = \prod_i P(A_i | pa(A_i))$$

where  $pa(A_i)$  is the parent set of  $A_i$ .

After giving this specification, the joint probability distribution can be calculated by the product:

$$P(x_1, \dots, x_n) = \prod_i P(x_i | pa_i)$$

Using this product, all probabilistic queries can be found coherently using probability calculus. There are a number of algorithms for probabilistic calculations in Bayesian Networks and the author has selected the one readily available from Chickering et al. (1997), which is explained in the next section.

### 3.1.1 Structure Learning

This section explains how it is possible to learn Bayesian Belief Networks from data. The process depends upon whether the structure of such networks is known and whether the variables are all observable. The structure of the network can be *known* or *unknown*, and the variables can be *observable* or *hidden* in all or some of the records in the data.

Therefore there are four cases of learning BBNs from data: (1) known structure and observable variables, (2) unknown structure and observable variables, (3) known structure and unobservable variables, and (4) unknown structure and unobservable variables. Furthermore the process of learning Bayesian Belief Networks can also be divided between the actual derivation of the network parameters and structure learning. The former is the estimation of the conditional probabilities (dependencies) in the network, while the latter refers to the estimation of the network topology (links).

In this study no records in the data have any missing value because about seven thousand incomplete responses were filtered from the data, and the only unknown is the topology of links. In this case, the algorithm used for determining the network structure is given the set of variables in the model as an input, and it then derives the links between them and estimates the parameters. The algorithms used to solve this type of problem are combinatorially expensive and are usually reduced to the problem of conducting a heuristic search over the space of BN structures.

The two main approaches to structure learning in BBNs are:

- **Constraint based:** performs tests for the conditional independence among the variables in the data, and search for a network that is consistent with the

observed dependencies. Constraint-based approaches usually start with a fully connected graph, and progressively remove the relational links connecting the variables if certain conditional independencies are not supported by the actual data. This has the disadvantage that repeated independence tests lose statistical power, and therefore is used less often.

- **Score based:** Evaluates how well the dependencies in a structure match the data, and searches for a structure that maximizes the score. In the more commonly used search-and-score approach, the main step is a search through the space of all possible directed acyclic graphs, which is intended to return one network or in some cases a set of possible sample networks that represent an approximation of the ideal dependency structure in the data. Unfortunately, the number of possible dags is a function of the number of nodes  $G(n)$ , and it is super exponential with respect to  $n$ . There is no known closed form formula for  $G(n)$  but the first few values for  $n=1,2,\dots,10$  are listed in Table 3.1 below (Glymour and Cooper. 1999).

$G(n)$	Dags
1	1
2	3
3	25
4	543
5	29,281
6	3,781,503
7	$1.1 \times 10^9$
8	$7.8 \times 10^{11}$
9	$1.2 \times 10^{15}$
10	$4.2 \times 10^{18}$

Table 3.1: Number of directed acyclic dags as a function of the number of nodes

The space of Bayesian networks is a combinatorial space, consisting of a super exponential number of structures, so it is not necessarily known that the highest score

found is the absolute maximum possible value. In general, the problem of finding the highest-scoring network structure is NP-hard.

Fortunately however, mathematics and Artificial Intelligence algorithms make it possible to search a combinatorial space with the goal of optimizing the functions returning the network structure. Consequently, the resulting approach is to define a search space, and then do a heuristic search. Therefore a structure-learning algorithm determines the following components:

- i) Scoring function for different candidate network structures.
- ii) The definition of the search space: operators that take one structure and modify it to produce another.
- iii) A search algorithm that does the optimization search.

The scoring function used is the *log-likelihood* function which is simply the log of the likelihood function. That is,

$$l(D | B, \theta_B) = \log L(D | B, \theta_B)$$

The log-likelihood is easier to analyze than the likelihood, because the logarithm turns all the products into sums. Therefore,

$$L(D | B, \theta_B) = \prod_m P(d[m] | B, \theta_B)$$

so that:

$$P(x_1, \dots, x_n) = \prod_i P(x_i | pa_i)$$

The search is resorted to for this optimization problem. A search space is defined as a container with a list of variables and a set of operators denoting the manipulations that can be performed upon these variables.

Because the number of possible networks is exponential to the number of nodes, it is not feasible to examine exhaustively the entire search space; therefore a local search algorithm (e.g., greedy hill climbing) or a global search algorithm (e.g., Markov Chain Monte Carlo) is generally employed. This method effectively searches the space of all possible dags because it makes use of the Markov chain property, which causes the search space dimensionality to be polynomial instead of exponential. This approach is difficult for practical applications requiring the use of more than ten variables.

Selecting the network with the maximum posterior probability as derived by the search and score algorithm is quite a brute force approach to structure learning because of the need to generate and score all possible dags. This approach, however, provides a baseline for comparing the performance of other algorithms used to generate BBNs.

The most basic procedure used for this task is the K2 algorithm, which tries to find the best structure by recursively selecting the best set of parents for each node, independently. This implies that the total ordering of the variables is known, a situation that may not always be true. If the variable ordering is unknown, a search over the most likely orderings is usually more efficient than searching over dags (Friedman and Koller 2000).

More effective than the K2 algorithm, the hill-climbing algorithm searches all points in space and their nearest neighbors, defined as all graphs that can be generated from the current graph by adding, deleting or reversing a single link (Chickering et al. 1997); it then moves to the neighbor that has the highest score and if no neighbors have a higher score than the current point, the algorithm stops. The best practice is

then to restart the procedure at a different point in space  $n$  number of times until the scores converge.

The search and score approach used here is the one developed by Chickering et al. (1997) at Microsoft Research. Similar to the hill-climbing approach, this algorithm adds, deletes, and reverses the possible arcs between the variables but it does so in the context of decision graphs used to represent the relationship among each pair of variables. This algorithm also integrates aspects of the Expected Maximization algorithm, which requires the calculation of the *expected sufficient statistics*<sup>1</sup> for the data. The expected sufficient statistics are then used to ensure the convergence of the results obtained from using a dataset with missing values with the results generated from a complete dataset.

With this technique, the analysis begins with the observation that the local distribution for variable  $X_i$  in a dependency network is the conditional distribution  $p(x_i | X \setminus x_i)$ , which can be estimated by any number of probabilistic classification techniques (or regression techniques, if we were to consider continuous variables) such as generalized linear models, neural networks, probabilistic support-vector machine, or embedded regression/classification models (Heckerman et al. 2000:55). The chosen method in this case is a probabilistic decision tree, where for each variable  $X_i$  in domain  $X$ , the classification algorithm independently estimates its local distribution from data. Once all estimates for the local distributions are obtained,

---

<sup>1</sup> In statistics, one often considers a family of probability distributions for a random variable  $X$  (and  $X$  is often a vector whose components are scalar-valued random variables, frequently independent) parameterized by a scalar- or vector-valued parameter, which let us call  $\theta$ . A quantity  $T(X)$  that depends on the (observable) random variable  $X$  but not on the (unobservable) parameter  $\theta$  is called a statistic. Sir Ronald Fisher tried to make precise the intuitive idea that a statistic may capture all of the information in  $X$  that is relevant to the estimation of  $\theta$ . A statistic that does that is called a sufficient statistic.

the structure of the Bayesian network can be constructed from the (in)dependencies encoded in these estimates (Heckerman et al. 2000:55). Each variable is modeled as a multinomial distribution and the learned decision tree corresponds to the Bayesian network.

The algorithm searches each row of data for unique combinations of categorical data. Each unique combination is called a “case” and it forms the basis of the following analytical steps, where the algorithm greedily grows decision-trees using the Bayesian scoring criterion. This is a greedy algorithm that combines a global search over the structure’s relational links with a local search over all of the nodes in the decision graphs. It begins with one node (variable) and evaluates its relationship to the other nodes (variables) by means of decision trees; then, it scores the corresponding Bayesian structure based on its posterior probability (of such a network) considering the given cases. The procedure is as follows:

- 1 Score a generic network structure. For each node  $x$  (variable) in the graph:
- 2 Add every non-descendant that is not a parent of  $x$  to the parent set
- 3 For every possible operator  $O$  in the graph:
  - i. Apply  $O$  to  $BS$
  - ii. Score the resulting structure
  - iii. Un-apply  $O$
- 4 Remove any parent that was added to  $x$  in step 3
- 5 If the best score from step ii. is better than the current score
  - Let  $O$  be the operator that resulted in the best score
  - If  $O$  is a split operator (either complete or binary) on a node  $x$  that is not in its set of parents then add a new node to the parent set
  - Apply  $O$  to  $BS$
  - Go to 1
- 6 Otherwise, return  $BS$

Three operators ( $O$ ) are allowed:

- Complete split adds a child node to a set of parents.

- Binary split adds two children to a set of parents.
- Merge split combines two or more children in a single new node inheriting all of their parent nodes.

To learn a decision tree structure for  $X_i$ , the search algorithm is initialized with a single root node having no children. Then, each leaf node is replaced with a binary split on some variable  $X_j$  in  $X \setminus X_i$ , until no such replacement increases the score of the tree. The *binary split* on  $X_j$  is a decision tree node with two children: one of the children corresponds to a particular value of  $X_j$ , and the other child corresponds to *all other* values of  $X_j$  (Chickering et al.1997).

As well as the Bayesian Belief Networks, this algorithm generates two other types of output associated with each dag: decision trees and conditional probability tables or distributions (CPT or CPD). Decision trees can be thought of as encoding the variable values through which significant relationships change their nature; each variable can be assessed for its moderating and mediating influence on all other variables and queried to provide quantitative assessments of how they influence travel mode choice. Conditional probability tables instead capture the sign and strength of relationships and present them in easy to read tabular form.

The researcher needs to read interactively all three types of outputs to interpret the overall results. For each variable it is necessary to look at the parent set of variables in the dags, the role that each of the parents has in the decision tree, and the sign and strength of their relationships with other variables and, in this case, with travel mode choice. The importance of this step cannot be stressed enough. In fact there is a field of enquiry entirely devoted to improving the understanding of Bayesian Networks. Work done at the University of Pittsburgh by Druzdzel (1996)



identified two main areas crucial to the proper interpretation of dags, the assumption made with the methodology and, more specifically, the explanations of the reasoning involved, which should focus on describing how conclusions can be extracted based on the assumptions coded in the data and the observed evidence. His conclusion is that efforts in explaining probabilistic networks should focus not on their entirety but on the most relevant parts of the graphical outputs and the translation of numerical probabilities into straightforward phrases. His advice is followed in this study.

### **3.2 Methodology**

The derivation and scoring of possible BBNs representing the land use and travel behavior relationships do not rest entirely on the application of the above algorithm to survey data and the interpretation of its outputs; this is the core of the method but the overall methodology includes many more steps and procedures.

In an attempt to illustrate the usefulness of BBNs for the study of land use and travel behavior, this study has employed both advanced statistical techniques and Geographic Information Systems, used together in a methodology that is based on the:

- 1) Collection of land use, travel and socioeconomic parameters in a database with discrete records or aggregated values at the tract and five digit ZIP code areas. The study uses the 2001 National Household Travel Survey Data for the Baltimore Metropolitan Area and information from the Maryland Property and Transit View databases.
- 2) Use of meta-heuristic search engines to inductively learn a graph-like structure of the causal relationships (or dags) between land use and travel

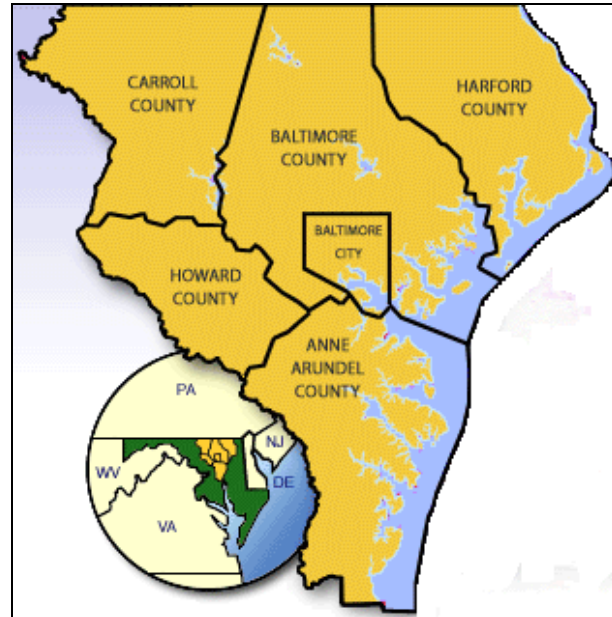
mode choice (auto, walking, and transit). The search engine uses Bayesian scoring metrics to identify the structure that most closely represents the information available in the data. Other types of outputs obtained are decision trees and conditional probability distributions. The former are used to establish the role of variables and quantify their influence on travel mode choice, while conditional probability distributions are used to determine the strength and direction of the relationships among significant variables in the study.

- 3) Comprehensive interpretation of the resulting relationships graphs, decision trees and conditional probability distributions. This is an interactive step that starts with the selection of parent-child pairs of variables from the network of relationships, together with their local dags. For each pair, their decision trees and conditional probability distribution are interpreted against existing theoretical tenets and empirical evidence.
- 4) Use of map overlap and spatial queries to isolate those areas where qualitative, rather than quantitative, characteristics make a difference in terms of human response.

Step three is repeated many times before a clear picture emerges from this methodology, and the results presented in Chapters 4 and 5 provide as rigorous as possible an interpretation of the dags, decision trees and conditional probability tables obtained using national and local NHTS data.

### ***3.2.1 Study Area and Data***

The study area covers the Baltimore metropolitan region, which includes the counties of Carroll, Howard, Anne Arundel, Baltimore, Harford and Baltimore City. These, as shown in Map 3.1, fall under the jurisdiction of the Baltimore Metropolitan Council.



Map 3.1: Study area: Baltimore Metropolitan Area (source: [www. baltometro.org](http://www.baltometro.org))

The city of Baltimore and its hinterland were chosen because they all share a set of interesting characteristics, in particular:

- The city of Baltimore is not a metropolis, nor a small city. With a resident population of about 650,000 people, the city ranks 18<sup>th</sup> among the biggest 50<sup>th</sup> cities in America as of 2004, and is comparable in population with Memphis TE, Forth Worth and Austin TX, and Charlotte NC.
- The city has a variety of public transportation options that very few other American cities have, such as both subway systems and surface trolleys, extensive bus routes and even water based public ferries. This variety of transportation means is important when studying travel behavior and mode choice.

- The area under study has a dense urban core represented by the Central Business District in Baltimore city and an almost continuous variety of other land use combinations, large industrial parks, waterfront areas, rural and semi rural areas, greenbelts, satellite towns, as well as a diverse topography.

Detailed data for households and individuals were obtained from the 2001 National Household Travel Survey (NHTS), which included the additional households surveyed in the Baltimore area. The survey is sponsored by the Federal Highway Administration (FHWA), the Bureau of Transportation Statistics (BTS), and carried out by the U.S. Department of Transportation by means of computer assisted telephone interviews (CATI). The survey provides a national sample of daily and long-distance travel and it includes demographic characteristics of households, people, vehicles, and detailed information on daily and longer-distance travel for all purposes by all modes. NHTS survey data are collected from U.S. households and expanded to provide national estimates of trips, trip purpose, and a host of household attributes. This study integrates the NHTS survey with local profiles of typical land use patterns, and road density for each tract and zip code in the study areas. Detailed information on the NHTS survey is provided by the Oak Ridge National Laboratories both on line at [http://nhts.ornl.gov/2001/html\\_files/introduction.shtml](http://nhts.ornl.gov/2001/html_files/introduction.shtml), and in print.

Information about the National Household Travel Survey Add-On can be found in the Baltimore Regional Transportation Board publication “Comprehensive report on the NHTS data”, which explains the methods used to carry out the local survey, and how the data were processed (Baltimore Regional Transportation Board 2002). Also see <http://nhts.ornl.gov/2001/usersguide/index.shtml>.

The variables from all datasets have been reclassified into categories that characterize each trip reported in the data. In many cases, the number of classes within each variable has been reduced to simplify the analysis. For example, the variable “age” for the respondents has been aggregated to four classes, with an important separation for teenagers at sixteen years of age to reflect the possibility of acquiring a driver’s license. The response variable “race” has been reclassified into four categories. More importantly, the race of the respondent has also been assigned to the remaining members of the family, an assumption that might not always hold true. The personal income variable was derived from assuming equal access to the household income resources. Transportation mode choices have been reduced to just three classes: private vehicle, walking and public transit. Private vehicle trips include the use of private cars, trucks, motorcycles, etc. Walking trips include bicycling, wheelchair mobility, jogging and any other non-motorized trip. The choice for transit includes all public transportation systems except for ferry and water taxi which, given their limited presence in the data, have not been analyzed in this study. Table 3.2 has information for all the variables used in the study and their detailed classification.

The land use variables were derived from the Maryland Property View Data; in particular, the 1997 Land Use/Land Cover GIS information layer was used, updated to the year 2000. Each land use polygon was assigned to a zip or a tract and its boundaries reshaped to fit into such administrative units. Based on the total area of each administrative boundary, land use variables were then calculated as a percentage of the total area and then reclassified into ten discrete classes of land use covers for each type of residential, commercial or other land use. Trip origins were then associated with these specific land uses, although a more comprehensive approach

would have been to distinguish between land uses at the origin and destination of each trip; in this instance, however, the data requirements for the algorithm would have escalated. The road network, sourced from the publically available Tiger Files, was subject to similar processing, where each road segment was assigned to a tract or zip and its length recalculated accordingly. A discrete ratio of the total road length within each administrative unit over the total area for such units created an index of road density, which was also reclassified into ten discrete categories.

Of notable absence among the variables used for this study are derived measures, such as land use mix or entropy, internal and external tract connectivity, gravity measures and other, more specific, measures of urban form. More notable is the absence of vehicles miles traveled (VMT). This is a very important variable for studies of travel mode choice, one that must be controlled for trip frequency and trip costs, including gasoline prices, all variables, which are not included in the study at this stage. Furthermore, the individual being the unit of analysis of choice, vehicle miles would have had to be sorted by individuals rather than by vehicles, adding another layer of abstraction beside the classification of continuous variables into discrete classes. Given the experimental nature of this study, the author purposely limited the complexity of the information to be analyzed so as to ensure the tractability of the input dataset by the algorithm of Chickering et al. (1997). However, given the successful results obtained so far there appears to be no reason to include more derived measures in future studies.

Land use variables are available as continuous percentage values but the decision was made to classify them into discrete categories, as for the other NHTS data. From a methodological point of view however, BBN algorithms can deal with

both discrete and continuous variables; their implementation is considerably more complex but the confidence in the resulting belief networks, depending on how dense are the continuous data, can be even higher. The real drawback of using continuous data as input for the derivation of BBNs is that it is not always possible to compute conditional probability distributions without further processing, such as when using WinMine 2.0 (Chickering 1997) as is the case here.

The NHTS data can be organized in at least five different ways for analysis with Bayesian Belief Networks. Each data framework has its own advantages and disadvantages as summarized below. In the analysis presented here, individual trip records were used as the unit of analysis for the transportation data.

- Individual trip records allow for the maximum number of cases that the search score can use to generate the most compelling networks. In the case of this study, 22,000 trip records were used to generate a model linking land use variables, socioeconomic factors and other variables to transportation mode choice. The drawbacks of a trip-level analysis are found to be the inability to analyze total numbers of trips by mode, and the loss of information on the individual responses or household interactions.
- Tracts or Zip codes could also be used as the basic unit of analysis. For the study area here, there are just over 600 tracts and just over 150 zip code areas covering the six counties. With this data structure, the characteristics of each spatial area could be summarized and the transportation mode choices could be analyzed in terms of overall aggregation of trips made by each mode. For the technically inclined, this data structure is the transpose of the case above, and although it results in a considerably lower number of records, it could be considered a more

geographically based approach. However, the number of trips in any given census tract may be limited due to the sampling structure of the NHTS, and insufficient to yield robust results. Actual census tables could be used to obtain a more uniform distribution of records in each tract, but the analysis would be limited to the information provided by the journey to work questions.

- Households could form the basis for analysis. For the Baltimore Add-On, there are approximately 5,000 household records to analyze for the entire area of interest. To base the analysis on households has the advantage of examining the full array of trips (or trips by specific modes) made at the household level, which may be the preferred decision making unit as many travel outcomes are the result of household responsibilities. Community-level data on land use and transportation characteristics could be included and aggregated at the appropriate spatial unit. The disadvantage of households as the unit of analysis is the loss of individual autonomy in decision-making and thus the role of individual circumstance, resources and constraints.
- Similarly, individuals' travel choices could be the subject of examination, with attention to those individual characteristics that influence mode choices or travel demand. This would yield considerably more records to analyze using the Baltimore NHTS data, with 7,825 individual records. The analysis of individual trip makers allows for exploration of personal situations, such as work location; however, it may miss some of the household-level interdependencies that may exist.
- Finally, trip tours could be constructed and analyzed to understand the interdependencies that occur between a sequence of trips and their relation to



personal, household and land use characteristics. Considerable effort would be required to construct trip tours but this remains a very promising and relatively new area of investigation.

### **3.2.2 Model Runs**

For this analysis the Baltimore Metropolitan Area offers an interesting mix of land use and transportation patterns, urban, semi-urban and rural conditions. People's transport choices, as derived from the 2000 National Household Travel Survey Add-On for Baltimore, are considered as the human response to transportation availability and land use conditions, parameters which in turn are studied at two different geographical scales. The same dataset provides information on the socioeconomic characteristics of all households and individuals included in the survey, while land use and transportation characteristics were derived from the Maryland Property View and Maryland Land Use datasets.

Only three different aggregation schemes that include both geographic scales, and individual level trip characteristics are investigated in this study. The three resulting Bayesian Belief Networks of land use and travel behavior relationships derived as part of this research are:

Model 1: Bayesian Belief Networks generated from individual level trips. The eleven socioeconomic characteristics of the person taking the trip are those of the household he/she belongs to, while the land use and transportation variables are aggregated at the Census Tract level. This data structure is the one with the highest number of

records, almost 20,000. Outputs for this model run, besides BBN, include binary decision trees and Conditional Probability Tables.

Model 2: Bayesian Belief Networks generated from land use, transportation, personal socioeconomic characteristics and trip counts, all aggregated and tallied at the Census Tract level. This dataset is limited to 615 records, one for each tract covering the area under study. As for Model 1, besides BBNs, binary decision trees and Conditional Probability Tables were also computed for this model run.

Model 3: Bayesian Belief Networks generated from land use, transportation, personal socioeconomic characteristics and trip counts, all aggregated and tallied at the five digit Zip level code. This dataset is limited to just 150 records, one for each zip area covering the study area. As for Models 1 and 2, besides BBN, binary decision trees and conditional probability tables were derived also.

### 3.2.3 Measurements

In this study the information and variables provided by the NHTS efforts, the Maryland Property View and transit View data, and the US. Census Tiger files, were categorized as follows:

Variables and variable name in BBN	Variable Description and categories	Equivalent variables in NHTS and MD Property View	Scale of aggregation for model 1
Driver Status (DRIVER)	Driver status of respondent, licensed, not licensed or not appropriate	DRIVER	Discrete
Worker (WORKER)	Worker status of respondent, working non working or not appropriate	WORKER	Discrete
Age (Age)	Age of respondent, 0 to 5 years, 6 to 16 years, 17 to 65 years, 66 years and more	R_AGE	Discrete
Vehicle Count (vcount)	Number of vehicles in household, 1 to 9 or more	HHVEHCNT	Discrete
Household Size (Hhsize)	Household size, 1 to 9 or more	HHSIZE	Discrete
Race (Race)	Race of the head of household: White, African American, Hispanic, Other (including Asian)	HHRACE01	Discrete
Driver Count (dcount)	Number of drivers in household, 1 to 6 or more	DRVRCNT	Discrete
Personal Income (Income)	Household income divided by number of persons in household. Eleven classes from \$0 to \$100,000 or more	HHFAMINC	Discrete
Transportation Mode Choice (modechoice)	Transportation mode choice among motorized options, walking and biking, or transit (including buses, metro and rail)	TRPTRANS	Discrete
% Low Residential	Low-density residential – Detached single-family/duplex dwelling units, yards and associated areas. Areas of more than 90 percent single-family/duplex dwelling units, with lot sizes of less than five acres but at least one-half acre (.2 dwelling units/acre to 2 dwelling units/acre)	_11	Census Tract

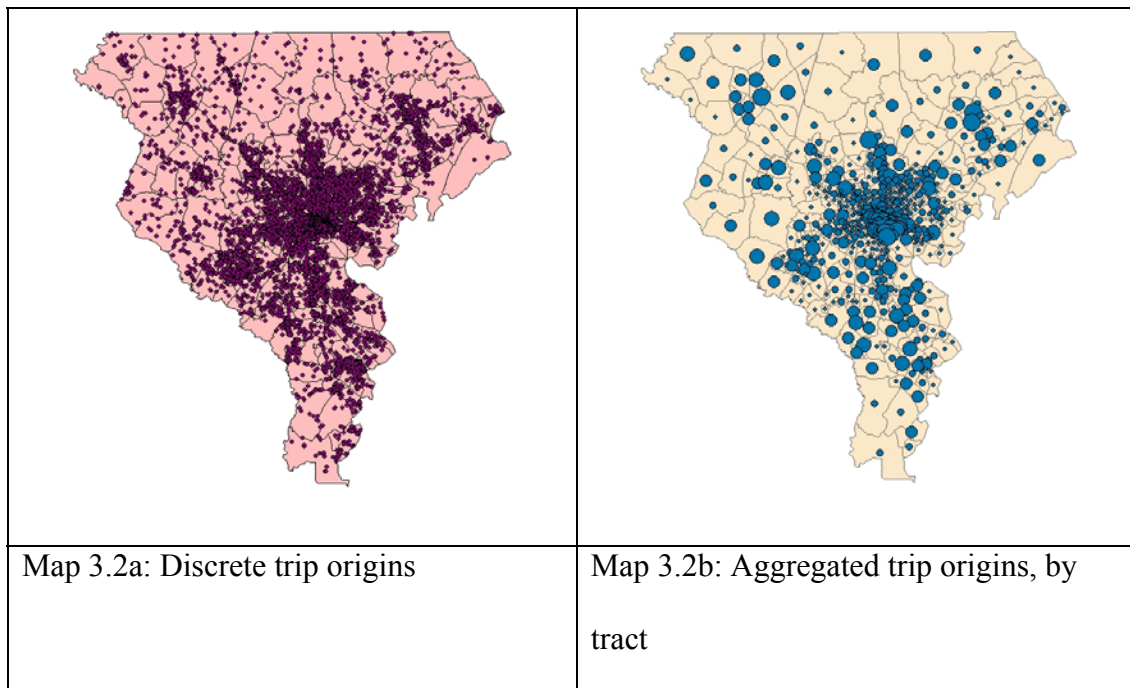
Table 3.2 Variables and data classification

% Medium Residential	Medium-density residential - Detached single-family/duplex, attached single-unit row housing, yards, and associated areas. Areas of more than 90 percent single-family/duplex units and attached single-unit row housing, with lot sizes of less than one-half acre but at least one-eighth acre (2 dwelling units/acre to 8 dwelling units/acre)	_12	Census Tract
% High Residential	High-density residential - Attached single-unit row housing, garden apartments, high-rise apartments/condominiums, mobile home and trailer parks. Areas of more than 90 percent high-density residential units, with more than 8 dwelling units per acre	_13	Census Tract
% Commercial	Commercial, retail and wholesale services. Areas used primarily for the sale of products and services, including associated yards and parking areas	_14	Census Tract
% Industrial	Manufacturing and industrial parks, including associated warehouses, storage yards, research laboratories, and parking areas	_15	Census Tract
% Vacant	Vacant land, such as water bodies and parks	_18	Census Tract
% Other	Other land uses	All other codes	Census Tract
% Institutional	Institutional land uses	_16	Census Tract
Road Density Index	Linear road length over square km of area, ten classes, with a lower value indicating more dense road networks		Census Tract
Residential Accessibility to Transit (AR)	Proportion of a tract or zip within 402 m. (0.25 miles) of a transit stop		Census Tract
Overall tract accessibility to transit (RR)	Proportion of residential areas in a tract or zip within 0.25 miles of a transit stop		Census Tract
Commercial accessibility to transit (CR)	Proportion of commercial areas in a tract or zip within 402 m. (0.25 miles) of a transit stop		Census Tract
Gender (R_sex)	Male, female	R_SEX	Discrete
Education (Education)	Educational attainment on an eight-class scale from no high school to graduate school.	EDUC	Discrete
Trip purpose (trppurpose)	Work and non-work	WHYTRP	Discrete
Trip length (trplength)	One mile or less, up to three miles, three to ten miles, more than 10 miles	TRPMILES	Discrete
Children in Household (kidsinhh)	Number of children under the age of 15 present in each household	Derived from household size and age variables	Discrete

Table 3.2 (Continued) Variables and data classification

The transportation mode choice variable is the one that the study focuses on with particular emphasis. For each entry in this variable a trip is made originating somewhere in the study area so that the local land use characteristics can in fact be associated with such a trip. Map 3.2a shows the spatial distributions of such trips and it shows how some census tracts have fewer trips originating from them, especially those tracts in the rural fringes of the study area.

If the same trips are aggregated into pies sized proportionally to the number of trips originated from each tract, Map 3.2b is obtained. Here it can be seen that large pies representing tracts generating 200 to 250 trips are more evenly distributed and do in fact cover a whole variety of areas, from downtown Baltimore to the valleys of Baltimore County, to rural Howard County and all other areas characterizing the Baltimore metropolitan region. It is such tracts that provide the most reliable evidence for the successful execution of the learning algorithm.



In conclusion the methodology for this study focuses on how to organize data so that they can be used as the input for a heuristic search algorithm. It also covers the logic behind the algorithm's data handling. The scope is to find what relationships are hidden in the data, and Chapter 4 presents the search results for the Baltimore metropolitan area.

## **Chapter 4: Findings**

As outlined in Chapter 2, significant questions remain about the land use travel behavior relationships and their interdependencies. A variety of approaches and data sources have been applied to this problem with varying results, and often conflicting findings. This chapter presents the expected results from using Bayesian Belief Networks, coupled with automatic learning; as seen earlier, these are in the forms of dependency networks, decision trees and probability distributions. This methodology was successfully applied to the analysis of land use and travel behavior and the three types of outputs discussed in this section provide new insight into the subject. In fact this study has also generated relevant information pertaining specifically to the Bayesian heuristic approach when applied to the analysis of spatial data; in the future such information might be used to refine and improve such methodology. More importantly the land use and travel behavior relationships presented here are derived analytically and without the specification of a priori statistical or conceptual models.

### **4.1 Cross-Tabulations and Chi Square Statistics**

Basic descriptive statistics show that the total number of trip records derived from the National Household Travel Survey is just over 19,000, a number that excludes all the records for which respondents did not know what to answer, or chose not to do so. Unfortunately, when performing the same analysis by aggregating all the trip information to the tract and five digit zip levels, the total number of records available to derive Bayesian Belief Networks precipitously drops to 615 census tracts and 150 zip areas. Given that the heuristic algorithms used in this study are susceptible to the

amount of data used for the analysis, the author has tested the data for its robustness and the outcomes are presented later, in section 4.7 Model 1 Stability Tests.

Because of the categorical nature of the data as discussed in Section 3.2.1, Pearson’s Correlation coefficients, although easy to derive, do not have any meaning in this instance. The appropriate measure of association for this dataset is in fact given by cross tabulation frequencies which were derived between each variable and mode choice. Chi Square tests and Cramer’s V tests can also be used to statistically evaluate such associations.

The cross tabulation tables among all variables and transportation mode choices provide information about private car trips, walking trips and transit trips in relation to particular instances of other variables. For example, in Tables 4.1 and 4.2 only 24% of the trips were ten or more miles in distance and they were largely undertaken with private vehicles (22%). For trips of less than three miles, 46% of the total, car is still the prevailing mode choice but transit has a significant share at about 10%.

Trip distance	Percent	Cumulative %
<1 mile	28.24	28.24
1-3 miles	18.31	46.55
3-10 miles	29.13	75.68
10 > miles	24.32	100

Table 4.1: Frequency distribution for trip distance

Trip distance by mode choice	Car	Walking	Transit	Total
<1 mile	18.68	1.81	7.75	28.24
1-3 miles	15.83	0.92	1.56	18.31
3-10 miles	26.03	1.39	1.71	29.13
10 > miles	22.00	1.08	1.24	24.32
Total	82.55	5.20	12.25	100

Table 4.2: Cross tabulation of trip distance by trip mode



Of interest also is the fact that most trips, more than 84%, were for purposes other than the daily journey to work, and the large majority of these were carried out with no children as passengers. Only 8% of all trips were made by people lacking access to a private car, while people in numerically larger households did not show any greater tendency to choose the vehicular mode than those in smaller households. With respect to income, the percentage of car-based trips for low-income households is 87%, only 3 percent more for the highest income class, a small difference indeed.

Also interesting is the fact that more trips by car take place as the percentage of low density residential increases. As seen in Table 4.4 below, in a tract with just 10% low residential, 71% of the trips are by private vehicle vs. 5% walking trips and 11% transit trips.

Mode choice	Percent	Cumulative %
Car	82.55	82.55
Walk	5.20	87.75
Transit	12.25	100

Table 4.3: Frequency distribution for mode choice

Mode choice by Low residential density	Car	Walk	Transit	Total
10 %	71.38	5.01	11.60	87.98
20%	6.83	0.13	0.43	7.38
30%	2.29	0.04	0.13	2.46
40%	1.50	0.02	0.06	1.57
50%	0.06	0.00	0.01	0.16
60%	0.09	0.01	0.01	0.11
70%	0.40	0.01	0.03	0.44

Table 4.4: Cross tabulation of percentages of low residential densities with mode choice

Yet for the equivalent 10% of observed land use in medium and high residential densities, car trip frequencies are only 49.84 and 60.13% of all trips, so residential density does appear to have an influence on transportation mode choice, but the

direction of this relationship is undetermined because as density increases car usage first drops then increases again.

For commercial land use Table 4.5 shows a more definitive situation, where most of the car based trips are generated where this type of land use is a minimal proportion of the overall land use in the originating tract.

Mode choice by Commercial density	Car	Walk	Transit	Total
10 %	66.21	3.54	8.31	78.06
20%	9.41	0.71	1.66	11.78
30%	3.35	0.24	0.47	4.06
40%	1.60	0.37	0.80	2.78
50%	0.80	0.17	0.38	1.36
60%	0.43	0.08	0.23	0.74
70%	0.75	0.08	0.40	1.23

Table 4.5: Cross tabulation of percentages of low residential densities with mode choice

Note that walk and transit trip percentages are steady for all degrees of road density except for those instances where road density is minimal.

Chi-Square tests and Cramer's V tests have also been run for a number of variables. Table 4.6 below shows the results for the association test between travel mode choice and income; here a commonly established theory is that the greater the income, the higher the tendency to use privately owned vehicles as preferred transportation mode. The tables however show a different picture; the p-value for the Chi-Square statistics is significant at the 0.001 level so that we can safely assume an association between income and mode choice. However for the highest income classes car usage is consistently lower than, or at best similar to, that of the tracts with lower household incomes, so that the postulated theory appears not to be true.

Mode choice by income class	Car	Walk	Transit
100,000+	87.11	1.39	11.50
90,000-100,000	81.08	2.84	16.08
80,000-90,000	82.49	3.93	13.58
70,000-80,000	88.95	2.29	8.76
60,000-70,000	79.88	7.17	12.96
50,000-60,000	70.91	9.48	19.61
20,000-30,000	79.39	6.60	14.01
0-10,000	91.16	2.31	6.53
Statistics	DF	Value	Prob
Chi-Square	14	512.3904	<.0001
Cramer's V		0.1161	

Table 4.6: Chi Square and Cramer's V tests for income vs. travel mode choice

Other similar observations can be made interpreting the other cross tabulations but there are  $26^2$  categorical frequency distributions and Chi-Square statistics to analyze, one more reason to use inductive reasoning to narrow the search space in which to look for significant relationships.

#### **4.2 Bayesian Belief Networks for Travel Mode Choice with Individual Trips and Tract Level Aggregated Land Use Measures – Model 1**

The Bayesian Belief Network generated from the dataset containing all discrete trip records and tract level aggregated land use measures results in the graphical representation of relationships of Figure 4.1. In this instance the node “modechoice” is already selected so that the nodes in purple indicate the variables that belong to the parent set of “modechoice”, that is, those variables which strongly influence the choice of transportation means by the person making a trip.

Each node relates to “modechoice” with a different strength; in descending order the variables which exercise the strongest influence are:

- 1) vehicle count (vcount),
- 2) residential accessibility to public transit (RR),
- 3) whether or not the person making the trip has a driver’s license or not (driver),
- 4) the race and ethnicity of the person making the trip,
- 5) the amount of residential density, specifically the density of detached single-family/duplex, attached single-unit row housing, yards, and associated areas or the equivalent of 2 dwelling units/acre to 8 dwelling units/acre (medres),
- 6) the amount of commercial, retail and wholesale services available at the point of origin for a trip (comm.),
- 7) the age of the person making a trip (age),
- 8) educational levels (education),
- 9) number of children in the household (kidsinhh),
- 10) road density (access), and
- 11) income (income).

Like the commonly perceived truth that income is a great determinant in selecting a private vehicle instead of public transit, in this instance family income has indeed a relationship with “modechoice”. However it is weaker than a number of other variables. The same argument applies to the number of children in the household and to the educational attainment; both occur in the parent set of “modechoice” but are less influential than some land use variables such as residential

density and accessibility to transit. Also surprising is the fact that road density is a weaker variable than many other socioeconomic characteristics, a good hint that this is a possible confounding variable.

The relationship of vehicle count in the household to “modechoice” is as expected; so is the fact that the non availability of a car for a household and its members dictates whether a trip can be taken by private vehicle or not. This is a physical barrier (lack of a car) that the study takes as a matter of fact. It would have been interesting to have information on whether a household voluntarily decides not to have a privately owned vehicle, because this fact could influence the resulting networks. A similar argument can be made for the variable “driver” indicating the licensed status of a person taking a trip; people ought to have a driver’s license to drive but might choose not to do so. To be noted also is that not having a driver’s license does not prevent a person from using a private car as chosen transportation means, because they can also be passengers.

The cluster of nodes for residential density, commercial services and race/ethnicity that progressively influence “modechoice” is just one example of the nested arrangements of variables measuring socioeconomic characteristics and variables which indicate physical properties of the built environment. Another cluster would be residential accessibility to transit and number of vehicles in the household.

The node for the variable “trip distance” is highlighted by a pink color, indicating this is a variable that is determined by the selection of transportation means, as confirmed by the causal link departing from it. To most researchers, this arrangement seems incongruous because the more rational explanation is that a person would select the appropriate transportation means as a function of how far the

destination is, and not vice versa. The author subscribes to this view, and section “4.3.1 Cause and Effect or Feedback Loops?” addresses this apparently irrational result more in detail. Interestingly enough however, this arrangement proves that the variable “modechoice” is considered both a dependent and an independent variable at the same time, as in fact are all other variables in the present Bayesian Belief Network.

Figure 4.1: Simplified BBN for all available transportation mode choices considered in “modechoice”

What the BBN really highlights is the fact that the effects of socioeconomic characteristics and the properties of the physical landscape interact on many levels.

To demonstrate how this is the case, the decision tree associated with mode choice can be queried to provide insights about what interactions among the variables are responsible for generating the above BBN. Furthermore, it is possible to analyze the tree to investigate for what variable states – or instances – the interactions among the variables cause the relationships to change strength and direction [e.g. for how much land use mix, residential accessibility to transit, for how many vehicles in the household etc.].

In technical terms the parent variables of “modechoice” converge on the node representing “modechoice”, and in the “Causal Networks” section of Appendix A it has been established that in such cases nothing else can be learned about the child variable – “modechoice” – other than what may be deduced from the knowledge of its parents which are said to be independent. Such independence means that evidence on one of the parents bears no effect on the certainty of the others. Yet if there is any other kind of evidence influencing the child variable, then the parents become dependent because of the principle of explaining away.

Again, it follows that evidence may only be transmitted through a converging connection if either the variable in the connection or one of its descendants has received evidence, which in this study is entirely sourced from actual data.

#### **4.2.1 Model 1, Decision Tree**

Decision trees can be used to map the actual values of variables in a Bayesian network for which significant changes in the relationship among variables occur. Each tree is a dag containing exactly one root node and every node other than the root node has exactly one parent, so that each node split is only binary. Each leaf node contains a table of k distinct variable’s values that collectively define a conditional

probability distribution for that node and for those values only. The same variable node, with a conditional probability distribution for a different set of values can appear elsewhere in the decision tree.

The binary decision tree of Figure 4.2a shows the interrelations among the variables affecting “modechoice” by means of nodes, as in the above BBN, and by means of non-directed links. The structure of a decision tree is a hierarchical one where the first node, in this case vehicle count (vcount), is the root and the last nodes, those generating no more links, are called leaves. In Figure 4.2a the leaves are: age class of the person making the trip, availability of driver’s license, and road density.

The similarities with the graphical output of a BBN, however, end here, because each node also provides the probabilities that a given trip will be carried out by privately owned vehicles (**p1**), on foot (**p2**), or by public transit (**p3**) as per Figure 4.2b as a function of the other variables related to such a node. The lower a node is in the hierarchy, the more variables will condition/influence these probabilities. If the node “Medium density residential” is queried a probability pane presents the actual values for p1 to p3 based on the influence of the other variables considered up to that point in the tree. For example, the probability pane of Figure 4.2b, the same as in Figure 4.2a, shows that for vehicle count other than 0, good accessibility to transit, ethnicity other than African American (2), and with a minimum of residential density the probabilities of car usage, walking or transit patronage are, respectively, 63, 0.03 and 29 per cent. All other nodes can be queried in a similar fashion and the analyst can detect the change in mode choice probabilities as a function of what set of conditions are being considered by the tree. Therefore it is possible to find out for what instance of each variable these probabilities change, so that questions such as



‘how much does transit usage increase should a variable change from one class to another’ can be answered.

The tree is read by following the node/variable hierarchy and the tree of figure 4.2a can be interpreted as: depending on whether the person making the trip has a privately owned vehicle at his/her disposal, if not, race/ethnicity comes into play as shown by the upper part of the decision tree of Figure 4.2a. Class 2 for the variable “race” is the populations of African American descent for which, limited to the case of the Baltimore Metropolitan Area, a given trip has a 51% probability of being carried out on foot, and a 31% probability of being made by public transit, should there be no other variables considered. The overall propensity is therefore for more trips to be made by walking or transit which is in line with expectations due to the lack of cars available for the persons making this decision. Of course this is not to say that African Americans in general walk and use transit more than do other groups, because the tree so far has only considered the case of vehicle count = 0.

Interestingly, these probabilities are moderated by whether the person taking the trip departs from a location with a semi sparse residential environment; in this case if he/she lives in an area with 2 to 8 dwelling units per acre, the probability of a trip made by public transit goes down to 6% and the probability of a trip made by privately owned car goes up to 39%. Again this is a valid result because the sample data includes the possibility that someone might be a passenger, rather than the driver for a car trip.

If the race/ethnicity of the person making the trip is not in class 2, that is not African American, then mode choice is moderated by residential accessibility to transit, given that vehicle count is still zero. In cases of good accessibility the

probability that a trip will be made on foot is 20% and by transit 62%. In other words even for other racial and ethnic groups, the probability of using transit, when lacking private cars, is high.

More interesting is the case for vehicle count other than zero, which means one or more cars are available to the household of the person taking a trip. The probability that such a trip will be made by privately owned vehicle is 89%. This changes if the trip maker has no access to transit at the origin of the trip because p1 (privately owned vehicles) goes up to 93%, a difference of 4%, and decreases only in areas of high commercial spaces (p1 down to 60%). At the same time these percentages for car vs. walk vs. transit trips vary according to some socioeconomic characteristics of the person making the mode choice, especially age and number of children in the household.

If transit accessibility is good, race and ethnicity play a role again: if the trip is taken by an African American, the probabilities of a car trip, walk and transit trips are respectively 83%, 6% and 10%. If the person taking the trip is not himself/herself a driver chances are he/she will walk 23%, take transit, 31% or still take a ride in a car 45%.

Again, for a race and ethnicity class other than 2, auto use stays high at 75% as one would expect, and in the case of medium residential densities there is a good probability (up to 25%) that the trip will be on foot. We are now down to the middle of the upper branch departing from residential accessibility to transit (RR).

By following this branch the scenario above changes according to the number of vehicles in the household. Even with a high number of vehicles, where street networks are dense the selection for trip mode is split between car 50%, and walking

48%. Where road densities are not so great, the probability for a trip made by transit decreases to 21% with some moderating effects on auto use if surrounded by at least some commercial spaces. Therefore people would be as likely to walk as they are to use privately owned vehicles in such conditions.

When car availability is limited to one vehicle, when the person taking a trip has a low educational attainment and does not have a driver's license, the probability of a trip being made on foot or by transit is 15% and 32% respectively, even if the previous conditions hold true. So we witness the influence these socioeconomic factors are having on the decision process for transportation means.

Finally, income appears to exercise some influence on trips made by privately owned vehicles but its influence is weak. In terms of probabilities, low income (class 7 and above – 0 to 12,000 dollars personal annual income) moderates the probability of a car trip down to 24%, while the probability of a transit trip stays unchanged and the probability of a walking trip increases to 71%, an indication that privately owned vehicles are too expensive a proposition.

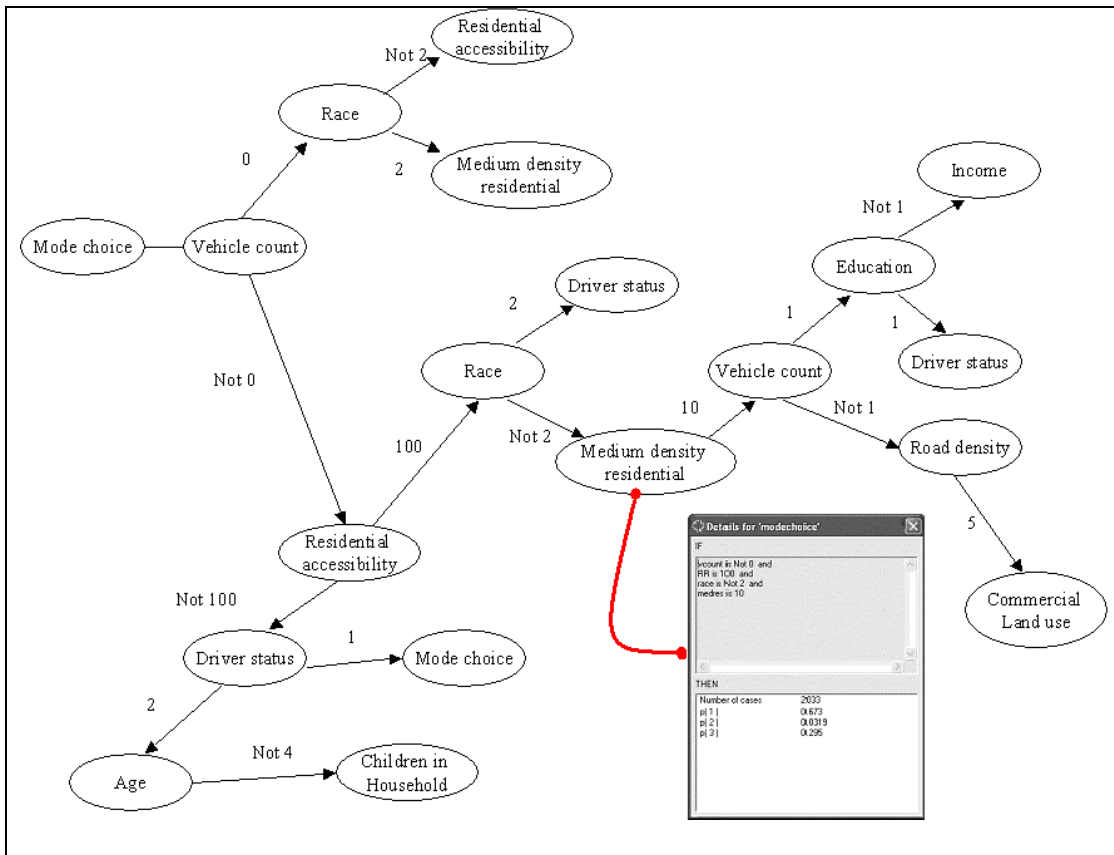


Figure 4.2a: Decision tree for Model 1

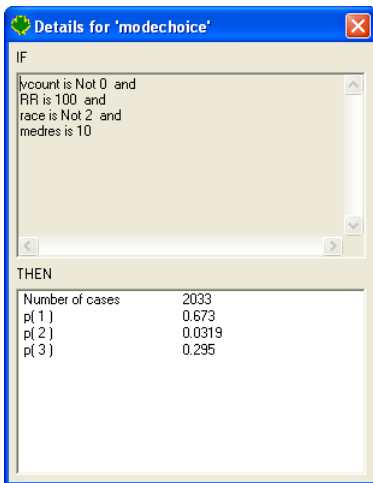


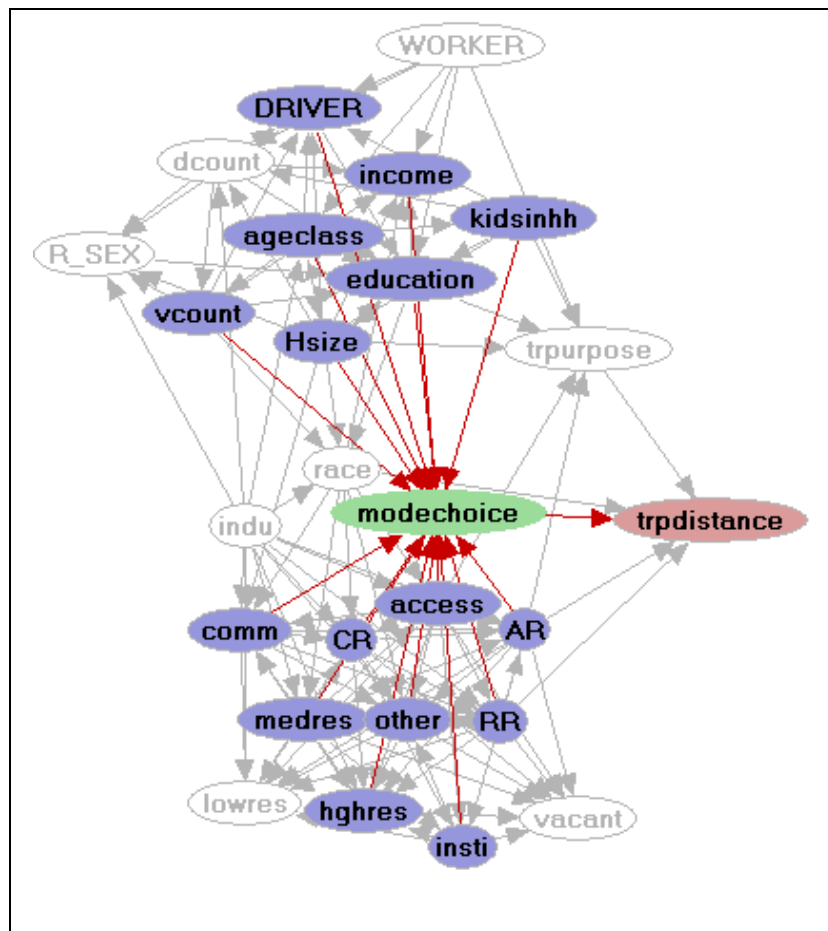
Figure 4.2b: Sample probabilities query for the decision tree of Model 1.

#### **4.2.2 Binary Mode Choice Modeling: Privately Owned Vehicles, with Individual Trips and Tract Level Aggregated Land Use Measures**

The above BBN and binary decision trees illustrate the concurrent influence that land use and socioeconomic variables exercise on people's decision process for transportation means. It is valuable as much for what is included as for what is not included in the parent set. For example, a number of a priori assumptions commonly made would expect household size and number of drivers in the household to exercise great influence on "modechoice", when in fact this is not the case for this dataset. Also missing from the parent set are a number of land use variables such as commercial space and high density residential.

The question then arises as to whether some of these variables could more directly influence some specific transportation means but not all of them. For this reason, the variable "modechoice" has also been processed as if it had a binary distribution of values, once for car trips vs. all other trips, once for walking trips vs. all other trips and also, transit trips vs. all other trips, thus generating three subsets of Bayesian Belief Networks modeling the relationship between land use and travel behavior which is more specific to the three different transportation means analyzed in this study.

The binary modeling attempt for privately owned vehicles results in the BBN of figure 4.3a.



DRIVER: availability of driver's license  
 WORKER: worker status  
 Dcount: driver count  
 R\_Sex: gender  
 Vcount: vehicle count in household  
 Ageclass: age class  
 Income: personal income  
 Hsize: household size  
 Education: educational attainment  
 Kidsinhh: number of children in household  
 Trpurpose: trip purpose  
 Race: race and ethnicity  
 Modechoice: travel mode choice  
 Trpdistance: trip distance  
 CR: Commercial accessibility from transit  
 RR: Residential accessibility to transit  
 AR: overall transit accessibility for tract  
 Access: road density  
 Comm: commercial land use  
 Medres: medium residential density land use  
 Highres: high residential density land use  
 Lowres: low residential density land use  
 Indu: industrial land uses  
 Insti: institutional land uses  
 Vacant: vacant land or parks  
 Other: other land uses

Figure 4.3a: Binary BBN for privately owned vehicles

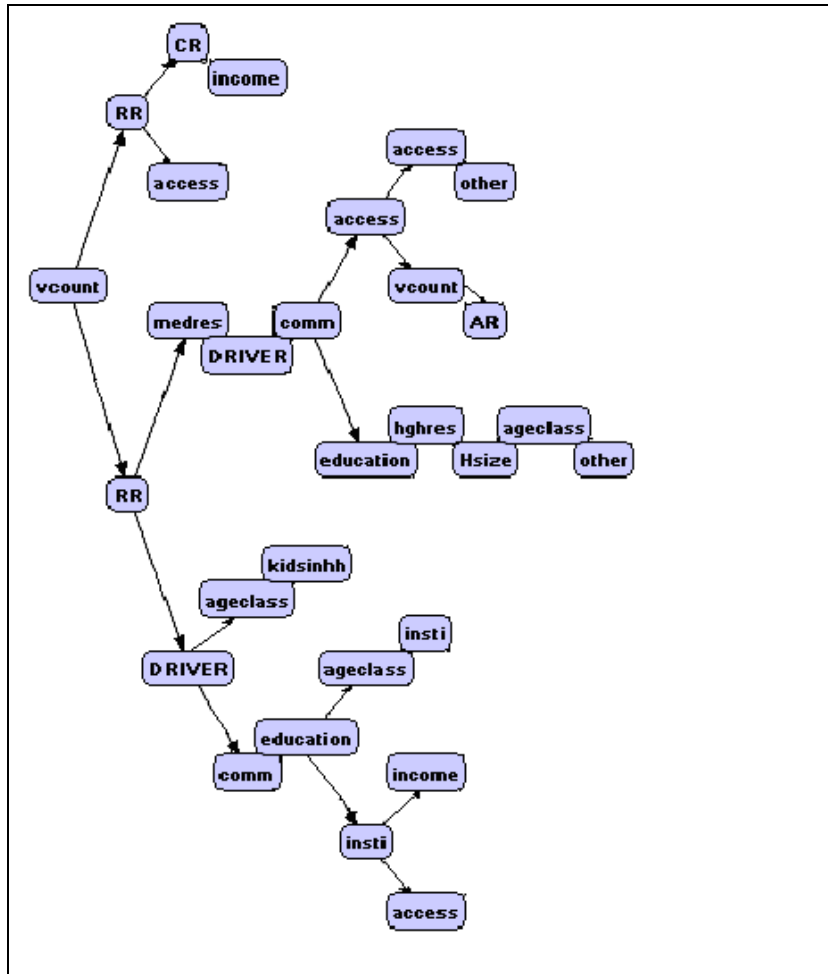


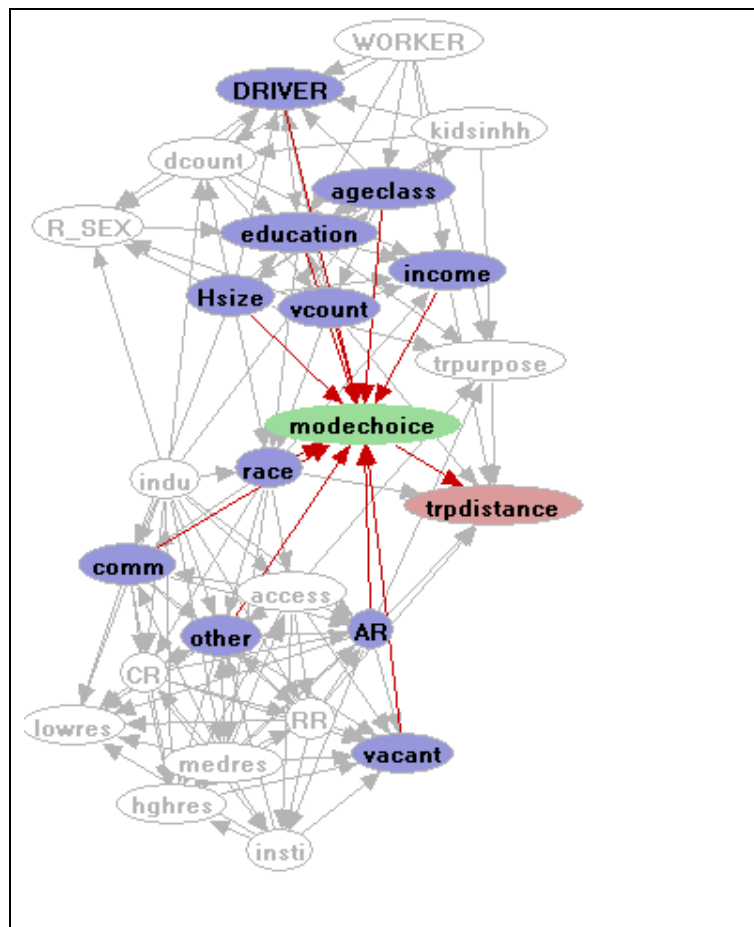
Figure 4.3b: Decision tree for privately owned vehicles

The complexity of this BBN is greater than for the general case presented in figure 4.2a. Could this indicate that the flexibility of the privately owned vehicle lends itself to a greater variety of uses and to a more complex relationship structure when it comes to deciding about its use? Probably yes, but the question remains as to the reasons why variables like high density residential, various other land use and transportation variables, such as other land uses, commercial spaces, accessibility to transit etc. are now included.

On examination of the decision tree associated with this model, presented in Figure 4.3a, the network of binary splits reveals that the variable for race and ethnicity is no longer a major factor but other socioeconomic variables are now very influential, among them education, income and age; of course the availability of a driver's license figures preeminently in this tree too, because it is a necessary condition for driving a privately owned vehicle. The probabilities of choosing a car as a transportation means are mediated by the land use variables already encountered in the previous case. In particular, the middle branch of the tree indicates that for poor road densities but good accessibility to transit, a relatively dense neighborhood and a mix of commercial space, the probability of a car trip stops at 73%, with the remaining portion largely taken up by transit (23%) and little else left on foot, an indication of poor accessibility associated with low road densities. For even lower road densities, the probability for p1 (privately owned vehicles) goes up to 84%, but the middle branch of the tree offers a different picture. In this case the probability of a trip made by car is 54% when medium and high-density residential and large commercial spaces are present. The associated probability for a transit trip in this instance is a very high 39% when the person taking the trip is middle aged, from a small household, with "medium" educational attainment. Finally, the lower branch of the tree indicates similar percentages with the exception of a very high likelihood of car usage for trips originated in areas with large percentages of land use devoted to institutional use such as schools and government areas.

#### ***4.2.3 Binary Mode Choice Modeling: Walking, with Individual Trips and Tract Level Aggregated Land Use Measures***





DRIVER: availability of driver's license  
 WORKER: worker status  
 Dcount: driver count  
 R\_Sex: gender  
 Vcount: vehicle count in household  
 Ageclass: age class  
 Income: personal income  
 Hsize: household size  
 Education: educational attainment  
 Kidsinhh: number of children in household  
 Trpurpose: trip purpose  
 Race: race and ethnicity  
 Modechoice: travel mode choice  
 Trpdistance: trip distance  
 CR: Commercial accessibility from transit  
 RR: Residential accessibility to transit  
 AR: overall transit accessibility for tract  
 Access: road density  
 Comm: commercial land use  
 Medres: medium residential density land use  
 Highres: high residential density land use  
 Lowres: low residential density land use  
 Indu: industrial land uses  
 Insti: institutional land uses  
 Vacant: vacant land or parks  
 Other: other land uses

Figure 4.4a: Binary BBN for walking

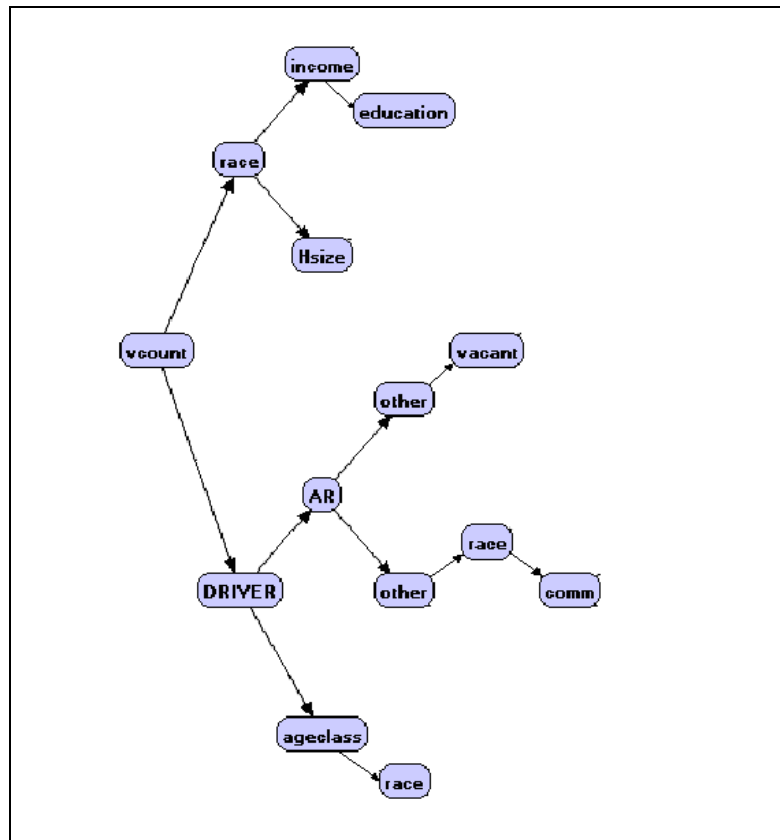


Figure 4.4b: Decision tree for walking

The binary BBN modeling walking trips – Figure 4.4a – is less complex than the one for motorized travel and, surprisingly, it does not indicate that the built environment and its land use have much influence on our decision to make a trip, except for the amount of vacant land, commercial space, and non-urban land use such as agriculture. In this instance, the middle branch of the tree in Figure 4.4b reveals that for remote areas not entirely served by transit, and not densely populated, the likelihood of a trip being made by privately owned vehicle is a large 92%, as we would expect. The decision tree for this BBN also reveals that the only significant percentage of persons choosing walking as their transportation choice happens

because of lack of driver's license, for all ages except for seniors. When race and ethnicity are included, the likelihood of a walk trip by African Americans increases from 19 to 27 percent. However the biggest jump in the likelihood of a walking trip happens when we also consider the amount of commercial space. When "comm." equals 60, meaning a range between 51 to 60% of the tract is used for commercial and service uses, then p2 (transport on foot) grows to an interesting 36%. The state of the other variables associated with this result sees race and ethnicity other than 2, marginal amounts of "other" land use types, and good accessibility to/from transit. The point here is that residential density appears to affect the use of privately owned vehicles more than walking trips. On the other hand, how handy commercial land uses and services are influences the decision to walk more. Finally, as seen from the tree in Figure 4.4b, for the case of no vehicles available to the household, and depending on household size, the likelihood of using transit or walking are in the 50% range, as found earlier when discussing the aggregate mode choice.

#### ***4.2.4 Binary Mode Choice Modeling: Transit, with Individual Trips and Tract Level Aggregated Land Use Measures***

The BBN for transit trips, figure 11a, is largely influenced by transportation and land use variables, with high and medium density residential, commercial space and accessibility to transit all having a strong influence on mode choice.

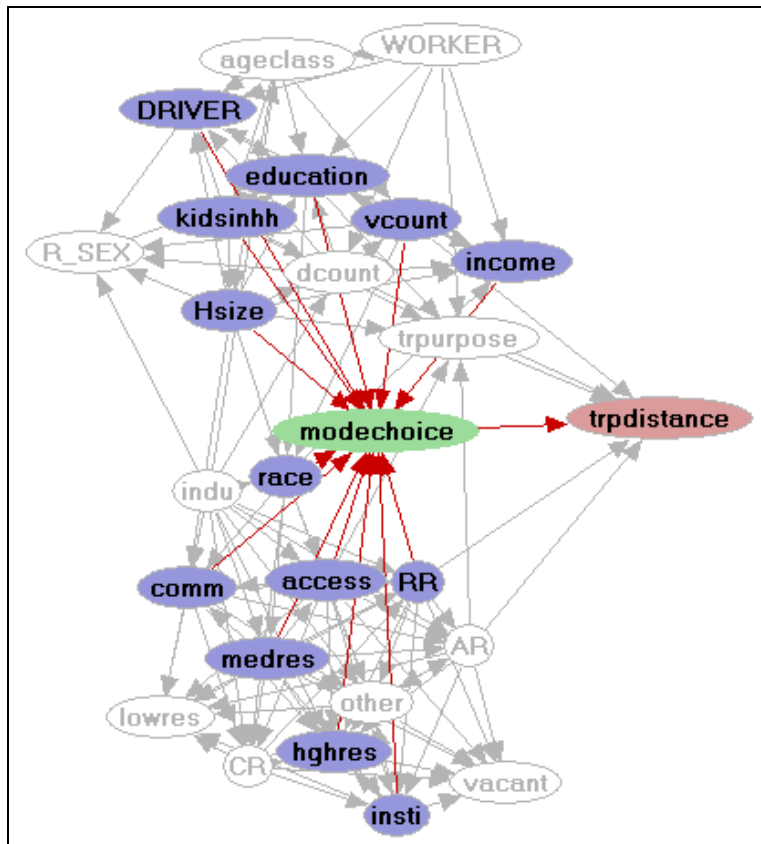
The biggest difference that can be observed in its associated tree presented in Figure 4.5b is that, unlike all other trees encountered so far, the sine qua non variable is no longer vehicle count in the households (vcount) but accessibility to transit, specifically, residential accessibility to transit in the tract (RR).

For this variable a significant difference occurs among those tracts entirely serviced by transit and those that are not. This is not surprising because bus, train and metro stops in Baltimore are so well dispersed that about 80% of the city is within one quarter mile of such a facility, a distance that can easily be covered on foot. As we move along to the suburbs, however, the frequency and accessibility to transit stops drops considerably and dramatically just outside the city limits.

Not surprisingly the variable “driver”, indicating whether or not a person taking a trip has a driver’s license, is more relevant for those tracts with limited accessibility to transit, as can be seen following the upper branch of this tree. At this point the probability that a trip will be carried out by transit drops from 11.00% to 0.04% if the person has a driver’s license, or it increases to 16% should the person making the trip have no driver’s license.

However the probability of auto use when this is a feasible alternative, as in the upper branch of the tree, stays at 94% regardless of all the other variables, including amount of commercial space, medium residential density and accessibility. This shows how the lack of accessibility to transit is a key barrier to its use, as expected, but also that where transit is accessible it may be able to draw considerable use.

In fact, for those tracts that benefit from full transit coverage, the probability that a trip will be based on public transit increases as medium density residential increases from 25% to 36% when high density residential is also present in the tract; for low income classes the probability of a transit trip increases only marginally by 4 percentage points to 40%, even if there is at least one privately owned vehicle in the household.



DRIVER: availability of driver's license  
 WORKER: worker status  
 Dcount: driver count  
 R\_Sex: gender  
 Vcount: vehicle count in household  
 Ageclass: age class  
 Income: personal income  
 Hsize: household size  
 Education: educational attainment  
 Kidsinhh: number of children in household  
 Trpurpose: trip purpose  
 Race: race and ethnicity  
 Modechoice: travel mode choice  
 CR: Commercial accessibility from transit  
 RR: Residential accessibility to transit  
 AR: overall transit accessibility for tract  
 Access: road density  
 Comm: commercial land use  
 Medres: medium residential density land use  
 Highres: high residential density land use  
 Lowres: low residential density land use  
 Indu: industrial land uses  
 Insti: institutional land uses  
 Vacant: vacant land or parks  
 Other: other land uses  
 Trpdistance: trip distance

Figure 4.5a: Binary BBN for transit

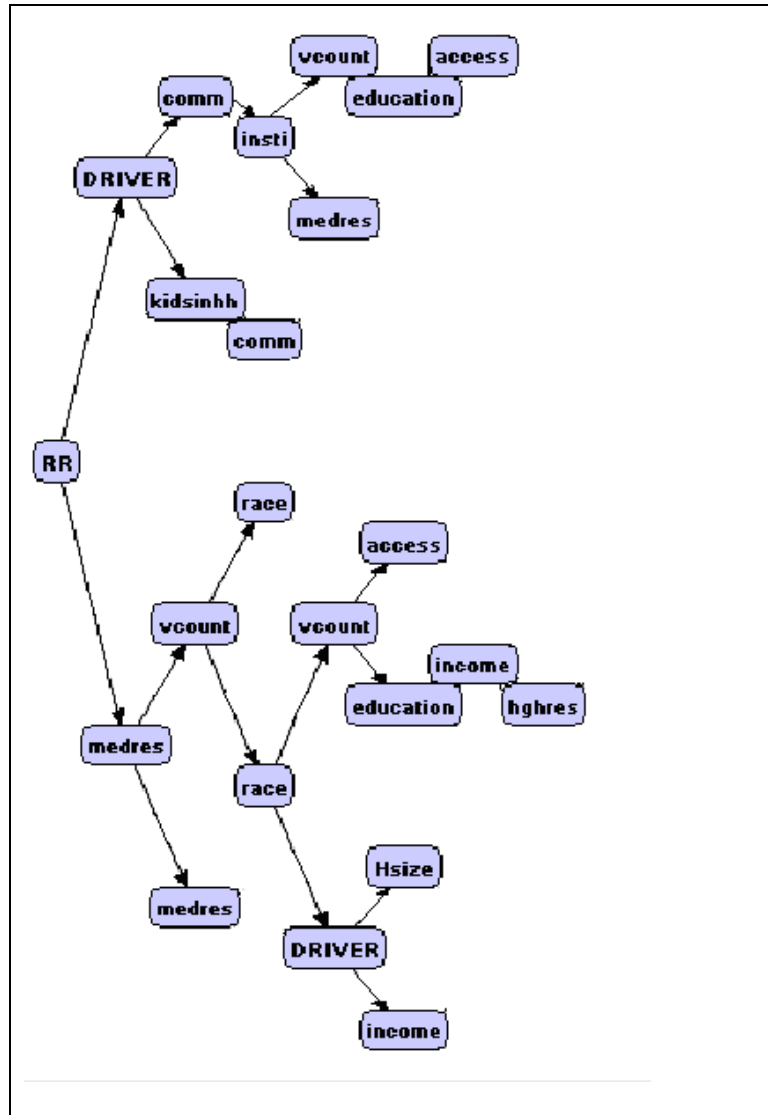


Figure 4.5b: Decision tree for transit

The lower branch of the tree in Figure 4.5b also indicates the effects of household size on the likelihood of using transit; the probability of transit use drops from 20% for a single person household to a negligible 1 % for all other cases with larger household size. In conclusion, the combination of 30% of a tract's land use in medium residential density (2 to 8 dwellings per acre) together with full access to transit stops can potentially draw up to 22% of the transportation demand.

### 4.3 Building a Conceptual Model

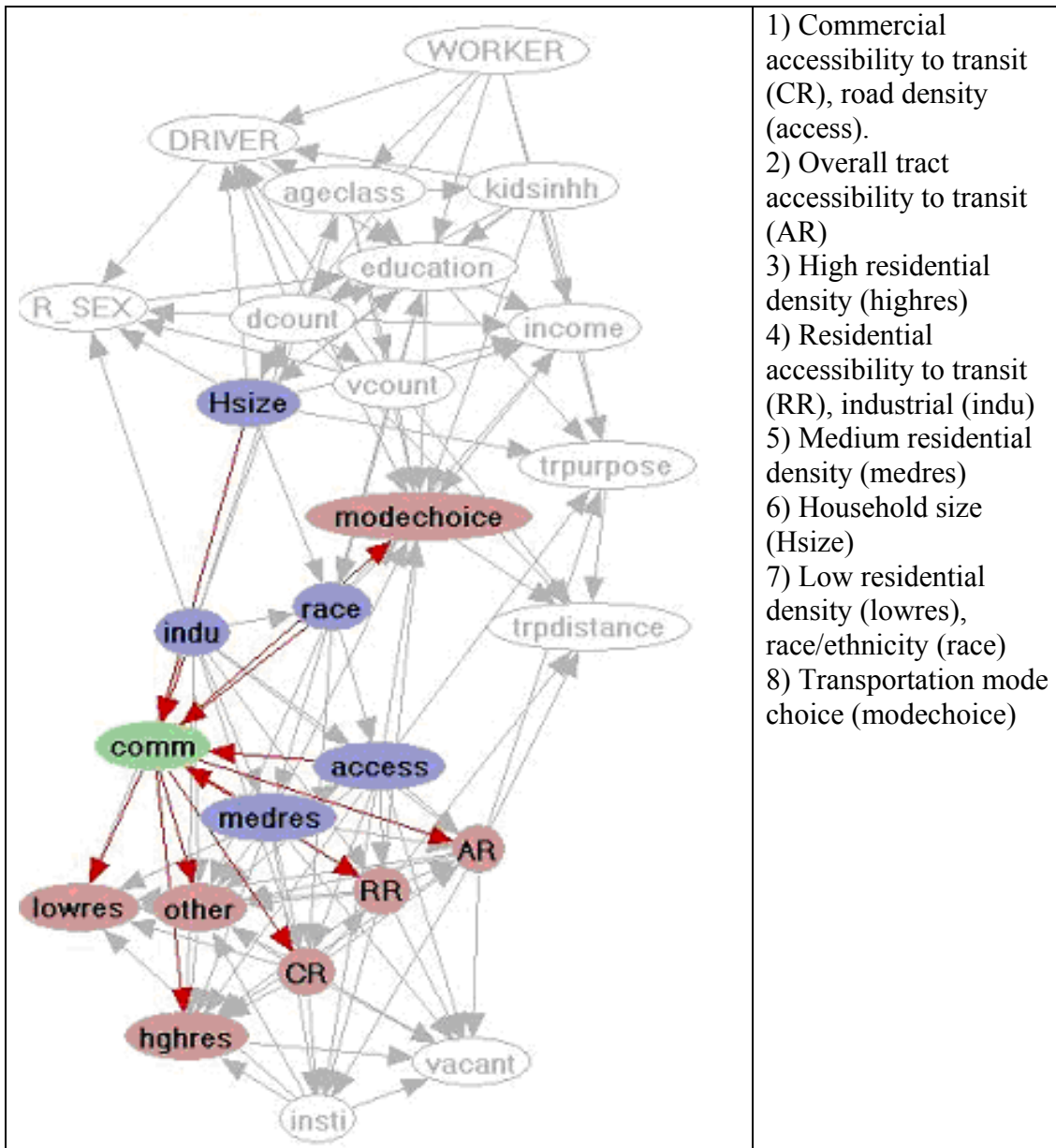
From the BBN outputs of Figures 4.4a and 4.5a it is possible to learn what variables influence people's choice of transportation means and which variables have the greatest direct influence on "modechoice". Still unknown however are the relationships among the other variables and what roles they have in influencing "modechoice" in an indirect way. After all, high density residential (highres), and trip purposes (tripurpose) do not figure in this Bayesian Belief Network, yet we would expect them to be of significance in people's decision making processes.

In fact, the analysis of this BBN cannot be limited to the variables belonging to the parent set of "modechoice", but for all the variables in the network their relative causal paths need to be examined as well. This section highlights these relationships for all other variables so that a comprehensive conceptual model of the relationship between land use, transportation and travel behavior can be derived. This is accomplished by mapping the causal paths from each parent variable to its child until the mode choice variable is reached. For example, from the variable "worker" to driver count to "education" to "modechoice" we learn that whether or not a person is employed is of relatively little importance to the selection of transportation means, a finding that undermines the view that commuting to work is a significant parameter for the determination of transportation mode choice. The job status of a person taking a trip in relation to "modechoice" is mediated, first, by the number of people in the person's household holding a driver's license, and then by his/her educational attainment.

The same reasoning can be applied to all other variables presented in the following table and the ones of Appendix B; in each, the same BBN of Figure 4.4a is

presented, but different variables are selected, one at a time, so as to cover the entire set of parent variables for mode choice. Again, the nodes are color coded so that the nodes in green represent the variables being investigated, purple nodes are the variables that have causal influence, and the nodes in pink are those variables that can be thought of as being dependent on the status of other variables. In each case the variables influencing and being influenced by the selected node are listed in order of strength. Comments are made about interesting relationships, but a more comprehensive discussion of the entire output can be found later in Section 5.1 “Implications for Current Theories of the Relationships among Land Use and Travel Behavior”.





- 1) Commercial accessibility to transit (CR), road density (access).
- 2) Overall tract accessibility to transit (AR)
- 3) High residential density (hghres)
- 4) Residential accessibility to transit (RR), industrial (indu)
- 5) Medium residential density (medres)
- 6) Household size (Hsize)
- 7) Low residential density (lowres), race/ethnicity (race)
- 8) Transportation mode choice (modechoice)

Figure 4.6a: BBN for commercial space

### **4.3.1 How To Interpret Dags and Decision Trees**

The graphical output from the algorithm is quite substantial and still in need of interpretation. The parents set of mode choice and mode choice itself constitute the local dags of relationships pertaining to travel mode choice yet, as seen above, each parent of mode choice has its own local dag. This means that to learn about all the variables affecting a child the analyst has to go through all hierarchical levels of the networks and reconstruct all direct and indirect relationships between variables.

Rebuilding the path of all relationships is a necessary starting point, but little is yet known about the magnitude and direction of such relationships. The former is easy to gauge when using the WinMine's option to display relationships by order of strength, while the latter can be learned from an analysis of the decision trees, node by node.

Here is an example of how to interactively process the algorithm outputs to build and understand the overall conceptual model for travel mode choice. This example is more complex than that involving the variable "worker" and begins by focusing our attention on two variables, rather than one, vehicle count and income, from Figure 4.1 which is the local dag for all discrete trips and all possible travel mode choices. Two local networks emerge, representing the parents and children for the two variables. As the interpretation of these networks and their associated trees is complete, the analyst can add variables to the existing discussion so as to exhaust all possible parents of mode choice. Alternatively he or she might choose to select other pairs, proceed in an analogous interpretation of the outputs, and synthesize the findings in a later step.

The selection of the variables income and vehicle count is not arbitrary, but follows the theoretical assumption that as income increases, so does the number of vehicles in a

given household. Figure 4.7 below shows that the parent sets of income and vehicle count are “age class”, “gender”, “driver status”, “worker status”, “education” “race and ethnicity”, “household size”, and “driver count”. Note how these last two variables are parents of both income and vehicle count, which in turn, affect the node for mode choice.

Figure 4.7 could be used to derive a conceptual model of mode choice as a function of income and vehicle count. It has two of the problematic relationships encountered above, which makes it interesting. Should the causal link from vehicle count to mode choice be reversed? The evidence processed by the algorithms does not support this thesis and there is empirical evidence that households with access to a privately owned vehicle have more access to the job market and potentially to better paying jobs than those depending on transit (Shen 2001).

Yet to be answered are the questions about the direction of relationships among variables. Decision trees can be followed and the relative variation of classes within variables examined to determine just that. For example in the decision tree for income the variable vehicle count is considered eight times, under a number of different conditions. The probabilities that a person will have a given income, according to the direction of the relationship discussed above, depend on how many vehicles are available for a trip. Therefore the analyst can examine those eight query panes and observe that the likelihood of a person having low income increases as the number of vehicles available to that person decreases. It can also be observed that the magnitude of this variation, as a likelihood change, is in fact small, which could explain why income is found to have a weak effect on mode choice.

Unfortunately decision trees do not always consider the same variable as many times as discussed above, so in some instances it might be difficult to find enough evidence to uniquely determine a sign for a relationship between two variables. In this case the numerical outputs in the conditional probability tables can be looked at to see if they present a more comprehensive numerical analysis.

The lesson to be learned from this conceptual model is that we could model mode choice simply as a direct dependent variable of vehicle count, or moderate this variable by noticing that vehicle count is related to race and ethnicity, educational attainment, then income and, only then, travel mode choice.

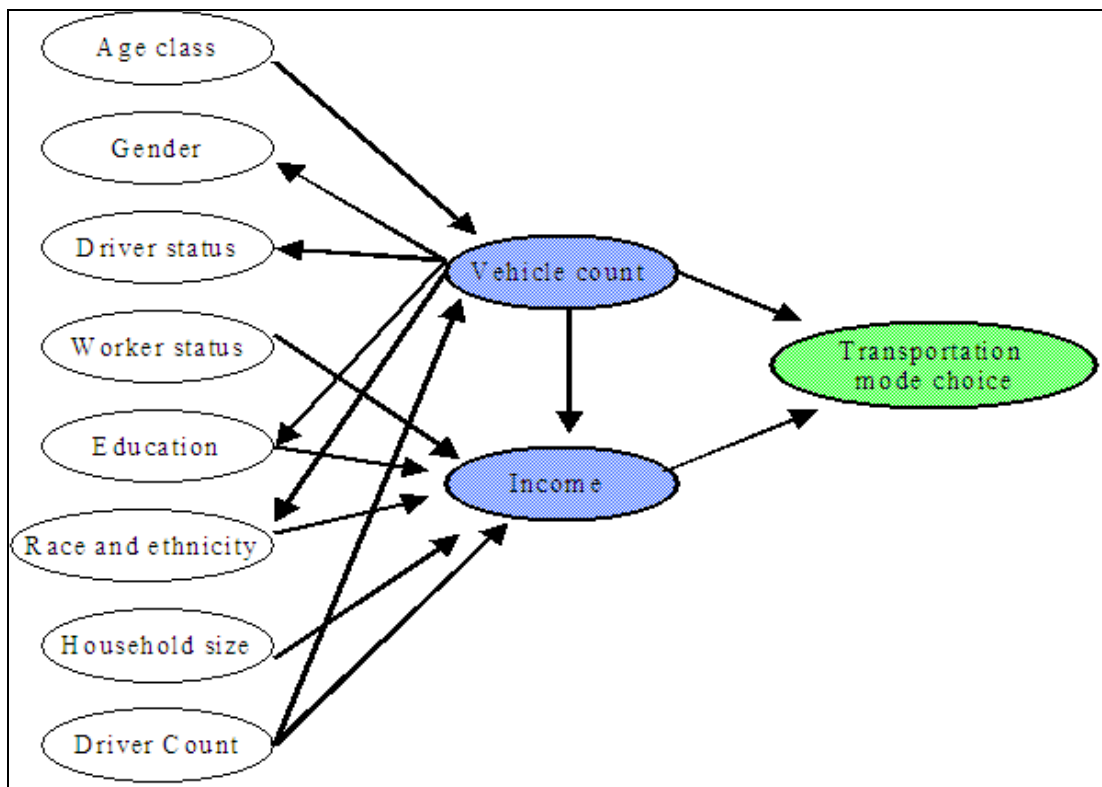


Figure 4.7: Building a conceptual model for income, household vehicle count and travel mode choice.

### **4.3.2 Cause and Effect or Feedback Loops?**

When it comes to the Bayesian Belief Network presented in Figure 4.4a, it is difficult to interpret the cause and effect linkage between mode choice and trip distance. The link originates from the node representing transportation choice and it terminates at the node measuring trip distances. This directed vector implies that “modechoice” is a determinant of trip distance and not vice versa. In fact our logical understanding of this relationship is that a person taking a trip would select the transportation means as a function of the distance to be traveled. The network effectively suggests, instead, that one would first select a transportation means and would subsequently travel for a certain distance normally associated with such transportation means.

The decision tree for the BBN mapping the relationships between trip distance and the other variables, in Figure 4.8 below, does suggest that in the case of privately owned vehicles trip distances tend to be longer, and overall one would agree with the proposed relationships linking trip distance to vehicle count, driver status, and various accessibility measures at the origin of a trip, but the direction of the various cause and effect links could arguably be the opposite.

To address these issues, an alternative type of network was generated to see whether the vector in question should in reality have only a module of strength but no direction. This was possible because of the development by Heckerman et al. (1995) of a similar algorithm to the one used to create Bayesian Belief Networks, but one that does not enforce the notion of acyclicity in the networks. This means that feedback loops can be modeled among variables that are at the same time determinants of and determined by other variables.

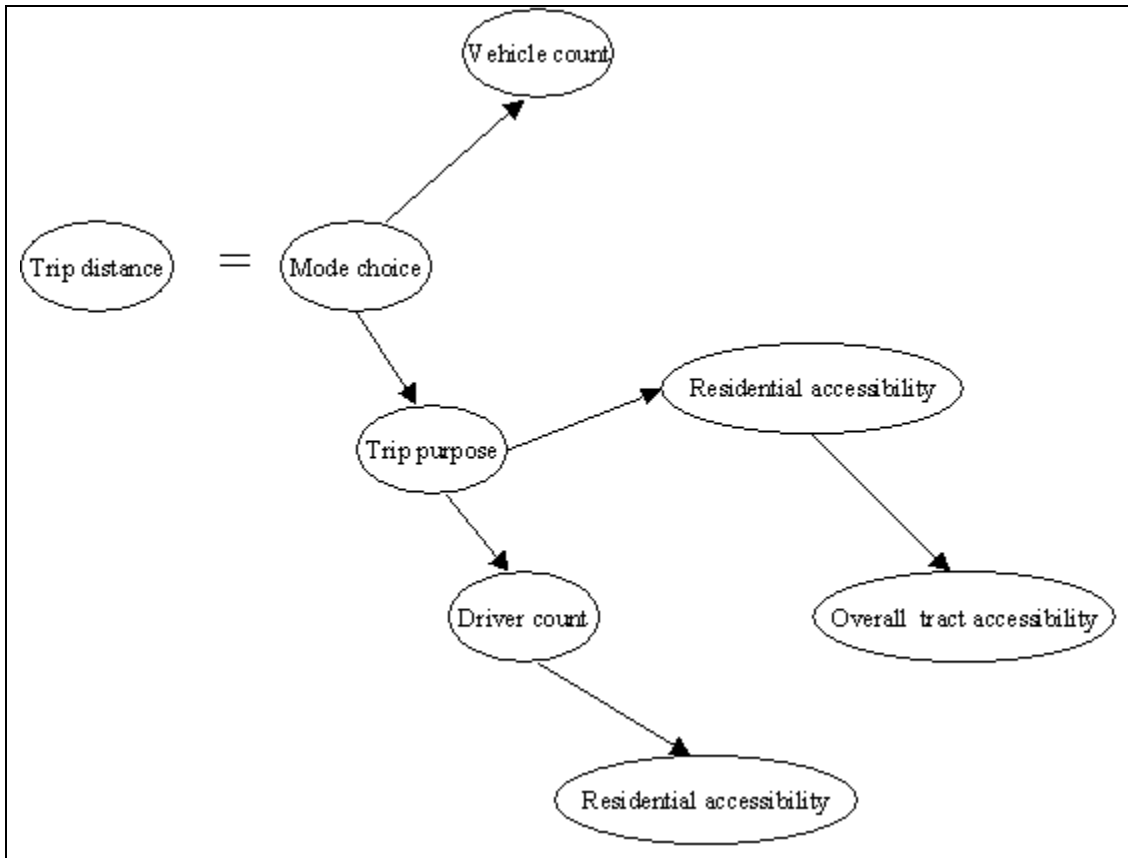


Figure 4.8: Decision tree for trip distance

The case of the relationship between “modechoice” and trip distance was found to be a feedback loop, and as such modeled as one instance, where the selection of transportation choice is related to trip distance and where trip distance influences the decision of transportation mode choice.

The algorithm used to derive such feedback loops is referred to as a “Dependency Network” and it results in a directed-less graph of relationships among variables where no notion of causality can be inferred; however because of the experimental nature of this algorithm and its limited use in applied research, and the overall untested effectiveness, the author has deemed it too immature to be used in this study.

#### **4.4 Bayesian Belief Network for Both Travel Choices and Land Use Characteristics Aggregated at the Tract Level – Model 2**

The second set of BBNs mapping the conceptual model of the relationships among travel behavior, transportation and land use variables was derived by aggregating all available sample information at the tract level. This means that not only were the socioeconomic parameters of the population and the physical characteristics of the landscape spatially aggregated at the tract level, but also all the information relating to trip mode choice was recounted to tally transportation usage, by mode, at the tract level as well. The implication here is that although the new data structure is nothing more than a transpose of the dataset used in Model 1, where the occurrence of a specific value within a variable is now a variable in itself, the total number of data records available for the analysis is limited by the total number of tracts in the area under investigation, 615.

From a heuristic computational view, this is a limiting factor to the reliability of the resulting networks. A model testing exercise was conducted to evaluate the reliability of the three models derived from the three different data structures used. In the case of disaggregate trip records of Model 1, 30% of the original dataset was set aside and it was found that the network structure of Figure 4.1 is able to predict correctly the value in this test dataset 85% of the time. This test results in a parameter equivalent to the coefficient of determination (R square) and it indicates the goodness of fit each model has.

In the case of Model 2, created by using the aggregate information of 615 tracts, the R square coefficient is no better than 60%, indicating not only a great loss of predictive power but also a loss in the variance of the original information, as expected due to the data aggregation.

In terms of the Bayesian Belief Networks and decision trees obtained, the major differences found, in comparison with Model 1, relate to the lower significance of land use variables, confirming the increased generality across the physical landscape.

The third type of output from Model 2 is in the form of conditional probability distribution (CPD) tables of mode choice as a function of the status of other variables. These types of tables have not been presented for Model 1 because they resulted in a poor combination of conditional variables and travel mode choice, due to the non direct match between the many discrete trips and the much fewer land use aggregates at the tract level.

Unlike the graphical representation of BBN networks and binary decision trees, which do not provide the changing probabilities of “modechoice” for all given combinations of variable states, conditional probability distribution tables are obtained from the trees to include all such possible combinations. Conditional probability distribution tables, CPD, can only potentially answer this question, because they are a derived product of the actual BBN, where the original dataset is now used to try and calculate all the remaining probabilities. In the words of Max Chickering, one of the authors of the WinMine software, the probabilities in the CPD tables are a “blown out” version of the probabilities presented in the decision trees, and allow the use of inference engines that require complete information about the probability of any event and all other events.

What is important is that “all of the context-specific independence relationships that are learned in the tree still hold in these blown-out tables, although these relationships are now hidden in the parameter values”, that is, they are not visible as parent child relationships.



The results of the CPD tables of Model 2, and later of Model 3, typically have much fewer edges because the number of parameters in a table grows exponentially with the number of corresponding parents. In contrast, when deriving decision trees, the number of parameters is proportional to the number of leaves (which can be linear in the number of parents). This is definitely a major limitation to the specific use of the WinMine algorithm because it precludes obtaining a full conditional distribution of mode choice as a function of all other variables, which would be quite valuable for policy evaluation. Section 4.5.1 “Threshold Analysis” explains this point in more detail.

Value instances for land use and income variables			Associated probabilities that a trip will take place by a specific travel mode		
Income	% Medium Residential	% High Residential	Probability Motorized Trip	Probability Non-Motorized Trip	Probability Transit Trip
1	20	60	86	6	8
1	30	10	88	5	7
1	10	60	60	11	29
1	10	70	60	12	28
2	40	20	86	6	8
2	20	70	60	11	29
5	20	50	37	17	46
5	10	80	60	11	29
6	10	50	60	11	29
6	90	10	86	6	8
8	10	90	37	17	46
8	90	10	88	6	6

Table 4.7: Sample conditional probability table for land use and income variables. The probabilities of using a private vehicle are high, regardless of the income level, which indicates that even persons on low incomes tend to use private vehicles for their mobility needs. Personal income classes are from 1, high income, to 11, low income.

The algorithm used in this research calculates the probability for the state of a class in all variables so the resulting tables are quite large. For example, in the case of the classes for variable driver count, which report the number of drivers per household, these are

assigned a probability of occurrence based on the occurrence of all other classes in all other variables. This is quite useful but the algorithm has no knowledge that the land use variables should all add up to 100% of the land use cover for a given area; it follows that the CPD tables for land use variables include situations where the occurrence of an 80% high density residential is compared with the occurrence of a 60% commercial land use, a case which clearly does not happen in reality.

#### **4.5 Bayesian Belief Network for Both Travel Choices and Land Use Characteristics Aggregated at the zip Level – Model 3**

The third data structure used for the analysis is similar to that of Model 2 but it aggregates socioeconomic characteristics of the population, physical land use and travel behavior at the five digit zip code level, resulting in only 150 spatial units available for the analysis. From a heuristic point of view this appears to be a significantly insufficient number of records to produce a statistically robust network. Of interest is the fact that land use variables are now even less relevant to the decision making process of the person taking a trip, and socioeconomic characteristics and other variables such as income appear to have gained more relevance. The land use variable “other”, which includes all other possible land uses, is also difficult to interpret, and is the one indication that at this scale of data aggregation the algorithms are now incapable of discerning physical characteristics of the built environment as determinants of mode choice, in part also because of the indeterminate nature of the “other” land uses. The figure below shows the resulting BBN, as it appears with just three parent variables shown (Figure 4.9) and as it appears with its entire parent set of variables (Figure 4.10).

One interesting outcome derived from the analysis of the models' CPDs is that high probabilities of transit share and non-motorized trips have occurred either when the land use variables have shown a large concentration of residential land use (as in downtown) or when there were small percentages of a mix of different land uses. This result would quantitatively support the argument of those who favor mixed use and higher densities as a means to improve transit ridership and abate private vehicle use.

Value instances for land use variables only				Associated probabilities that a trip will take place by a specific travel mode		
% Medium Residential	% High Residential	% Commercial	% Other Land Uses	Probability Motorized Trip	Probability Non-Motorized Trip	Probability Transit Trip
10	10	20	60	90	5	5
10	10	40	10	52	16	33
10	40	30	20	69	18	13
10	80	10	0	62	8	30
20	10	10	10	78	11	11
30	10	10	10	78	11	11
10	20	50	10	53	14	33
10	30	20	10	53	14	33
10	40	20	10	63	7	30
20	60	10	10	78	10	12
10	30	30	10	52	15	33

Table 4.8: Sample conditional probability table for land use variables only. Notice how a small mix of uses has high probabilities of transit and non-motorized trips. The same happens with extremely high percentages of residential densities. Could this be the quantitative proof about land use mixes?

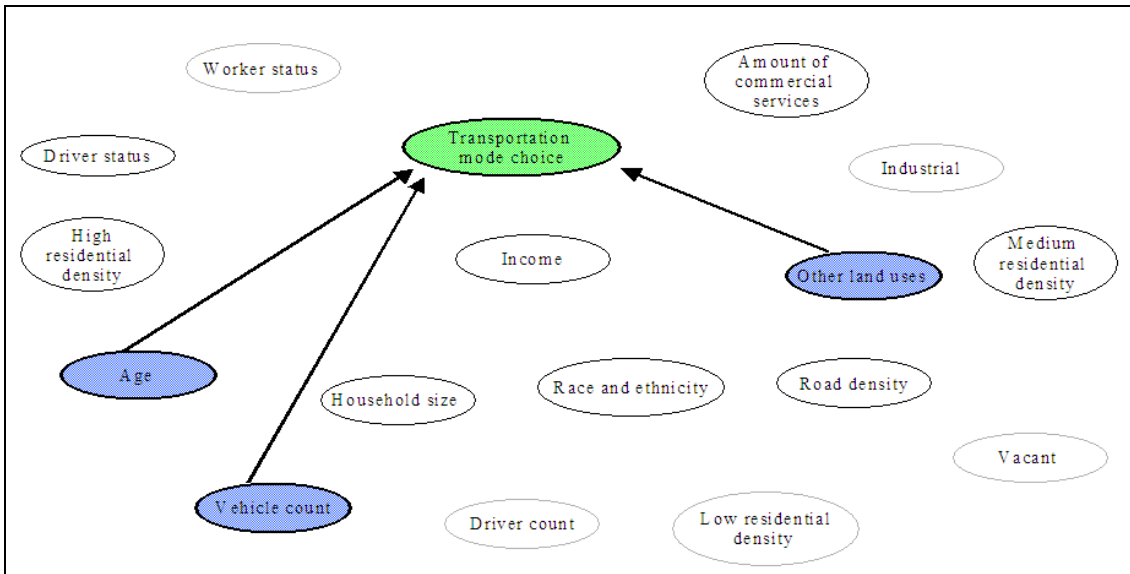


Figure 4.9: Model 3: five digit zip areas, three parents of “modechoice”

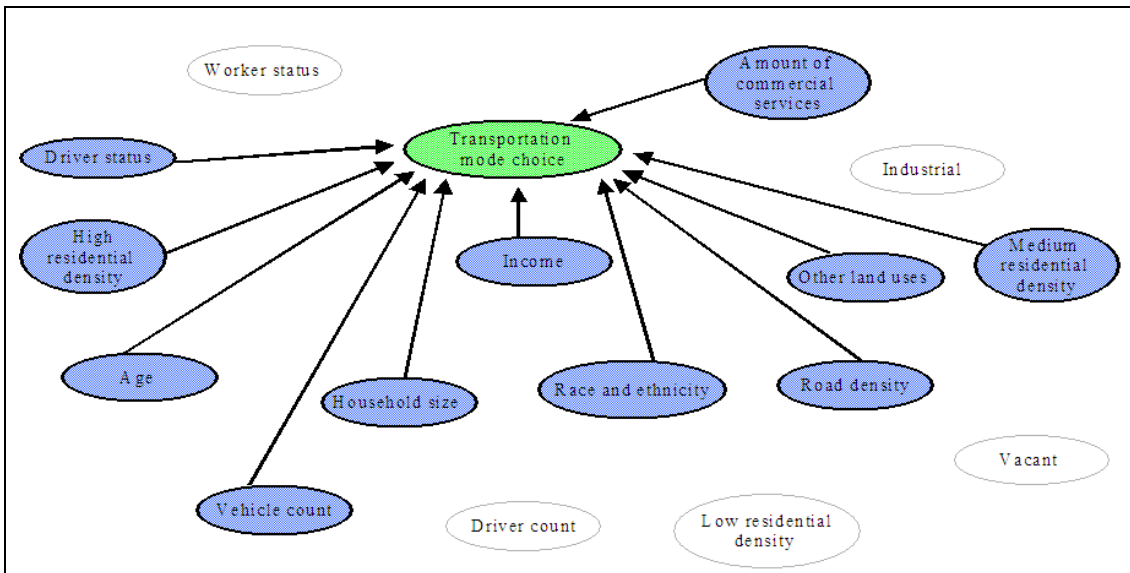


Figure 4.10: Model 3: five digit zip areas

#### 4.5.1 Threshold Analysis

If each transportation travel mode obtained from the conditional probability distribution table is independently reorganized, the analyst can select a minimum likelihood for a specific travel mode and see what conditions are associated with it. In table 4.9 for

example, the land use conditions that underpin a minimum likelihood of 90% that a trip will be made by car are listed together with the land use conditions for a minimum likelihood of a walking trip of 14% and the land use conditions for a minimum likelihood of a transit trip of 30%.

Value instances for land use variables only				Associated probabilities that a trip will take place by a specific travel mode		
% Medium Residential	% High Residential	% Commercial	% Other Land Uses	Probability Motorized Trip	Probability Non-Motorized Trip	Probability Transit Trip
10	10	20	60	≥89		
10	10	10	70	≥89		
10	60	10	20	≥89		
40	10	10	20	≥89		
10	10	10	20	≥89		
10	10	10	30	≥89		
10	10	20	60	≥89		
10	20	50	10		≥14	
10	10	40	10		≥14	
10	40	30	20		≥14	
10	30	30	10		≥14	
10	20	50	10		≥14	
10	30	20	10		≥14	
10	20	40	10		≥14	
10	10	40	10			≥30
10	80	10	0			≥30
10	20	50	10			≥30
10	30	20	10			≥30
10	40	20	10			≥30
10	30	30	10			≥30
10	20	20	10			≥30
10	50	20	10			≥30

Table 4.9: Thresholds for mode choice likelihood and land use mix

The table shows that car usage is more likely where the percentage of “other” land uses is the greatest compared to all other land use variables. The same high probability of a trip being done by privately owned vehicle is found where residential is present as an exclusionary or predominant land use.

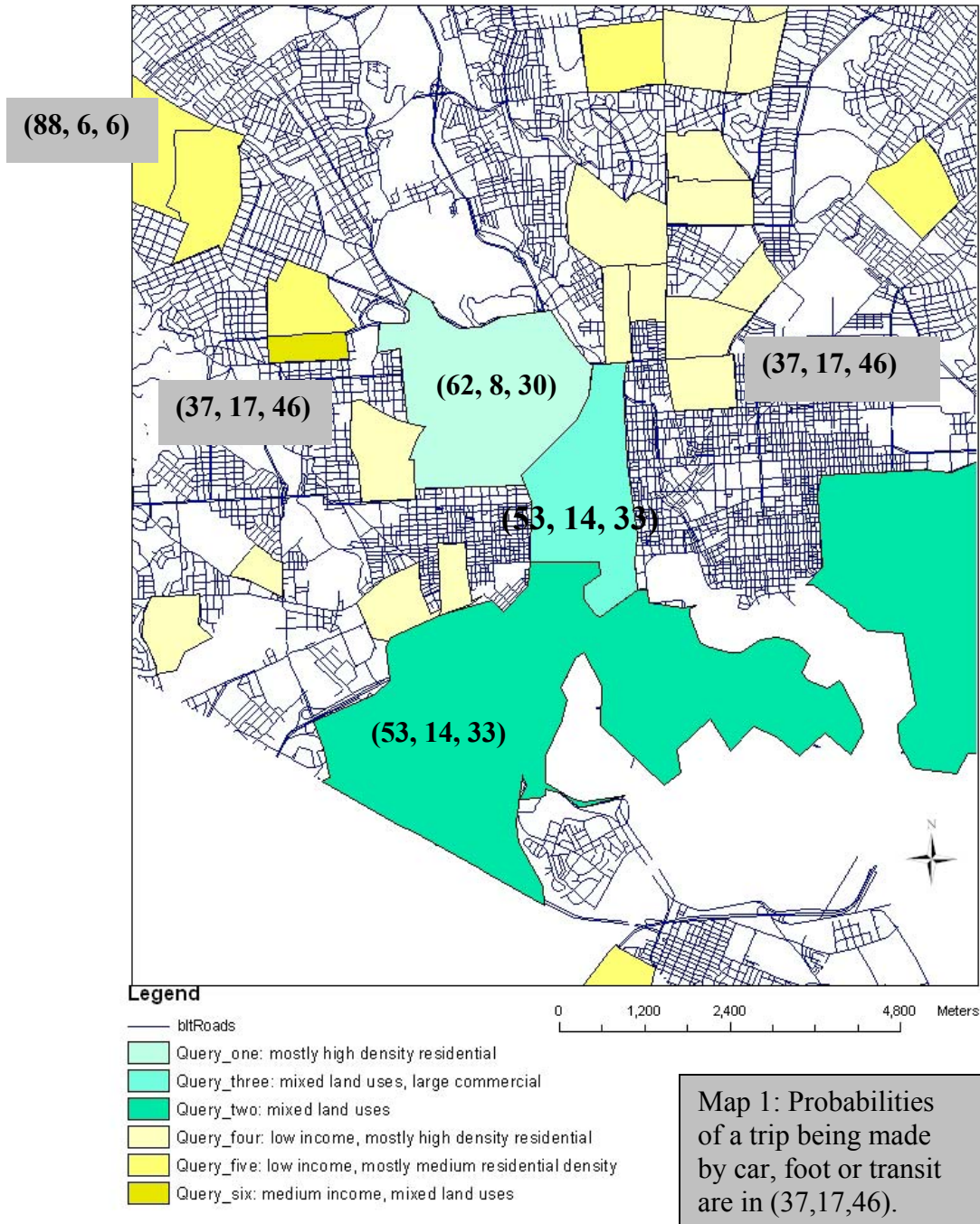
On the contrary, the ideal land use conditions for significant probabilities of walking and transit trips consists of a higher degree of land use mix, where large proportions of commercial land uses are significant for both, but especially so for walking.

These trip mode probabilities are directly linked to physical locations that show a combination of specific land uses, so it would be quite valuable to be able to map where these numbers apply.

By joining the data used to derive Bayesian Belief Networks to a spatial dataset containing spatial identifiers for zip and tract areas, spatial queries can be executed to establish, on the ground, which areas and neighborhoods are more likely to generate higher percentages of a particular transportation mode. The value in identifying such neighborhoods is that the decision to study one locality rather than others can be justified quantitatively because such localities may have certain physical properties that facilitate the use of a specific transportation means over all various alternatives.

Policymakers can also identify, on the ground or on the maps, what combinations of land uses are more conducive to a particular travel mode choice and include this quantitative piece of evidence into their policies.

## Query samples



Map 4.1 shows the results of six attribute queries that identify the probability of transportation mode choices for car, walk and transit, and locates them in space based on the instance of variables such as income and residential densities.

From Table 4.8, at the zip level, the following queries were run:

Query 1: Select Where Medres = 10 And Highres = 80 And Comm = 10 And Other = 0

Query 2: Select Where Medres = 10 And Highres = 30 And Comm = 20 And Other = 10

Query 3: Select Where Medres = 10 And Highres = 20 And Comm = 50 And Other = 10

From Table 4.7, at the tract level, the following queries were run:

Query 4: Select Where Income = 8 And Medres = 10 And Highres = 90

Query 5: Select Where Income = 8 And Medres = 90 And Highres = 10

Query 6: Select Where Income = 5 And Medres = 20 And Highres = 50

Income: personal income	Highres: high residential density land use
Comm: commercial land use	Indu: industrial land uses
Medres: medium residential density land use	Other: other land uses

Table 4.10: Spatial Query Legend

In the south central portion of the map, which is by and large an industrial and commercial area, as well as for the areas identified in Queries 3 and 5, the probabilities of a trip made by transit are higher, yet these zip and tract areas are not equally well served by public transportation. One could argue for a complete rethinking of which areas should be better served by our public transit services; such a policy shift however ought to be also based on the total number of trips because even the absolute certainty that just a handful of trips will be made by transit is less relevant than the possibility of having many not so probable trips made by transit, which in fact, could result in higher overall ridership numbers.



## 4.6 Model 1 Stability Tests

Given that in this particular research the term “heuristic” can be defined as progressively searching for evidence, one might wonder what exactly constitutes “evidence” and how much of it is enough. The fundamental unit of analysis, as mentioned before, is a case or string of values from all of the variables in the dataset, but one single instance of such a combination of values is hardly representative of a real process at work. By the same token are 100 equal cases enough? Also, how much should be considered given the size of our dataset?

In this study the author has set minimum threshold levels by which cases were deemed representative if they occurred in the data at least once every one hundred records. Other tests were run with frequencies of up to 6 cases per 100 records, but no variation in the complexity of the resulting networks was observed within these parameters. The complexity increased for thresholds of less than 1% so that the most extensive parent set of variables for mode choice was of seventeen variables; for higher threshold values the complexity of the resulting networks decreased to just a few strong variables, such as vehicle count, which significantly affected the outcome of transportation choices.

Another concern arose about the possibility that variables with fewer classes, hence less variation, might have had a stronger role in the resulting networks. Driver status for example has only two states, whether a person has a driver’s license or not. Could this variable substitute other variables in the resulting networks because of its strength?

All this testing was conducted by changing the “K” parameter, which is the WinMine command that determines complexity, defined as a continuous value between 0 and 1. The variation of this parameter provides the clues to some substitution effects among

the variables but there is another approach that was tested to isolate the relative strength among variables in relation to “modechoice”. This approach involves the derivation of Bayesian Belief Networks with a different set of variables so that the resulting parent sets of mode choice can be compared. In this instance three networks were generated, all similar in data structure to Model 1, and compared in Table 10.

The first column shows the parent set of variables for mode choice derived from Model 1, the second column represents the same parent set when derived from all the same variables less trip distance, trip purpose, education, number of children in the household, and gender; the third column is again the same parent set when derived from a dataset that is also missing all the accessibility measures to transit. In yellow are the variables that have lost their presence in the final BBN and which have been replaced by such variables as trip distance, education and number of children in the household. The count column shows the number of times the variables in Model 1 have appeared in the other runs, and shows that what has not been substituted was in fact a strong parent variable in all three cases. The counts with a value of one are from the newly added variables and if a test was conducted with more variables than those available, it would be normal to expect their count to increase as well, given the high strength demonstrated by these nodes in their respective networks of Figures 4.6 through 4.11.

Varying the K parameter so as to increase complexity would result in a larger parent set where it is possible to see the relative strength of each variable, but where the analyst would not know which variable is susceptible to substitution by what other variable. Instead with the results of this method outlined in Table 4.11, the analyst can be more confident that the list of parent variables do in fact exercise a significant influence on

the variable of interest and it is possible to discard other weaker variables. This might be a very useful exercise when trying to reduce the complexity of the resulting Bayesian Belief Networks. Also, the variables that were replaced were classified in both a high and a low number of classes, so no substitution effects can be attributed to the discrete aggregation of the data.

All variables in Model 1	Model 1 - 5 variables	Model 1 - 8 variables	Count
Vcount	Vcount	Vcount	3
Tripdistance	Driver	Driver	1
RR	RR	Highres	2
Driver	Ageclass	Medres	3
Race	Race	Ageclass	3
Medres	Medres	Race	3
Comm	Comm	Comm	3
Ageclass	Income	Vacant	3
Kidsinhh	Vacant	Access	1
Income	Highres	Hsize	3
Education	Insti	Income	1
Access	Worker	Other	2
K=0.01	K=0.01	K=0.01	

Table 4.11: Variable substitution summary for Model 1

#### 4.7 Internal & External Validity

The validity of this study rests largely on the accuracy of the two major datasets used as input data, the land use/land cover layer from the Maryland Property View product and the Baltimore Add-On information for the National Household Travel Survey 2001. The latter sample includes only 3519 households living in the area of interest, barely 0.5% of the total number of households in the Baltimore Metropolitan Council’s jurisdiction, and only accounts for 9200 individuals out of a total of two and a half million. By contrast, the Summary File 3 from the US Census Bureau directly samples 418,000 individuals, about

15% of the total population in the study area. The NHTS proponents have made some effort to ensure the representativeness of their products. Specifically, a detailed stratification was employed in the selection of households based on population density, number of household vehicles and the number of household members. Also, a list-assisted, random digit dialing (LA-RDD) method was employed to select the sample so that households without a listed telephone number had as much of a chance of being contacted for the survey as households with freely available listed numbers.

Unfortunately the biggest drawback to the use of NHTS data is that no questions were asked about the decision making process each household member uses to determine his or her choice of travel mode. No data were recorded with what factors influence people's behavior, and this study only looks at travel mode choice as a response to personal assessment for these factors.

Another source of concern is the discrepancy between the land use information analyzed in this study and the real physical landscape the data tries to capture. This is due to the minimum size of coherent and single uses that can be detected by relying heavily on small scale aerial and satellite image interpretation. A second problem arises from the question of how often is the land use/cover information updated over the years, although care was taken to use the 2001 version of such land use data to match the time the 2001 NHTS sample was carried out.

In relation to the validity of this study to other cities and localities, two issues should be considered. First, the study area encompasses a wide variety of land use mixes: an urban core, city neighborhoods, semi urban and rural areas as well as a number of "rusting belt" style industrial parks and factories. This is the same land use mix we would

encounter in most American cities along the eastern seaboard and, given the broad classification of land uses into large percentage classes, it is reasonable to expect that the same variety of land use combinations considered in this study would apply to other localities as well.

The second consideration is the differences between the socioeconomic characteristics of the citizens of the Baltimore Metropolitan Statistical Area and those of the rest of the nation. To establish whether the data used in the study, specifically variables such as education, income, race and ethnicity, gender, working status, number of drivers in the household, household size, transportation mode choice, and availability of a driver’s license, are statistically different from the NHTS national averages two tests were used. In the case of continuous variables, student’s “t” tests were used to compare national averages against the Baltimore sample, and Chi-Square tests were used to compare the distribution of classes within a variable, when dealing with discrete variables.

From the National Household Travel Survey the Consolidated Statistical Area of Baltimore was extracted into a local sample. The averages and class distributions for this sample have been compared to the remaining national data and the results show that only two variables, number of drivers in households and household vehicle counts are somewhat different from the national averages (Table 4.12), but that all other variables show that the Baltimore Area is in line with the rest of the nation.

Variables in Add-On data vs. national NHTS data	t-value	Pr >  t
Driver count in household	-1.92	0.0554
Household vehicle count	-10.03	<0.0001

Table 4.12: T-tests for the difference in number of drivers and cars in households between the local data and the national NHTS sample

The Baltimore area also has a lower average number of drivers per household, 2 instead of the national average of 2.1, and an average of 2.2 vehicles per household instead of the U.S. average of 2.34.

Other differences are specific to the transportation mode split, with 1.74% of the trips in Baltimore being carried out by public transit bus, versus a national average of only 0.67%. The share for the subway/rapid rail system in Baltimore is 1.22% versus the 0.22% nationwide. Walking is used by 10.56% of the people surveyed in Baltimore versus a national average, as per NHTS surveys, of 7.7%. In Baltimore therefore there appears to be a higher tendency to use transit and walking as transportation modes, but these differences with the national averages were not significant. These and other measurements are compared in Table 4.13.

Variable	NHTS ADD-ON only average	National NHTS average
Transit	1.74	0.67
Walking	10.56	7.7
Subway use	1.22	0.22
Drivers per household	2.1	2.2
Vehicles per household	2.2	2.34
High school	17%	23.63%
Graduate school	17.98%	9.25%
% African American	14.01	9%

Table 4.13: Selected percentage differences for variables in both the local NHTS Add-ON and the national NHTS data

#### **4.7.1 Graph Representation Validity**

Besides the validity and appropriateness of the data used in the research some testing needs to be done to verify the appropriateness of a method for the intended analysis. For this scope there are two quantities that measure how well the resulting networks model the input data. In the same way that a regression model is more accurate than a simple baseline model chosen in the form of the mean of the dependent measurements, the “lift over marginal” log score provides information on how well the model fits the data. As in all appropriate modeling attempts, it is possible to test the model on a subset of data to verify that its relationship construct and conditional probabilities still hold true. In fact, a small dataset set aside for this purpose has been correctly modeled at the 65 and 80 % levels which, for aggregate data is what we would expect. Also as the author worked towards covering a greater variety of data structures, the lift over marginal scores have all been greater than the marginal log scores; therefore we can rest assured that their construct is better at treating all variables equally.

#### **4.8 Synthesis**

The results presented so far prove that the Bayesian heuristic approach can be applied to map the relationships between land use and travel behavior and to generate new and interesting hypotheses about the nature of these relationships. There is no unique or true map of relationships among the dags introduced above, nor it is advisable to present the method as being capable of finding such true and universal causal links among physical landscape and human behavior. On one hand the study is too limited in its measurements to claim this ability and, more importantly, being heuristic, it is intrinsically an

approximation of reality, and one that relies heavily on data. However this due disclaimer should not impede the next interpretation of these findings, one which sees the theoretical body of evidence discussed in the literature review, reviewed under the lenses of the lessons learned from the analysis of the results. To do so all the dags, for all variables in all modes, are to be looked at critically and compared to the existing literature. In fact salient points will be made as learned from all the outputs above, whether dags, decision trees, conditional probability distributions or simple cross tabulation tables.



## Chapter 5: Discussion

There is a sort of parallel between researchers, with the data they use, and parents to be. They know a baby is coming, but can only assume what he/she will become after many years of hopefully good parenting and schooling. Like parents, researchers can only assume that certain measurements, collected to investigate a particular problem domain, will have certain characteristics, hopefully great variance among the variables, and also certain relationships among them. It is not until after the actual data analysis that researchers can really say anything conclusive about the data and the hypotheses they were used to prove or disprove. By then it is too late, short of starting anew, to modify our conceptual models about reality and the data collected.

This study uses a new methodology to query the data for relevant relationships in such a manner as not to influence the response, while still gaining useful knowledge of what are the relationships hidden among the variables under study. In the case of land use and travel behavior the results are a multifaceted window into the complex connections relating the physical landscape and human activity; the Bayesian heuristic method helps sort through them but, as seen previously, the analyst must properly assess the various types of outputs and draw his/her conclusions. This chapter discusses this author's findings in relation to the various tenets presented in Section 1.2 on this study's research hypothesis, specifically whether travel mode choices are more strongly influenced by socioeconomic factors than by land use characteristics, the effects of mixed land use on travel behavior, the variation of mode choice with respect to household income, what is the appropriate spatial scale at which to capture behavioral responses to land use, and whether there are unique land uses that influence certain specific travel modes.

## **5.1 The Heuristic Perspective on Travel Behavior and Land Use.**

In such a context as the one above, and as part of the research approaches in general as well, the ability to objectively “map” our datasets prior to any attempt to conceptualize them would be of invaluable benefit. In fact, what if tentative conceptual models could be easily and quickly derived for any kind of data, whether continuous or discrete, such that researchers would have an unbiased understanding of the information they are dealing with prior to any deductive thinking? One would be in a position to assess whether a particular dataset is in fact appropriate or not to answer specific questions. In the case presented here, the author would have expected some traditional views on income, household vehicle count and travel mode choice to be validated, when in fact the resulting BBN has found little difference in mode choice as a function of income. The algorithm also found no connection between income and amount of commercial space, as some would have expected. This finding is actually reassuring because the Baltimore area is characterized by a great number of small commercial entities well dispersed within the BMC jurisdiction, and the author would have found such a connection to be suspicious. This is not to say that such a relationship could not exist in another dataset, maybe one with a different partitioning of commercial space based on type of retail. However, when beginning an analysis of data with a Pearson correlation coefficient we become aware of associations among variables and their measurements but not of causal explanations, so there is no way to tell if our personal conjectures about conceptual models hold true or not, unless we proceed as suggested in this study.

Ideally should the analyst follow the research wheel of Figure 1.1, from data collection to data analysis, he/she would end up back at the establishment of theoretical axioms used to design deductive studies. What can we learn in following such a path when dealing with travel behavior? What else can be said from this inductive derivation of a conceptual model of travel behavior, transportation and physical landscape with respect to established common theoretical frameworks? Earlier, the four areas of research interest were stated as being a) the relative influence of land use and socioeconomic variables on travel behavior, b) the specific role of land use variables with respect to travel mode choice, c) their sensitivity, and d) issues of scale and aggregation. In relation to the four points above it was found that:

a) There is no reason to believe land use characteristics are less relevant to the travel mode choice outcome than all other socioeconomic characteristics of the person making a trip. In all three model runs, as well as when modeling mode choice strictly as a binary choice, all resulting networks and trees presented a nested structure of both socioeconomic variables and land use characteristics. In fact, in some instances, residential density and amount of commercial space were significant determinants of large percentage shifts in mode choice, even when controlling for race, ethnicity and income. Yet authors such as Cervero and Kockleman (1997), Kulkarni (1996), Handy (1996), Kitamura, Monktarian and Laidet (1997), Holtzclaw (1994), Messenger and Ewing (1996), and O'Regan and Quigley (1988), among others, have found socioeconomic characteristics or the land use mix to be major factors per se. The networks presented in this study, interpreted as conceptual models of the relationships between various measures of land use and socioeconomic characteristics,

show these relationships as interacting at all levels. In fact only vehicle count appears to be a crucial variable for the selection of privately owned vehicles as transportation mode, while all other variables interact at multiple levels.

Transit accessibility, for example, confounds both the influence of land use, specifically medium density residential, and the availability of a driver's license. At different hierarchical levels, all variables related to transportation, socioeconomic characteristics or land use interact; the decision trees show for which instance of these variables the relationships among them co-vary or diverge in their effects on "modechoice". Thus inferential statistics, if based on such complex hierarchical dependencies, could bear so much more information than can our current models, limited as they are by our bounded rationality.

b) Further evidence of the land use travel connection is independently detected by the heuristic approach used in the study, in that the same proportions of transportation mode share can be achieved in areas where the land use is either consistently above 70% high density residential or where the area is characterized by a variety of different land use mixes, each usually no more than 30% of the total land use. The results obtained in Table 4.8, from modeling travel mode choice and land use at the five digit zip level codes, indicate that the same probability for a transit trip, around 30%, can be obtained for both the case of high residential density equal to 80% of the land use, and the case of 30% high residential density and 30% commercial land use. This finding supports the argument for land use mix without any a priori attempt to do so on behalf of the researcher. The decision tree of Figure 4.2a, which is derived by using discrete trips and travel mode choice, but

tract level aggregation for land use features, also supports this thesis because when the query pane shows a mix of commercial and medium density residential uses the probability of trips made by private car decreases 33% compared with high density residential alone, or other single use. The presence of commercial uses is also particularly important in the decision process related to walking trips; by querying the binary modeling for walking, Figure 4.4b – also derived using discrete trips and tract level land use aggregation – one sees how the probability of a trip made on foot increases as the percentage of commercial land use increases.

On the other hand, when the predominant land use mix is not residential but either institutional, vacant, low density residential or other, the mode share for private auto use always appears to be higher, independent of income, as found by the steady probabilities of car usage in the tree of Figure 4.2a which is derived by using discrete trips and travel mode choice but tract level aggregation for land use features. The same fact emerges from Table 4.7 where, for zip level aggregation of both trip mode choice and land use characteristics, different income classes are not associated with any change in the probability of car usage. Overall income does appear to exercise little influence over the variable “modechoice”. In fact, in all the BBNs income has been portrayed as being one of the least influential variables in the parent set of mode choice, which leads to the conclusion that the level of household and personal income alone is not enough to induce or greatly reduce the use of transit or private auto. From the decision trees for Model 1 and the binary modeling of “modechoice”, both derived from analyzing discrete trip mode choice, it appears that the only exception is for very low personal incomes, of less than \$8,000 where owning a private vehicle became objectively difficult. For all other low income instances, this

variable had little influence over the determination of transportation mode choice, a very different finding from the more established theoretical belief that income is one of the most influential socioeconomic variables when it comes to the propensity to use privately owned vehicles.

It also appears that when vehicle count is other than 0 the tendency is for an overwhelming use of private vehicles; this fact, coupled with the observation that access to a vehicle can influence the level of personal income, is powerful evidence that changing travel behavior in favor of alternative transportation modes is going to be difficult. At least this is the picture emerging from the data and the location under study.

Low residential density does not induce higher rates of car trips. This is a strong statement but is supported by the fact that the node indicating low residential densities (lowres) is not directly linked to transportation mode choice (modechoice) in Figure 4.1, nor does it appear to be linked specifically to particular transportation means because the same can be said for the BBNs of figures 4.3a, 4.4a and 4.5a, where each travel choice was more specifically modeled. This means that low density residential, as a variable, was never found to be a relevant factor influencing travel mode choice in any of the three data structures used, whether using discrete or aggregated travel mode choice data. Instead, from the analysis of the BBNs presented in Section 4.3, Building a Conceptual Model, it appears that it is the lack of commercial spaces and services that characterizes low density residential areas, which in turn affects the probabilities of each transportation choice, so that the choice is the indirect consequence of having sparse residential units.

As presented by the BBN of Figure 5.1, the nodes for commercial spaces, medium density residential, other, institutional and industrial land uses, road density and

accessibility to transit, all point towards “lowres”, and help define what it really is. By using discrete trip mode choice and land use aggregated at the tract level, the resulting decision tree shows that the more commercial land use there is, the less likely is a tract to have a large proportion of low residential densities; at the same time, we have learned that the amount of commercial space does impact on “modechoice” so that one could conclude that the higher probability of trips made by privately owned vehicles in low density residential areas is due to the lack of services within reasonable reach, and not residential density per se. These results are intuitively correct, but the algorithms used to derive them, unlike previous studies using the K2 algorithm, are not compromised by researchers’ bias.

Furthermore, economic principles state that goods and services follow their customers so it is only natural to observe more of the former where demand concentrates, such as in dense urban cores or easily accessible locations. In fact, this is particularly true in the binary BBN derived to model the choice of walking as opposed to all other alternative transportation modes, Figure 4.4a. Again, using discrete trip mode choice and land use aggregated at the tract level, it appears that the amount of commercial spaces greatly influences people’s decision to walk, while the highest probabilities for such an outcome are associated with medium and high residential densities.

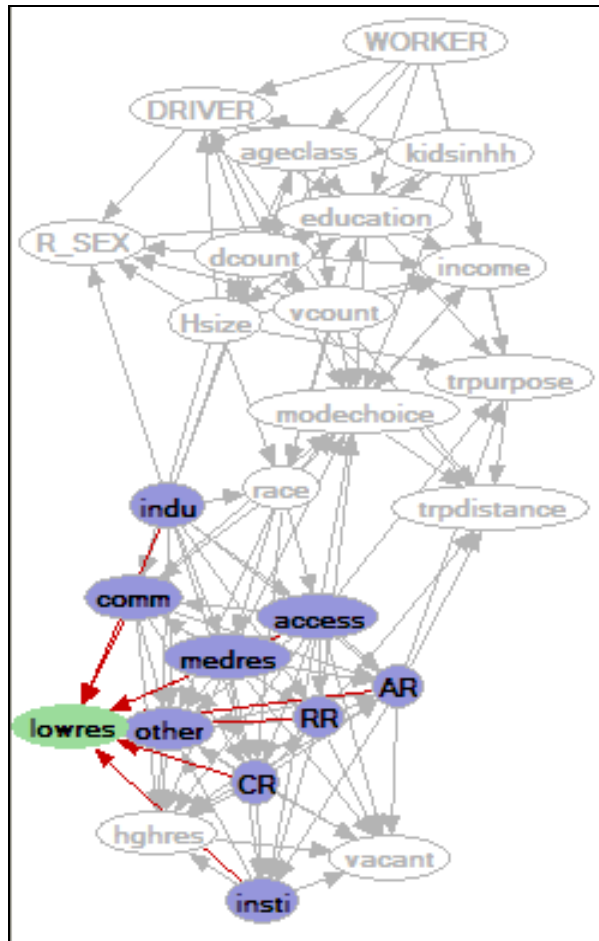


Figure 5.1: BBN for low density residential

c) Different land use types affect the type of transportation choice in unequal ways as seen in Table 5.1 which reports, in order of strength, all the variables that influence specific travel mode choices. These results are a further step into the understanding of what policy makers need to look at when trying to enhance walking and transit usage by means of altering land use characteristics.

Fore example, reading from Table 5.1, residential density enters into the binary decision model for the use of privately owned vehicles, but not into the one for walking trips, where it is the availability of commercial activities and services that has greater influence on the walking mode choice.



Auto	Walking	Transit
Vehicle count Trip distance Residential accessibility to transit Driver status Medium density residential Commercial Age Educational attainment Road density Household size High density residential Income Children in household Overall accessibility to transit Institutional Other	Vehicle count Driver status Race Overall accessibility to transit Age Trip distance Other land uses Household size Education Income Commercial Vacant land	Residential accessibility to transit Trip distance Medium density residential Vehicle count Driver status Race Commercial Children in household Income Road density Educational attainment Household size High density residential Institutional

Table 5.1: Sensitivity of specific travel mode choices to all variables

To be noted is how the amount of commercial space appears in the list as a weak variable, but that the tree associated to the binary modeling for walking found in Figure 4.4b, shows large variations in the probability of walking trips when this variable is considered as a condition for the calculation of the probability of a walking trip (p2).

These result are somewhat contrary to the theoretical notion that residential density is the key to entice people to walking more, while they also confirm that residential density is a proxy for private car usage, with lower densities experiencing greater private vehicular traffic.

d) The land use information has been aggregated to two levels of geography, US Census tracts and five digit zip codes. Zip level aggregation results in a loss of strength for the variables capturing the physical environments although they still appear in the resulting networks. The author suspects that too much variance is being lost in the land use variables

at the zip level compared to tract level aggregation, which suggests that smaller scales of aggregation should be employed.

Although the use of smaller scales of aggregation, or even better, discrete units of land uses, is also promoted by researchers such as Frank (2004), this suggestion is based not on the assumptions various researchers have made, but on the observation that the same heuristic algorithm cannot as efficiently benefit from data aggregated at large geographic levels as it does from smaller spatial units which provide more database records. In fact the key observation to be made is that the derivation of conditional probability distributions is more effective when using aggregates of travel mode choice because the input dataset has a one to one relationship connecting spatial locations, socioeconomic characteristics and travel choice. If discrete buffers of land use characteristics were to be used to match each trip origin, it is this author's opinion that the CPD for model one, derived from discrete units of analysis, would have proved more informative than the CPD for models 2 and 3 presented in this study.

The 615 discrete tracts that cover the study area provide a sufficient number of records to run a heuristic search algorithm, but the robustness of the results is somewhat compromised. In fact the best modeling results, with R square test scores of 80% or more, were found when using all 20,000 discrete trips that the National Household Travel Add-On Survey 2001 recorded for the Baltimore Metropolitan Area, with land use characteristics computed at the tract level. This means that, in the case of a perfect distribution, all tracts have at least 32 trips originating from them, a barely significant statistical level; if smaller units of aggregations had been used in this study, for example

blocks, each unit would have had even fewer trips originating from it, so that a greater number of records, in this case single trip origins, would have been needed.

The implications of the above seven points for this study's research question and working hypothesis are obvious. First, the question of whether there are links between land use and travel behavior is answered by the algorithm used here. This has found and quantified these links. If such relationships did not exist, the heuristic approach used in the study would be an appropriate approach to establish that fact as well.

More importantly, other implications for applied research follow from the study's working hypothesis. If one researcher were to deductively investigate the same problem domain, using the same data and assuming that income and trip purpose are significant determinants of trip mode choice, the results would be different from those obtained by another researcher who developed a different statistical model with a different choice of determinants.

A proper reinterpretation of the study results, following the dag of Figure 4.1 in a counterclockwise fashion, and based on the other numerical outputs, would suggest that a researcher approaching the same problem would now know that: a) the data do contain a relationship between travel mode choice and land use characteristics, b) worker status, household size, income and gender are not as strong determinants of mode choice as theoretically implied, c) trip distance should be modeled as a feedback loop, d) trip purpose as a distinction between work and non work travel is not a very effective variable and perhaps it should be included with a more meaningful classification scheme such as work, shopping, elective travel etc., e) although the influence of land use on travel behavior is

normally secondary to land use characteristics – e.g. availability of vehicles is always such a strong node in the networks – there are some instances, such as the inclusion of a mix of residential and commercial land uses, that affect the percentage shifts of travel mode choice in a significant way.

Furthermore, for those researchers focusing on a specific travel mode choice, for example transit usage, the dag of reference should be the one of Figure 4.5a and its associated tree in Figure 4.5b, where the heuristic evidence emerges with information specific to what socioeconomic and land use variables affect that particular travel mode.

Overall, running a quick inductive heuristic algorithm thorough the data, prior to the design of a conceptual model for more traditional deductive studies, yields some valuable information that allows us to know whether the data at hand do indeed contain the information needed to answer our research hypothesis, and can guide us towards the formulation of a more rigorous research hypothesis and study design. Oftentimes the biggest disappointment from research comes from discovering that the proposed hypothetical relationships under study are in fact not present in the data, or that they were modeled in an erroneous fashion and, worst of all, it is not until well after the research has been carried out that these problems emerge. Our theoretical background related to a problem domain is a valuable asset but it is not always applicable. A heuristically derived conceptual framework at the outset of any analysis integrates previous theoretical knowledge and suggests which ways it should be adjusted, if need be, so that our statistical constructs can better model the data used in the analysis.

This study therefore proves that a more inductive approach to the analysis of the relationships between land use and transportation behavior provides very valuable

knowledge, and can be used to better formulate our deductive studies and tailor them to the peculiarities of the data at hand, even if our theoretical knowledge of a particular process would say otherwise. It helps us to question the principles and theories we use to model phenomena and to provide *reasoned* alternative points of view that might otherwise be discarded as implausible, when in fact they may hide subtle complexities that could better explain travel behavior and the influences of land use. With the escalating costs of infrastructure development and their unintended negative consequences, both historical and current, we can ill afford to dismiss these alternative explanations of such phenomena.

## **5.2 Major Study Limitations and its Overall Applicability**

The most significant limitations of this study have their roots in the methodology, in particular its reliance on data to drive the analysis. This problem is typical of many bottom up scientific approaches and it can be summarized with the fact that a) this study can be thought to be only as good as the data used in it and b) the applicability of the relationships between land use and travel behavior, as learned from the Baltimore Metropolitan area, are difficult to meaningfully generalize to other cities, although the current findings are remarkably close to what would be expected given the existing theoretical knowledge.

With respect to point “a”, a series of internal and external validity tests were done to compare the local NHTS data with national averages but the data is still a small sample of the entire population and as such it should be treated with caution. In relation to point “b” it should be noted that these data driven inductive approaches, as well as data mining in general, can be scaled down. In fact if a study is to be conducted at the neighborhood level,

the land use travel behavior relationships can be learned from the entire surrounding region, thereby providing background information for the more localized study.

A third shortcoming is also typical of all heuristic approaches, that is, they all require large amounts of data. The results of Model 1 in this study, based on discrete trip origins and the surrounding land use characteristics aggregated at the tract level, did match the test dataset to a large degree, 85%. The use of larger datasets might improve the overall fitness of the model against a comparable test set, but the usefulness of using larger databases, with hundreds of thousands of records, lies in the algorithm's ability to better resolve the weakest relationships for which this study did not produce enough evidence.

The fourth and most relevant shortcoming, in the author's opinion, is due to the challenges researchers face when interpreting the results of such network or structure learning algorithms. In these cases, in-depth knowledge of the methodology and its assumptions might still not be enough for every researcher to reach the same interpretation or the same set of conclusions. On the other hand the results of such heuristic search exercises can be used as an unbiased platform for discussion and collective interpretation in all those instances, such as the study of travel behavior, where the complexity of the theoretical domain is too large to be uniquely and fully resolved by many distinctive and partially complete ad-hoc studies.

A fifth limitation that is specific to this study is identified in the deliberate use of a limited set of variables. As has been explained earlier, doubt existed about the possibility of a successful outcome from such a study, especially given the use of a relatively new and untested algorithm such as the one by Chickering et al (1997), Torres and Huber (2003), Jenssen et al (2003, 2004) all had the means to develop their own software platforms and

implemented specific strategies aimed at the reduction of complexity in the networks they generated, a possibility that was clearly beyond this study. Hence, the partial disappointment of having successfully applied the Bayesian Belief Network approach to the study of travel behavior and land use, while having to contend with a set of variables that are surely relevant to such a study but that are by no means complete, thus making the result interpretation even more challenging.

The point above touches on a related problem about heuristic search exercises. In beginning the analysis with the use of existing data, researchers might wonder what to include as relevant information and what to exclude, although this is less of a problem in this study because of the existence of so much literature on the subject. Technically there are no limits to the amount of data that can be analyzed, but this depends on accepting long computing times and complex outputs to interpret. Furthermore, the inductive search approach is used to assess problem domains where nothing is known a priori about them, and this is achieved by using specialized algorithms that look for missing variables that can potentially explain otherwise inexplicable associations. This study makes no use of such complex techniques but it can be seen that the inductive reasoning approach can be used as only one of many tools and techniques in the much wider context of research, and it can be applied as a complementary technique to more traditional deductive reasoning.

Lastly, the calculations for the probability of a given variable class to take place based on the conditions of other variables should be looked at carefully. When deriving Conditional Probability Distributions the algorithm made did not differentiate between land use variables – which are spatial in nature – and other variables. The results are CPDs with spurious conditional probabilities of land use mixes that do not take place on the

ground, so the use of spatial variables should be restricted to combinations these that make logical sense, as in the case of cumulative land use percentages. This could be achieved by making sure that a given set of conditions based on land use variables should not exceed 100% of the spatial coverage for a every tract or zip.

Also, the fact that so many probabilities are similar to each other in the CPDs presented in this study is a direct consequence of using discrete classes in the input data. A final issue related to the calculation of CPDs is that not all the parent sets of variables are calculated concurrently, which forces the analyst to compare different conditional probability distributions; this is a software issue rather than a theoretical one and the use of other, more flexible software platforms, such as Bayesnet Toolbox by Kevin Murphy (<http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html>) in future analyses may prove to be fruitful.



## **Chapter 6: Conclusion**

This study expands on the work by Torres and Huber (Torres and Huber 2003) and by Jenssen et al. (2003, 2004) by adding land use variables and a new search algorithm. It has found that:

- The heuristic search approach to derive relationships between land use and travel behavior does work.
- This technique needs to be fine tuned for the proper use of spatially explicit data.
- The research outputs are an unbiased representation of the land use travel behavior relationships but they need proper interpretation, especially in light of persisting theoretical questions still driving this research field.
- Not all the assumed relationships most often cited in the travel behavior theory hold true in this particular dataset, and that this methodology provides the means to find such discrepancies before more detailed studies are carried out.
- Interpretation of the results forces researchers to try explaining why and where these differences exist. While some commonly held assumptions about travel behavior and physical characteristics are in fact correct, the study also found new evidence suggesting alternative explanations for the influence of land use (residential density and commercial space in particular) and household socioeconomic characteristics (income and children in household) on the decision process related to travel mode choice.
- Land use parameters should be measured at large spatial scales.

### **6.1 Possible Benefits from a Heuristic Search Approach**

Can an inductive approach produce a reliable representation of the relationships between land use and travel behavior? This study offers an insight into what this methodology can deliver for planning research by identifying its strength and weaknesses. The creation of Bayesian Belief Networks provides the analyst with a model of the relationships among variables under study that is derived by means of meta-heuristic search methods. No statistical model needs to be specified a priori and there is no need to characterize variables as independent or dependent. It provides quantitative assessments of the occurrences of specific outcomes based on the status of all other variables, and it allows for the study of complex problems based on how the data capture them. Compared with existing theoretical models, and the information presented above, the analysis either confirms or questions our beliefs. However this is not all that the method is potentially capable of. Rather, it is the beginning of a much larger and more complex research methodology which uses BBNs as the first step to compute inferential statistics on conceptual models that are much more complex and complete than any researcher could possibly organize deductively or without external help. The derivation of elasticities is the next step, along with a comparison to other studies, which try to organize and understand the same set of relationships.

In the analysis of the travel diary data from the National Household Travel Survey 2001, each trip is considered unique and is characterized by the land use conditions of the tract or zip it originates from. This assumption implies that each trip is treated independently of all other trips, even if some were originally taken as a part of a trip chain; in practice the algorithm used in this analysis treated trips as discrete separate events, which is not always the case, as when multiple trips are made by the same person. This type of analysis is not necessarily based on the best assumptions but, as mentioned above,

a) future analysis can be undertaken with the trip chains being explicitly considered as such and b) the aim in this research was to focus on the land use conditions underlying the decision to use a particular trip mode, even as we recognize that there are interdependencies between sequential trips and their modal choices.

The biggest concern about the appropriateness of the discussion in the previous chapter however lies in the very strong reliance, in the Bayesian heuristic method, on the use of location specific data. That is, the networks that have been derived and upon which I have based my conclusions are based on a ground “truth” that is as good, or true, as the measurements used to capture it. Should the measurements be repeated elsewhere or even in the same location at a different time, or should a different set of households be included in the survey, we could obtain different representations of the relationships linking land use and transportation behavior. This is a common problem with all modeling attempts that too closely fit to the data at hand; these models are often well calibrated and tested but perform poorly when used for forecasting purposes.

This study is not about forecasting but the method used does have this potential. In this case the data used to derive relationships should be collected from a different and much larger area than the local data used to instantiate the network when using Bayesian inference for forecasting purposes. With such an approach, analysts could be sure that the overall representation of relationships is tested and appropriate for a given area, whereas the measurements used for forecasting apply only to the local area; bigger confidence intervals can be expected but the problem of over fitting the data is therefore avoided.

The Bayesian inferential method is not used in this study but it would help address some unanswered questions. For example, as presented so far it appears that at least in

Baltimore the variable for race and ethnicity, specifically African American (class 2), is a strong determinant of the variable mode choice, and the decision trees allow the analyst to learn about what combination of variables and their instances affect or change the likelihood of driving a car, walking or taking transit. Yet the trees hardly map the case for race and ethnicity other than for African American, so there is little to be learned for Whites (1), Hispanics (3) and other races (4). To investigate the relationships among the variables and these particular instances of “race” it is possible to use Bayesian inferential statistics, where the researcher can declare, or instantiate, the status of a specific variable and evaluate the available evidence to infer the relationship of other variables with the particular instance of the variable under consideration. By the same token, this study would greatly benefit from knowing what specific effects the number of vehicles have on “modechoice”, other than the one presented in the decision trees. In fact one platform to run this type of analysis is Microsoft MSNBX, a software package that predates WinMine that has only recently been made compatible with it.

While the resulting Bayesian networks in this study need not be taken as *the* map of land use travel behavior relationships, but as their possible representations based on the measurements and data at hand, for those planners developing heuristic inductive approaches the biggest prospects for research lie even further. The Bayesian approach can be developed as a tool for experimentation without the typical requirements of control groups, with intervention’s effects determined solely on the proper detection of the resulting changes. Unlike health practitioners who can administer a real medicine to a group of patients, and a placebo to their control group, planners do not have cities and their societies under control. However, cause and effect relationships can be examined as the

consequences of intervention in Bayesian inference, once basic causal explanation are learned by means of Bayesian structure learning, as this study has done.

## Appendix A: Bayesian Statistics

### Reasoning Under Uncertainty

The basic problem when reasoning under uncertainty is whether information on some event influences our belief in other events. As a comparison, rule-based systems cannot capture reasoning under uncertainty because the dependence between events changes with the knowledge of other events. The rules are fixed and cannot be changed when the necessity arises. The following example, from Pearl (1988), is provided to illustrate this point:

Leaving his house in the morning, Mr. Laurel notices that his grass is wet. “It must have rained last night” he reasons; then he looks into his neighbor Mr. Hardy’s garden to see if that is also wet to confirm his thinking.

Notice: the information that Mr. Laurel’s grass is wet has an influence on his belief of the status of Mr. Hardy’s grass, which was actually dry.

Mr. Laurel then checks his rain meter, and finds it to be totally dry. “It cannot have rained”, he thinks.

Notice: Mr. Laurel is now considering the information about Mr. Hardy’s garden to be irrelevant to his belief of his grass being wet.

What, beside rain, could be responsible for the wet lawn, Mr. Laurel ponders? “I might have forgotten to turn the sprinkler off last night, that’s why the lawn is wet”.

But, suppose that the next morning, Mr. Laurel notices once again that his grass is wet. Mr. Laurel’s belief of both rain and sprinkler increases. Then he observes that Mr. Hardy’s grass is wet, and he concludes that it had rained last night.

The last step in the above line of thinking is very difficult, almost impossible, to implement through decision rules, yet very natural for human beings, and it is called *explaining away*.

Explaining away is defined as the process of decreasing one’s belief in a causal event because of a concurrent increase in the belief of an alternative causal event. In the reasoning above, Mr. Laurel concluded that it had rained after he saw Mr. Hardy’s wet grass the next morning. Thus, it follows that Mr. Laurel’s wet grass has been explained by the rain. Consequently he no longer has any reason to believe that he forgot the sprinkler on the first night, regardless of the status of the rain gauge. Explaining away is another example of dependence changing with the information available (Jensen 1996).

Two events are considered to be dependent on each other when the probability of an event depends on the knowledge of the other event (Leonard and Hsu 1999:76). Once again, in the reasoning above, when Mr. Laurel knows nothing about the initial state of the problem, the variables “rain” and “sprinkler” are *independent*. This changes when the

information on Mr. Laurel's grass is updated so that the variables "rain" and "sprinkler" are now *dependent*.

Thus, a change in the belief on whether it rained or not, will change the belief of whether the sprinkler was on or off. If it rained, then the sprinkler was off, otherwise, the sprinkler was on. Of course, all these conditions are true only if there are no other possible variables that could cause Mr. Laurel's grass to be wet. Yet if we have no information on the condition of Mr. Laurel's grass, then we cannot relate the variables "rain" and "sprinkler". Dependence between events is the fundamental building block of causal networks, which conceptualize the information on a given status as prior certainties.

In the above example, it is obvious that if an event is known, the certainty about the other events must be changed. This fact is reflected in certainty calculus, where if the actual certainty of a specific event has to be calculated, the knowledge of certainties prior to any information is also required. For example, the certainty of the event rain is still very much dependent on whether rain at night is rare (as in Los Angeles) or very common (as in London), given that Mr. Laurel's grass is wet (Pearl 1988).

### Causal Networks

With the basic principles of reasoning under certainty and the notion of dependent events outlined, causal networks can be introduced. The thinking process followed by Mr. Laurel can, in fact, be described by a graph. In figure A.1 the outcomes are represented as nodes, and the two nodes *A* and *B* are connected by a directed link from *A* to *B* if *A* has a causal impact on *B*. Figure three is a graphical model of Mr. Laurel's small world of wet grass.

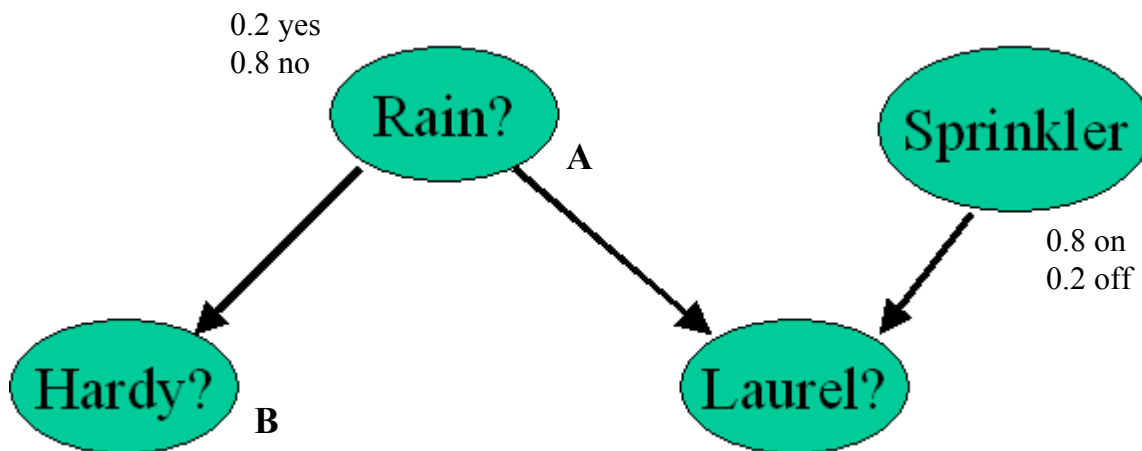


Figure A.1: A graphical model for the wet grass example (Adapted from Jensen 1996).

Figure A.1 is also an example of a causal network; a *causal network* is a graphical arrangement of a set of nodes representing *variables* and a set of *directed links* between variables. The variables symbolize events and, in mathematics, this arrangement is called a *directed graph*. The relationships between variables in a directed graph are described by using the terminology of family relations. If there exists a link from variable *A* to variable *B*, then *A* is called a *parent* of *B* and *B* is called a *child* of *A*. Every variable in a causal

network has two or more states (i.e. the color of a car: blue, green, red, and black) and in general, variables can have continuous and discrete states.

Attached to the directed graph of Figure A.1 is also a quantitative part derived from the calculation and combination of *certainty numbers* (Pearl 1988), in this case the probabilities historically derived for the event rain and sprinkler. The certainty numbers are the prior probabilities of the event (variables) given the data. From the graph in figure A.1, one can read about the dependencies and independencies in the small world of wet grass. If we have already learned that it has not rained, then information on Mr. Hardy's grass has no influence on Mr. Laurel's grass. All possible ways in which influence and causality may run between variables in a causal network have been analyzed by Pearl (1986) and Verna (1987).

Two variables are said to be *separated* if new evidence gathered about one of them bears no impact on our belief about the other. If the state of a variable is known, then we say it is *instantiated*. There are three types of connections in a causal network: *serial*, *diverging*, and *converging* connections. Figure A.2 shows all types of connections in a causal network:

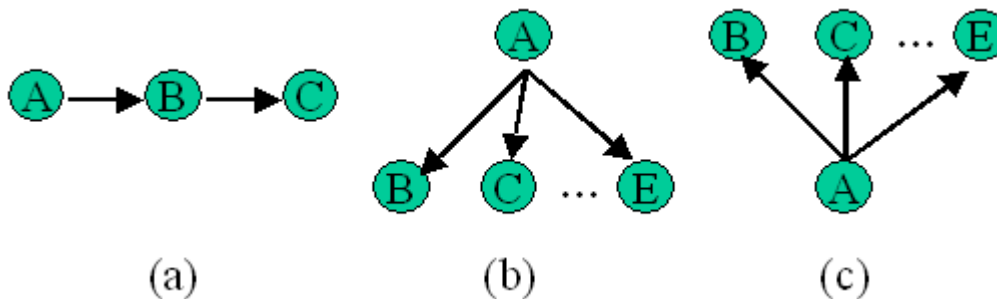


Figure A.2: Serial, diverging, and converging connections respectively.

In Figure A.2 (a) the variable  $A$ , the parent, has a control on the variable  $B$ , the child, that then has control over variable  $C$ , a child too. Apparently, the evidence about the variable  $A$  will affect the certainty of the variable  $B$ , that in turn affects the certainty of the variable  $C$ . Analogously, the evidence on the variable  $C$  will affect the certainty of the variable  $A$  through the variable  $B$ . On the other hand, if the state of the variable  $B$  is given, that is we are completely sure about its state, then the link is blocked, and the variable  $A$  and the variable  $C$  become independent because, no matter what  $A$  is, we already know  $B$  and therefore  $B$ 's influence on  $C$ . In other words, influence may run from  $A$  to  $C$  and vice versa unless  $B$  is instantiated.

In Figure A.2 (b), the case of a diverging connection shows that the influence can pass to all the children of the variable  $A$ , unless the state of the variable  $A$  is given. If the state of the variable  $A$  is known, then the variables  $B, C, \dots, E$  become independent from each other. Therefore, influence may run between  $A$ 's children unless  $A$  is instantiated.

Figure A.2 (c), shows a converging connection where, if there is nothing known about the variable  $A$ , other than what may be deduced from the knowledge of its parents  $B, C, \dots, E$ , then the parents are said to be independent. The independence means that evidence on one of the parents has no effect on the certainty of the others. Yet if there is any other kind of evidence influencing the variable  $A$ , then the parents become dependent because of the principle of explaining away. It follows that evidence may only be transmitted through a



converging connection if either the variable in the connection or one of its descendants has received evidence. The evidence can be direct evidence on the variable  $A$ , or it can be evidence from one of its children. In causal networks, this fact is called *conditional dependence*.

The three cases explained above consider all the possible forms in which evidence may be transmitted through a variable. By following the rule given below, it is possible to assess whether any pair of variables in a causal network are dependent, given the evidence entered into the network. Two variables  $A$  and  $B$  are said to be *d-separated* if for all paths between variables  $A$  and  $B$ , there is an intermediate variable  $V$  so that either:

- the connection is serial or diverging and the state of  $V$  is known

or

- the connection is converging and neither  $V$  nor any of  $V$ 's descendants have received evidence (Pearl 1988).

If variables  $A$  and  $B$  are not d-separated they are said to be *d-connected*.

For example, if the state of the variable  $B$  is given in Figure A.2 (a), then the link is blocked, and the variable  $A$  and the variable  $C$  become independent. Therefore, it is said that the variable  $A$  and the variable  $C$  are d-separated given the variable  $B$ . Similarly, in Figure A.1, “sprinkler” and “Hardy” are d-separated because the connecting trail is converging around the variable “Laurel”.

One should note that d-separation is a property of human reasoning (Jensen 1996), and therefore any calculus for uncertainty in causal structures must obey the principle that whenever  $A$  and  $B$  are d-separated then new information on one of them does not change the certainty of the other.

Finally, another important concept in causal networks is the conditional independence between variables. In the Bayesian calculus, the blocking of influence between variables is reflected in the concept of *conditional independence*. The variables  $A$  and  $C$  are independent given the variable  $B$  if  $P(A | B) = P(A | B, C)$ . In the case of serial and diverging connections, this expresses that if the state of the variable  $B$  is given then no information of the variable  $C$  will change the probability of the variable  $A$ .

### Probability Calculus

The quantitative part of the certainty assessment carried out by assessing d-separation as in the preceding section is based on Bayesian calculus, which is *classical probability calculus*. The basic concept in the Bayesian treatment of certainties in causal networks is the notion of *conditional probability*. When the probability of an event  $A$ ,  $P(A)$ , is known, then it is a probability which is conditioned by other known factors, say  $B$ , and it has the following form:

*Given the event  $B$ , the probability of the event  $A$  is  $x$ .*

The mathematical notation for this statement is  $P(A | B) = x$ . This does not mean that whenever  $B$  is true, then the probability for  $A$  is  $x$ . Rather it means that if  $B$  is true, and everything else known is not applicable to  $A$ , then  $P(A | B) = x$ .

The *fundamental rule* for probability calculus is given in the following way in Pearl (1988):

$$P(A | B)P(B) = P(A, B)$$

where  $P(A, B)$  is the probability of the joint event  $A \ni B$ . Because probabilities ought always to be conditioned by a context  $C$ , the formula should be written as;

$$P(A | B, C)P(B | C) = P(A, B | C)$$

It follows that  $P(A | B)P(B) = P(B | A)P(A)$  and this gives the famous *Bayes' rule*:

$$P(B | A) = \frac{P(A | B)P(B)}{P(A)}$$

And, by conditioning the Bayes' rule:

$$P(B | A, C) = \frac{P(A | B, C)P(B | C)}{P(A | C)}$$

By considering  $A$  as a variable in a causal network with the set of states  $a_1, a_2, \dots, a_n$ , then the  $P(A)$  is a probability distribution over this set of states:

$$P(A) = (x_1, x_2, \dots, x_n) \quad x_i \geq 0 \quad \sum_{i=1}^n x_i = 1$$

where  $x_i$  is the probability of  $A$  being in the state  $a_i$ .

Now, if the probability of  $A$  being in the state  $a_i$  is expressed as  $P(A = a_i)$  and if  $B$  is another variable with the states  $b_1, b_2, \dots, b_m$ , then  $P(A | B)$  is an  $n$ -by- $m$  table consisting of numbers  $P(a_i | b_j)$ . This table is called a conditional probability table (CPT) or conditional probability distribution (CPD) for  $P(A | B)$ .

The joint probability for the variables  $A$  and  $B$ ,  $P(A, B)$ , is also an  $n$ -by- $m$  table containing the probabilities  $P(a_i, b_j)$ . The joint probabilities,  $P(A, B)$ , can be computed by utilizing the fundamental rule of equation 1:

$$P(a_i, b_j) = P(a_i | b_j)P(b_j)$$

or equivalently,

$$P(A, B) = P(A | B)P(B)$$

The joint probability,  $P(A, B)$ , has  $n*m$  entries. The probability  $P(A)$ , can be computed from the table  $P(A, B)$ . With  $a_i$  denoting a state of the variable  $A$ , the table  $P(A, B)$ , has  $m$  different events for which the variable  $A$  is in state  $a_i$ , namely the mutually exclusive events  $(a_i, b_1), \dots, (a_i, b_m)$ .  $P(a_i)$  can then be calculated as;

$$P(a_i) = \sum_{j=1}^m P(a_i, b_j)$$

This operation is called *marginalization* and it is said that the variable  $B$  is marginalized out of  $P(A, B)$  (producing  $P(A)$ ). Thus, the notation can be written as follows:

$$P(A) = \sum_B P(A, B)$$

All the definitions above work only for classical (objective) probabilities but causal networks have yet another type of probability, called subjective probability. Subjective probability is one of the important features of causal networks and by association, important in Bayesian Belief Networks, because of its ability to explain one's belief of an event.

### **Subjective Probabilities**

Probability calculus does not require that the probabilities be based on frequencies of repeated experiments. In fact probabilities may also be completely subjective estimates of the certainty of an event. Subjective probability is also called Bayesian probability or personal probability in the literature and it is the probability of an event  $x$  based upon a person's *degree of belief* in that event. A Bayesian probability is a function of the person who assigns the probability (e.g., someone's degree of belief that a coin will land heads), whereas a classical or objective probability is a physical property of the world (e.g., the probability that a coin will land heads given a frequency of repeated experiments). Much work has been done about the shortcomings of relying solely upon personal probabilities, and as seen in Chapter 3, this study uses heuristic methods to replace personal probabilities with probabilities learned from data.

### **Bayesian Networks**

Bayesian Networks are based on causal networks as defined above. However they also make use of other theorems and provide researchers with the added notion of *strength*. This gives causal relations a quantitative side, expressed by attaching likelihoods to the links.

With the variable  $A$  being a parent of the variable  $B$  in a causal network, and by using probability calculus, the conditional probability,  $P(B | A)$ , can be thought of as the strength of the link between these variables. On the other hand, if the variable  $C$  is also a parent of the variable  $B$ , then conditional probabilities  $P(B | A)$  and  $P(B | C)$  do not provide any information on how impacts from the variable  $A$  and the variable  $B$  interact. They may cooperate or counteract in various ways so that the specification of  $P(B | A, C)$  is also required.

It may happen that the domain to be modeled contains feedback cycles. Feedback cycles are difficult to model quantitatively. For causal networks, no calculus coping with feedback cycles has been developed, although Jansen (2001) has suggested that differential calculus is theoretically capable of solving this problem. For the time being however, the use of Bayesian Belief Networks is limited to those cases that do not contain cycles, and directed graphs are often qualified as "directed acyclic graphs" – or dags.

A Bayesian network consists of the following elements:

- A set of *variables* and a set of *directed links* between variables,
- The variables coupled with the directed links construct a *directed acyclic graph* (dag),
- Each variable  $A$  with parents  $B_1, B_2, \dots, B_n$  has a conditional probability table

$P(A | B_1, B_2, \dots, B_n)$  associated with it (Pearl 1988).

For the graph in Figure A.3, the prior probabilities  $P(A)$  and  $P(B)$  have to be specified.

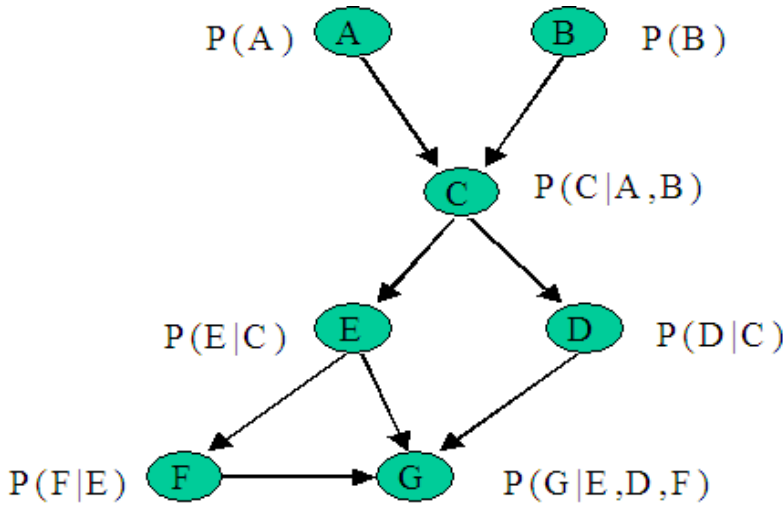


Figure A.3: A directed acyclic graph. The probabilities to specify are shown.

Prior probabilities are essential not for mathematical reasons but because prior certainty assessments are an integral part of human reasoning.

One of the benefits of Bayesian networks is that they admit d-separation. If the variables  $A$  and  $B$  are d-separated in a Bayesian network with evidence  $e$  inserted, then  $P(A | B, e) = P(A | e)$ . Therefore, d-separation can be used to explain away conditional independencies.

A basic example of a fully developed BBN, a converging one, is the one in the next diagram and Figure A.4, where event  $C$  can be affected by events  $A$  and  $B$ :

Assume the following probabilities for  $A$ :

True	False
$p(A) = 0.1$	$p(\sim A) = 0.9$

And  $B$ :

True	False
$p(B) = 0.4$	$p(\sim B) = 0.6$

And  $C$ :

$A$	True		False	
$B$	True	False	True	False
True	$p(C   AB) = 0.8$	$p(C   A\sim B) = 0.6$	$P(C   \sim AB) = 0.5$	$p(C   \sim A\sim B) = 0.5$
False	$p(\sim C   AB) = 0.2$	$p(\sim C   A\sim B) = 0.4$	$P(\sim C   \sim AB) = 0.5$	$p(\sim C   \sim A\sim B) = 0.5$

Organized in the BBN of Figure A.4.

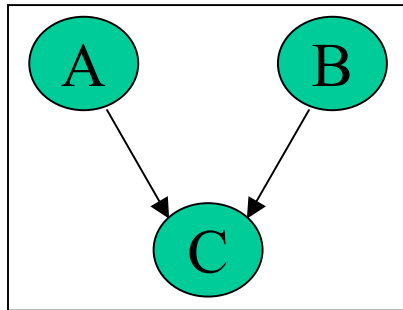


Figure A.4: Convergent Bayesian Belief Network with numerically complete example.

Using the known probabilities, it is possible to calculate the 'initialized' probability of  $C$ , by summing the various combinations in which  $C$  is true, and breaking those probabilities down into known probabilities which are:

$p(C) =$	$p(CAB) + p(C\sim AB) + p(CA\sim B) + p(C\sim A\sim B)$	
$=$	$p(C   AB) * p(AB) + p(C   \sim AB) * p(\sim AB) + p(C   A\sim B) * p(A\sim B) + p(C   \sim A\sim B) * p(\sim A\sim B)$	$0.8 * (0.1 * 0.4) + 0.5 * (0.9 * 0.4) + 0.6 * (0.1 * 0.6) + 0.5 * (0.9 * 0.6)$
$=$	$p(C   AB) * p(A) * p(B) + p(C   \sim AB) * p(\sim A) * p(B) + p(C   A\sim B) * p(A) * p(\sim B) + p(C   \sim A\sim B) * p(\sim A) * p(\sim B)$	$0.032 + 0.180 + 0.036 + 0.27$
$=$	$0.518$	$0.518$

As a result of the conditional probabilities,  $C$  has a 0.518 chance of being true in the absence of any other evidence. Inference is hence based on calculating the revised probabilities, that is, if for example  $C$  is true, the "revised" probabilities of  $A$  or  $B$  being true (and therefore the chances that they caused  $C$  to be true) can be inferred by using Bayes Theorem with the initialized probability:

$p(B   C) =$	$(p(C   B) * p(B)) / p(C)$
$=$	$((p(C   AB) * p(A) + p(C   \sim AB) * p(\sim A)) * p(B)) / p(C)$
$=$	$((0.8 * 0.1 + 0.5 * 0.9) * 0.4) / 0.518$
$=$	$0.409$

$p(A   C)$	$(p(C   A) * p(A)) / p(C)$
=	$((p(C   AB) * p(B) + p(C   A\sim B) * p(\sim B)) * p(A)) / p(C)$
=	$((0.8 * 0.4 + 0.6 * 0.6) * 0.1) / 0.518$
=	0.131

The conclusion is that given  $C$  is true,  $B$  is more likely to be the cause than  $A$ . Note that in this case, when dependencies converge, there may be several conditional probabilities to derive although some can be calculated from others because the probabilities for each state should sum to 1.

**Appendix B: Local dags for the parent set of travel mode choice**

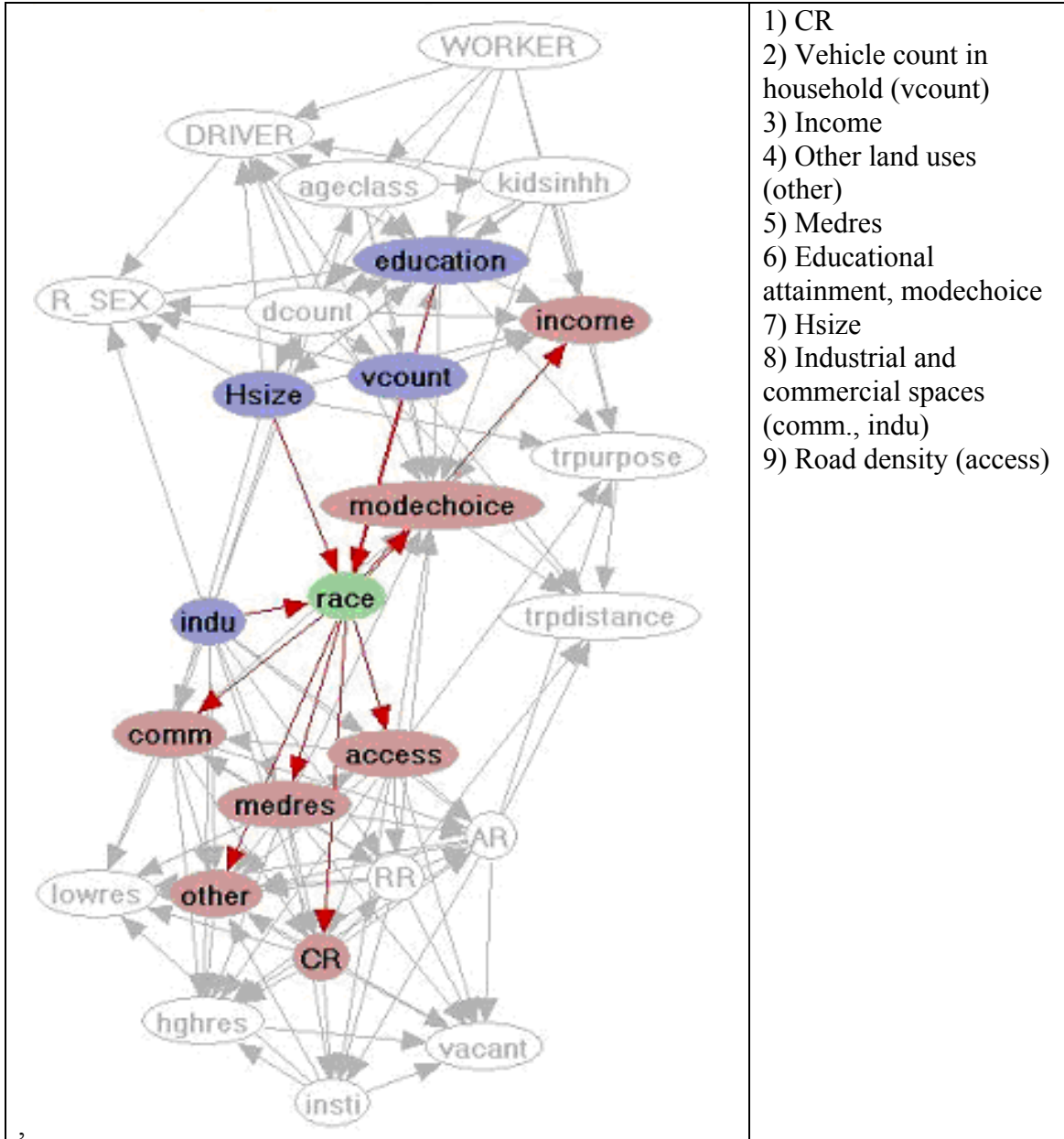


Figure B.1: BBN for race

Figure 4.6a, 4.6b: In the case of commercial and service areas, one could argue that the various land use variables as well as transportation variables (highres, lowres, other, AR, CR, RR) are not really influenced by the amount of commercial spaces, but rather, commercial entities locate in these areas because of their need to have access to customers, and customers prefer having these services close at hand. The directed links showing cause and effects would imply that the opposite case is true, but the network is still valuable in having correctly found that there is a relationship between overall accessibility to commercial spaces and what kind of other land uses commercial entities prefer in terms of

where they would want to locate. As for the effects of “comm.” on “modechoice” it appears as if the selection of auto, transit or walking trips is indeed influenced by the amount of commercial land use available at the origin of the trip. However this relationship is weak, except in the case of trips made on foot as seen above in the discussion about modeling walking mode choice as a binary choice.

The BBN for the variable race and ethnicity (BBN above right) shows that land use variables are influenced by what race/ethnicity is being considered. The author would argue that it is the large segregation of homogeneous cohorts in the population within the metropolitan area of Baltimore that might explain these relationships; in fact a more reasonable explanation would see the causal link to be wrongly directed again, given that it is more likely that minorities concentrate in downtown areas more than other socioeconomic groups and that they do so because of the local land use conditions (better accessibility to goods and services), cheaper rents etc. Once again, however, the fact remains that the BBN has correctly picked these relationships.



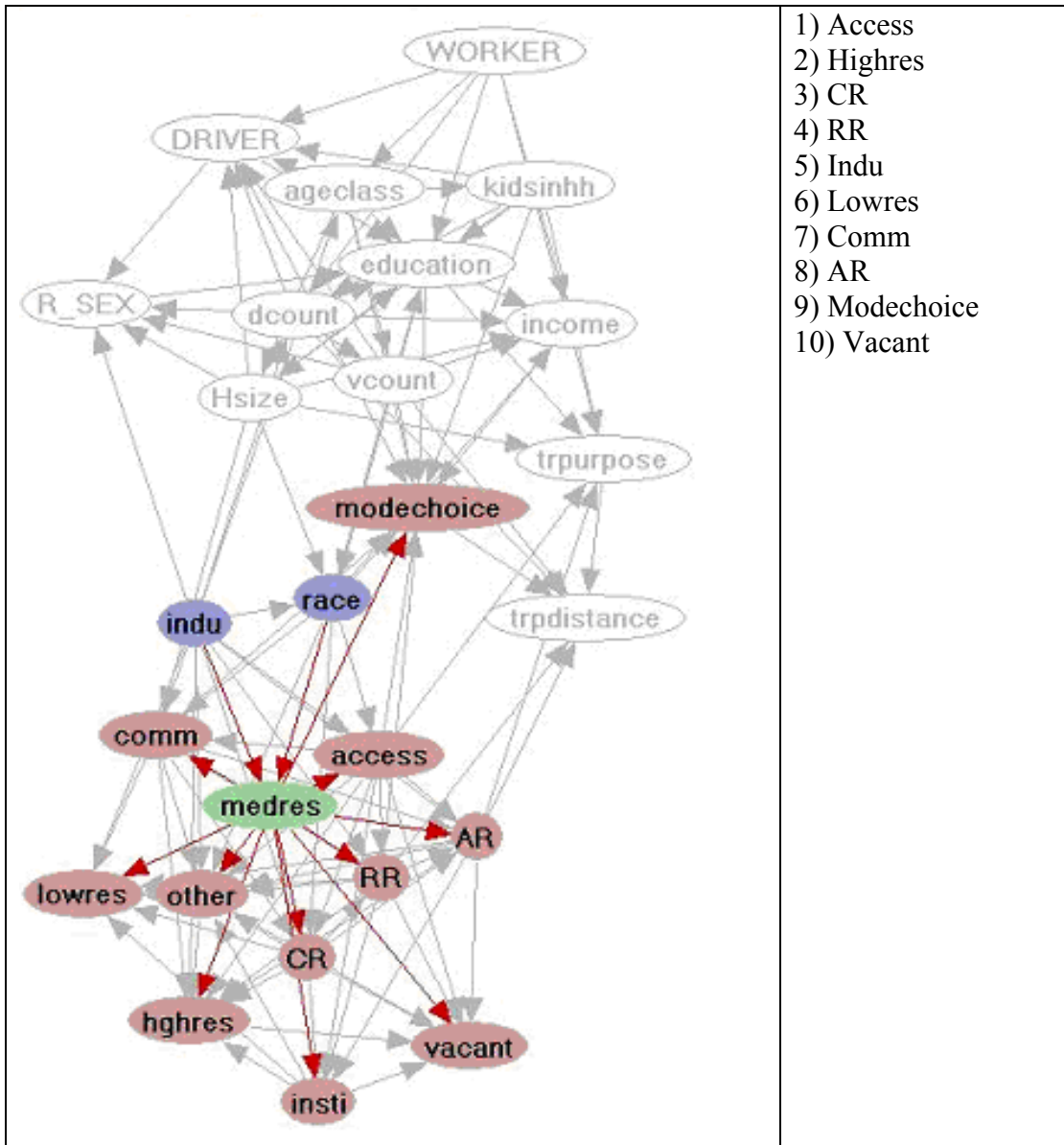


Figure B.2: BBN for medium density residential

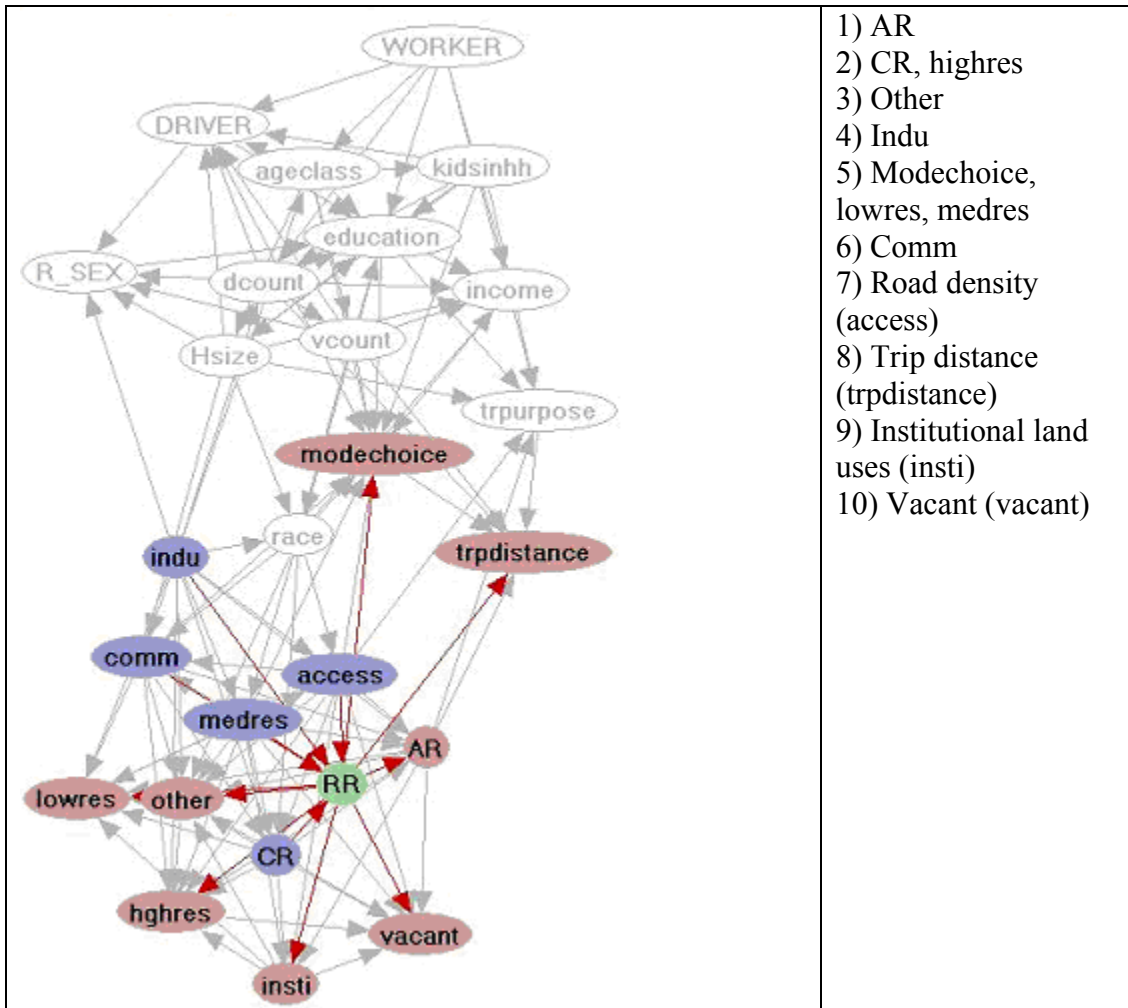


Figure B.3: BBN for accessibility to transit for residential areas

B.2, 4.7b: The BBNs for the land use variables medium density residential and residential accessibility to transit (RR) are more interesting in terms of their relationships to socioeconomic variables. Medres has only one, race and ethnicity, probably because of the underlying factors mentioned above about the same node. Medium density residential occurs, in the sample, together with large percentages of high residential densities, and it generally has good accessibility to transit (RR). This last node/variable has no connection to socioeconomic variables of any sort. The node is only characterized by its association with other land use characteristics so that transit accessibility does not change according to economic status or race. This is interesting because, given the large concentration of homogeneous socioeconomic groups in downtown hinted at by the relationship among the other variables, it would have been possible to argue that transit systems were more accessible to, and better serving, these specific population cohorts, and not all based on income and other socioeconomic differences.

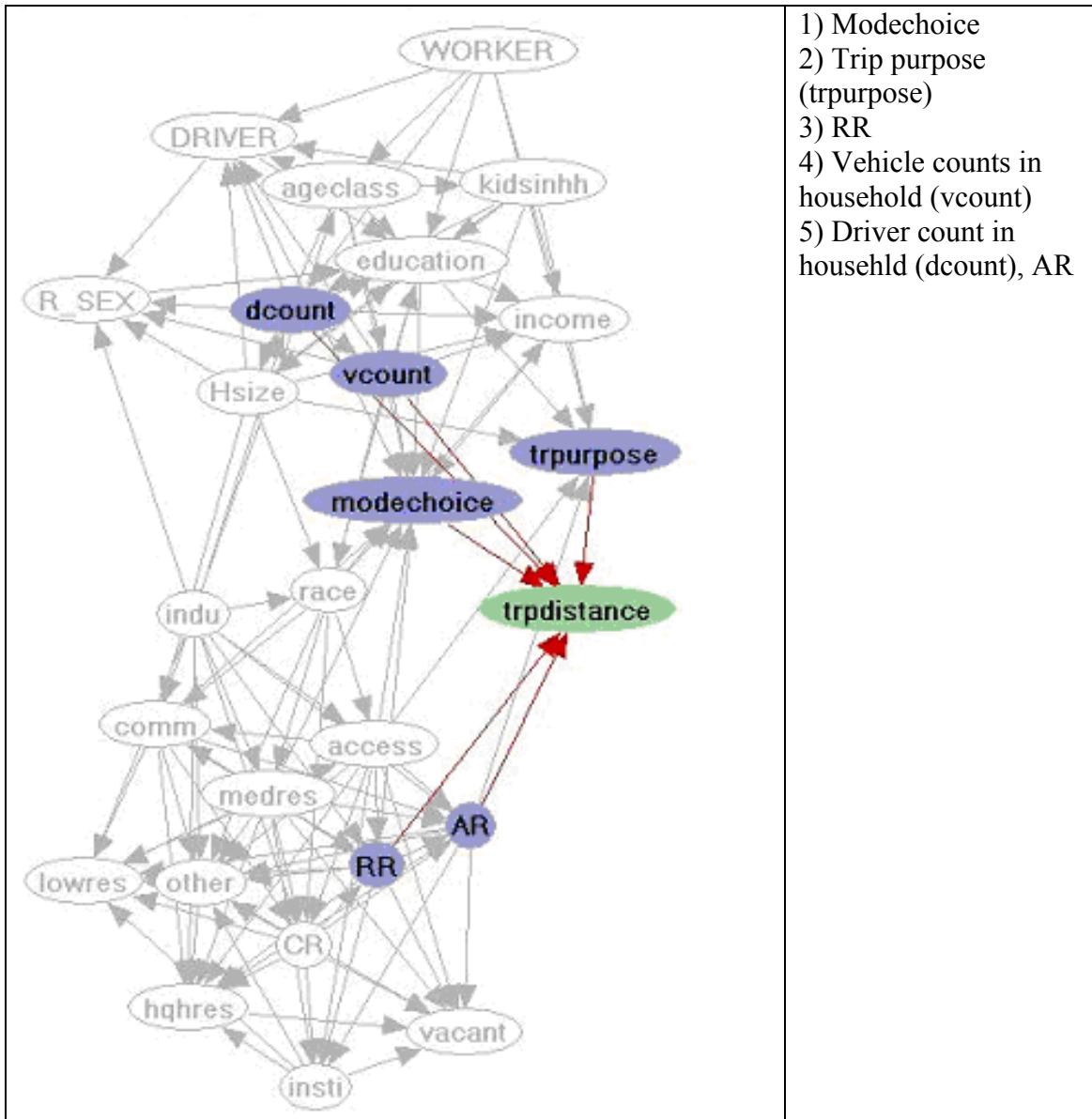
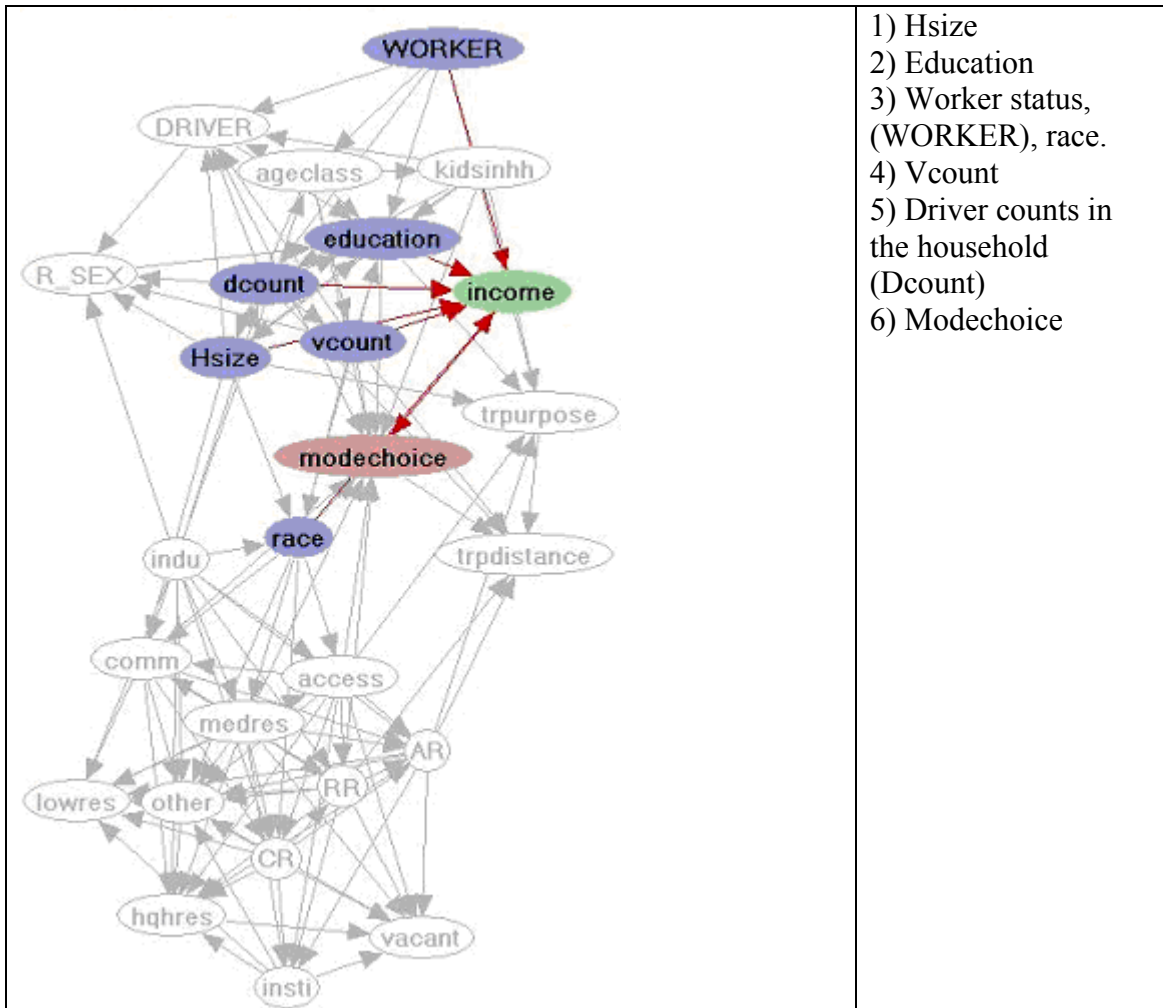


Figure B.4: BBN for trip distance



- 1) Hsize
- 2) Education
- 3) Worker status, (WORKER), race.
- 4) Vcount
- 5) Driver counts in the household (Dcount)
- 6) Modechoice

Figure B.5: BBN for income

B.4, 4.8b: As expected, trip distance has a relationship with mode choice, but a more in depth discussion about the direction of this relationship is warranted later in section 4.3.1 “Cause and Effect or Feedback Loop?”. To be noted however is how trip purposes are strongly linked to trip distances, as well as a number of other personal and household characteristics. The variable income (BBN above to the right) is related to household size and to the number of workers; it appears that the bigger the household, the more workers are available to generate income and the higher the income will be. The variable income, however, has a very weak causal relationship with transportation mode choice.

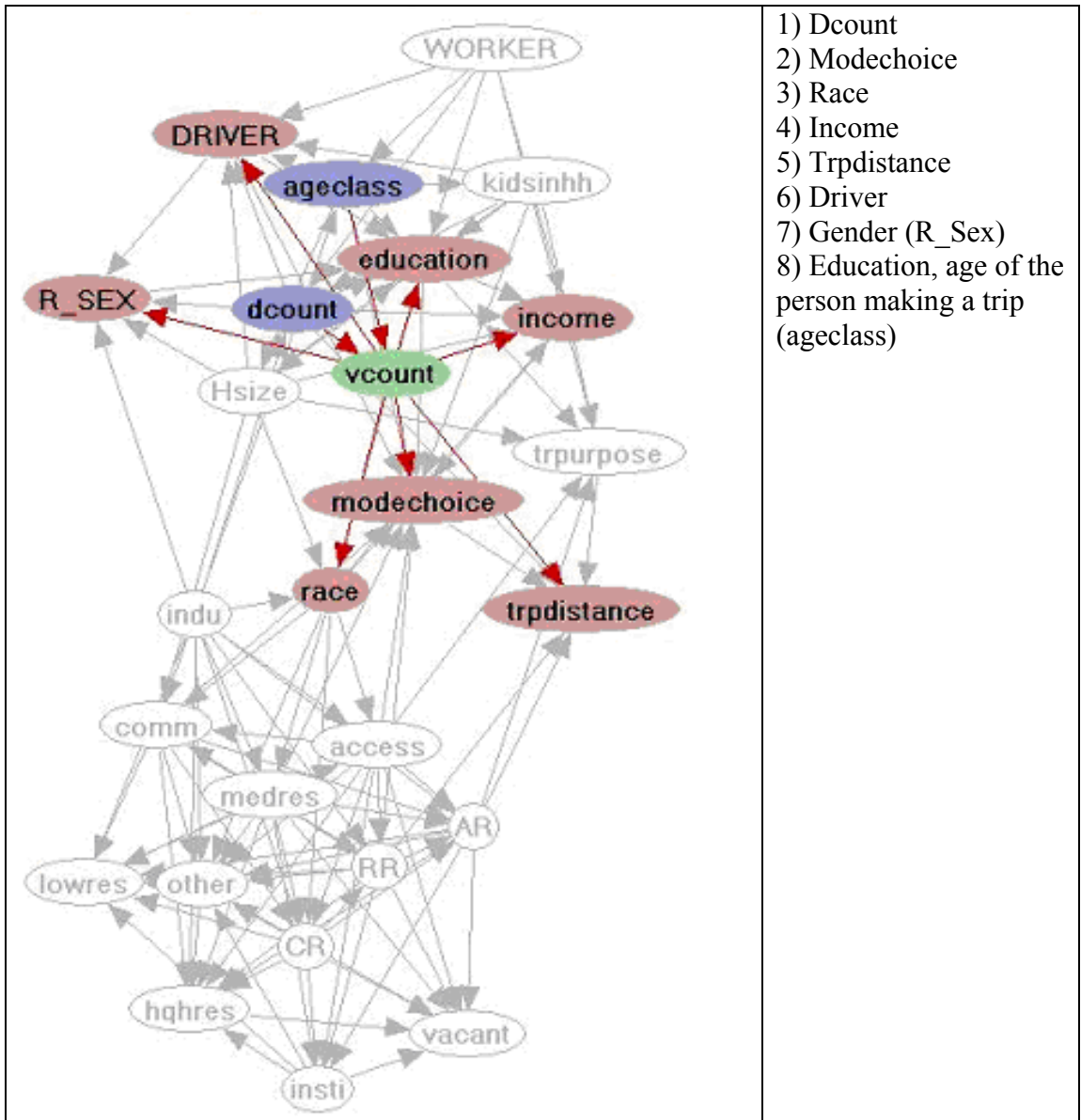


Figure B.6: BBN for vehicle count

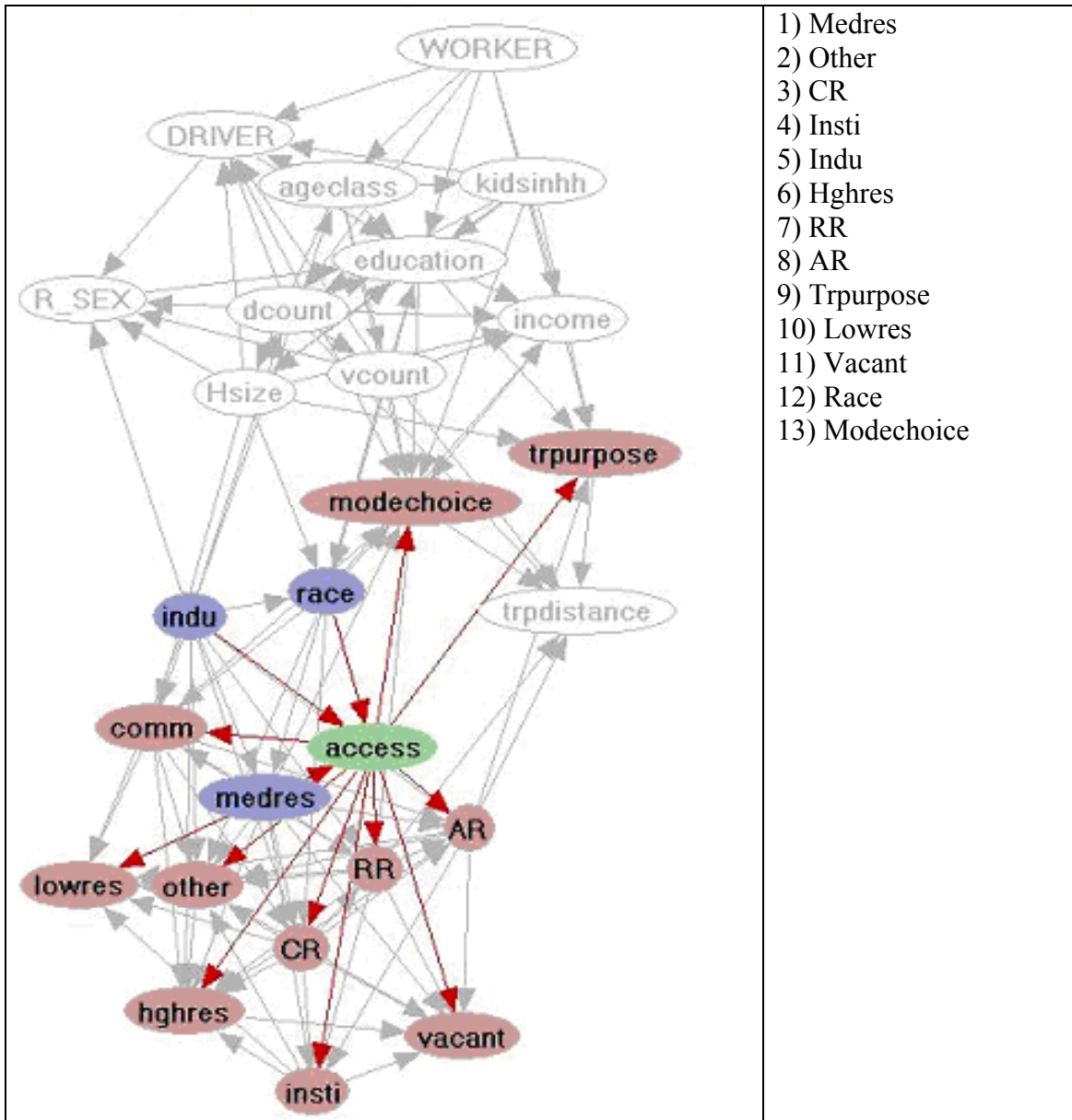


Figure B.7: BBN for road density

B.6, 4.9b: The variable vehicle count, contrary to income, has a fundamental moderating role for the variable “modechoice”. In the BBN it also appears that the availability of privately owned vehicles in a household has a relationship with the household level of income. The associated decision tree hints that for higher “Vcounts” income will be higher too, as if the accessibility to jobs increases because of that. Road density is characterized more by its association with land use variables than with race and ethnicity, indicating that different population groups live in tracts with significantly different levels of road density.

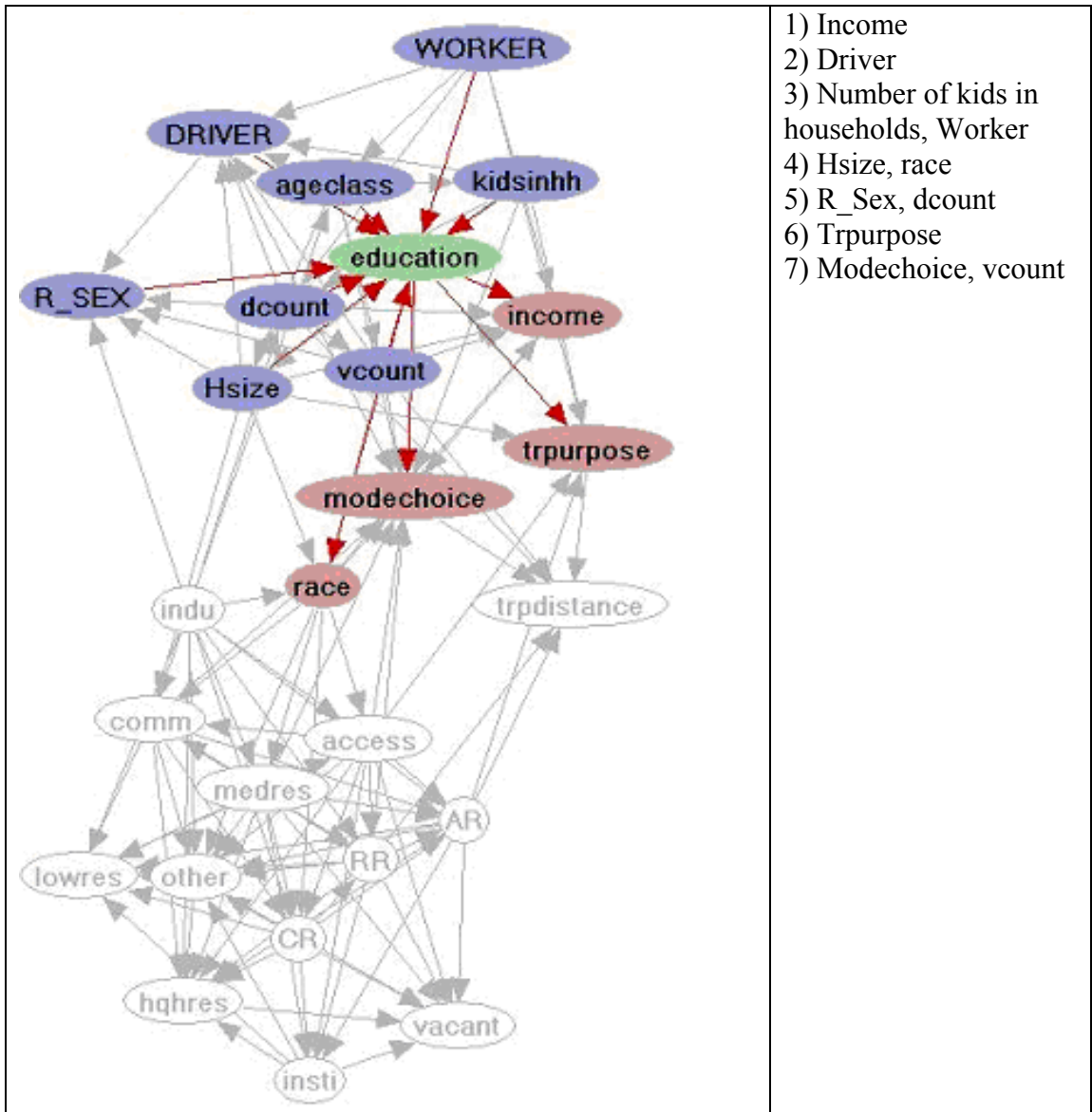


Figure B.8: BBN for education

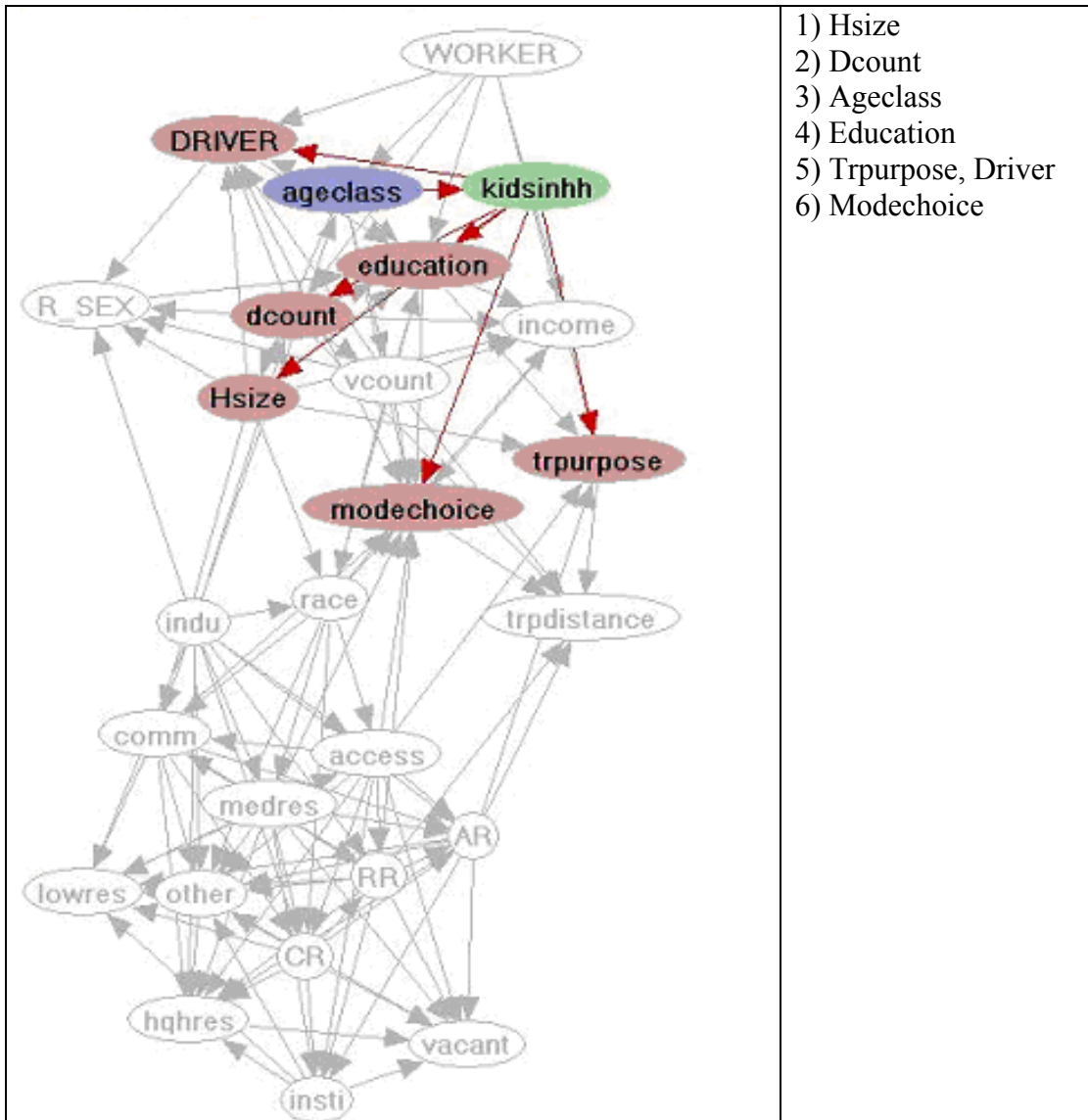


Figure B.9: BBN for number of children in households

B.8, 4.10b: Education is not a matter of land use! The BBN above left shows it as related to other socioeconomic characteristics with only income and number of children in the households exercising strong causal relationships. The associated decision tree was looked at in order to establish if lower income is associated with a higher number of children in the households, but it was not possible to establish a unique answer to that. Education has little causal influence on transportation mode choice; neither has the number of children in the household (BBN above right). In this case, the network correctly maps our expectations that depending on the number of children in the household, the number of drivers will change, the household size will change, the income will change, and so will educational attainment.



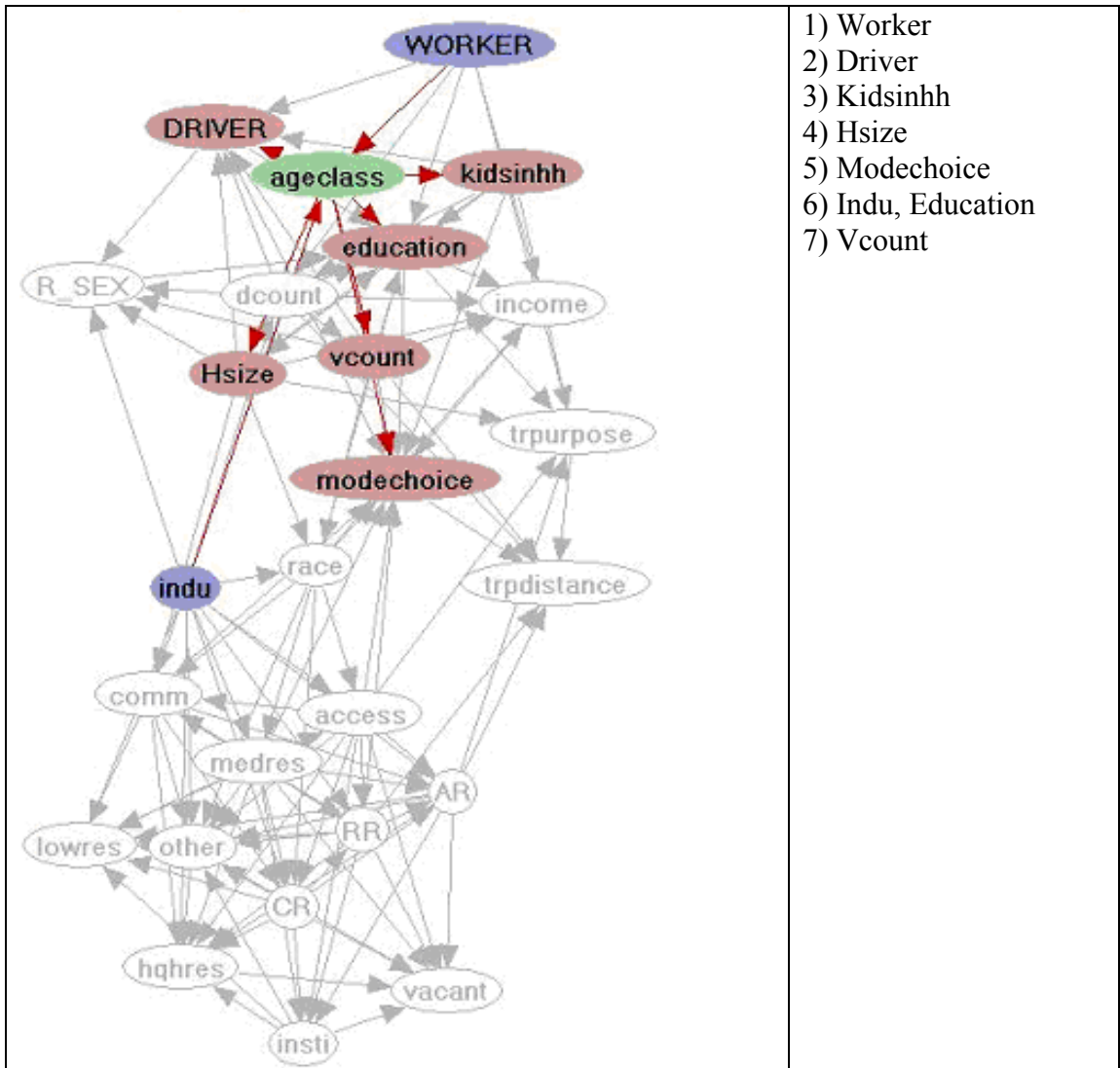


Figure B.10: BBN for age class

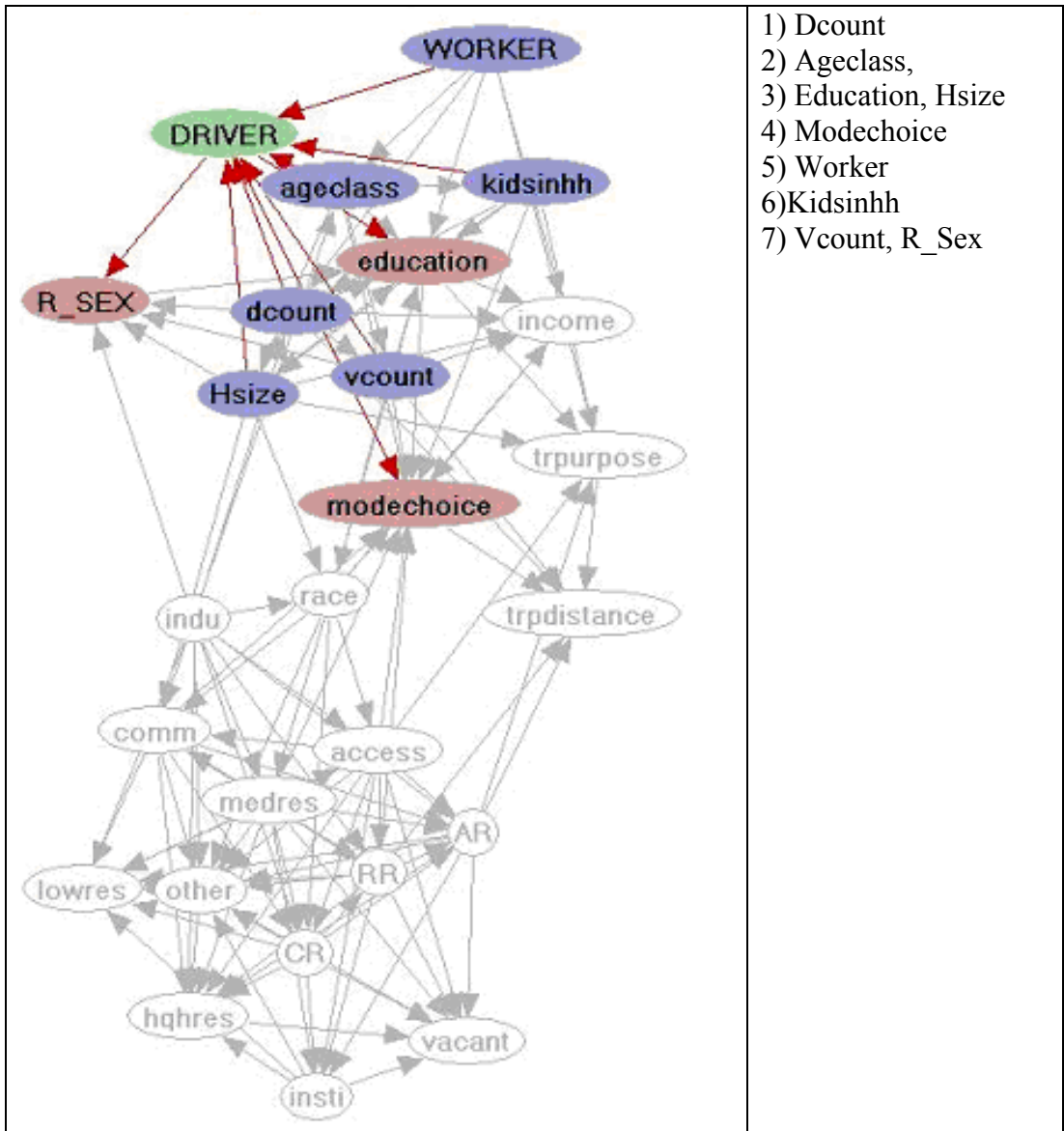


Figure B.11: BBN for driver status

B.10, 4.11b: Age class and driver status (BBNs above left and right) are mostly a function of socioeconomic status. Interestingly, it appears that greater proportions of older populations more than other age groups are living in industrial spaces. Debatable is the causal relationship between driver status and gender where one could infer that more males have driver's licenses or equally that females drive less. In the sample the ratio of females to males is 1.15 to 1, so there are more responses from females than males while the same ratio in the population is 1.07 to 1. As seen before, having a driver's license is one of the conditions to use privately owned vehicles as a transportation option, so it figures strongly in influencing the variable "modechoice"

## Bibliography

- 1000 Friends of Oregon. 1993. The Pedestrian Environment: LUTRAQ Report Volume A. Prepared by Parsons Brinckerhoff Quade & Douglas.
- Aliferis, C., Tsamardinos, I., Statnikov, A., Brown, L. E. (2003). Causal Explorer: A Probabilistic Network Learning Toolkit for Biomedical Discovery. In Proceedings of the 2003 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS), June 23-26, Las Vegas, Nevada, USA, CSREA.
- Alonso, W. (1964). Location and land use: Toward a general theory of land rent. Cambridge, Ma., Harvard University Press.
- American Society of Planning Officials (ASPO) (1951). The journey to work: Relations between employment and residence. Planning Advisory Service Report no. 26. Chicago: American Society of Planning Officials.
- Arentze, T.A., and Timmermans, H.J.P. (2004). Capturing the Role of Awareness and Information Search Processes on Choice Set Formation in Models of Activity-Travel Behavior1. Electronic conference proceedings of the 83rd Annual Meeting of the Transportation Research Board (CD-ROM), Washington, D.C.
- Badoe, D. A., and Miller, E. J. (2000) Transportation-land-use interaction: empirical Findings in North America and their implications for modeling. Transportation Research Part D n. 5, 235-263.
- Baltimore Regional Transportation Board (2002). 2001 National Household Travel Survey Add-On Program: Baltimore Regional Transportation Board, Final Report and Codebook.
- Baldi, P., and Hatfield, W. (2002). DNA Microarrays and Gene Expression. Cambridge University Press.
- Black, W. R. (2003). Transportation: A Geographical Analysis. The Guilford Press.
- Blunden, W.R., and Black, J. A. (1984). The Land Use/Transport System. Pergamon Press.
- Boarnet, M. G. and Crane, R. (2001a). Travel By Design - The Influence of Urban Form on Travel. Oxford University Press.
- Boarnet, M. G. and Crane, R. (2001b). The Influence of Land Use on Travel Behavior: Empirical Strategies, Transportation Research A, 35, pp 823-845.

- Boarnet, W. R., and Greenwald, M. J. (2000). Land Use, Urban Design, Nonwork Travel. Transportation Research Record 1722, Paper No. 00-1054
- Breda-Vazquez, I., Riberio-Ramos, P. (2003). Metropolitan Dynamics, Travel Behavior and the Importance of New Urban Policies to Achieve Sustainable Mobility: A Local Perspective. Urban Transport IX, Ninth International Conference on Urban Transport and the Environment in the 21st Century.
- Chatman D. G (2003). How workplace land use affects personal commercial travel and commute mode choice. Transportation Research Record: Journal of the Transportation Research Board 1831, pp 193-201.
- Cervero, R. (1984). Light Rail Transit and Urban Development. Journal of the American Planning Association 50, pp133-147.
- Cervero, R and Duncan, M (2003). Walking, Bicycling, and Urban Landscapes: Evidence From the San Francisco Bay Area. American Journal of Public Health Vol 93, No. 9 pp:1478-1483.
- Cervero, R., and Kockelman, K. (1997). Travel Demand and the 3Ds: Density, Diversity and Design. Transportation Research, Part D, Vol. 2, No. 3, pp.199-219.
- Cervero, R. (1986). Suburban Gridlock. New Brunswick, N.J., Center for Urban Policy Research.
- Cervero, R. (1989). America's Suburban Centers: The Land-Use Transportation Link. Unwin-Hyman.
- Cervero, R. (1994). Transit-Based Housing in California: Evidence on Ridership Impacts. Transportation Policy Vol. 1 No.3, pp174-183
- Cervero, R. (1995a). Why Go Anywhere? Millions of people could be liberated from their vehicles. Scientific American (September 1995) pp. 92-93.
- Cervero, R. (1996). Mixed Land-Uses and Commuting: Evidence from the American Housing Survey. Transportation Research A: Policy and Practice 30 (N5) pp. 361-377.
- Cervero, R., and Landis, J. (1995b). The Transportation-Land Use Connection Still Matters. Access 7(Fall 1995) pp. 2-10.
- Cervero, R., and Radisch, C. (1996). Travel Choices in Pedestrian Versus Automobile Oriented Neighborhoods. Transport Policy, Vol. 3, pp. 127-141.
- Cervero, R., and Landis, J. (1996b). Why the Transportation Land-Use connection is still important. TR News 187 (November-December) pp. 9-11.

- Cervero, R. (2002). Built environments and mode choice: toward a normative framework. *Transportation Research, Part D No. 7* pp. 265-284.
- Cervero, R. and Kockelman, K. (1997). Travel Demand and the 3Ds: Density, Diversity and Design. *Transportation research Part D: Transport and Environment 2 (N3)* pp. 199-219
- Cervero, R., and Seskin, S. (1995). An Evaluation of the Relationships Between Transit and Urban Form. Transit Cooperative Research Program, TRB, National Research Council, Washington, D.C.
- Chickering, D. M., Heckerman, D., Meek, C. (1997). A Bayesian Approach to Learning Bayesian Networks with Local Structure. Microsoft Technical Report MSR-TR-97-07
- Christian, P. R. (2001). *The Bayesian Choice*. Springer.
- Clifton, K., and Handy, S. L. (2001). Qualitative Methods in Travel Behavior Research. Prepared for the International Conference on Transport Survey Quality and Innovation, Kruger National Park, South Africa, August 5-10.
- Committee for Study of Impacts of Highway Capacity, Improvements on Air Quality and Energy Consumption (1995). Land Use and Urban Form. Expanding Metropolitan Highways - Implications for Air Quality and Energy Use, Transportation Research Board, Washington D.C., National Academies of Science, pp. 174-209.
- Cooper, G. F., and Herskovits, E. (1992). A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, Kluwer Academic Publishers, Boston 9, pp. 309-347.
- Council on Environmental Quality (1976). *The Growth Shapers*. Washington D.C.
- Crane, R. (2000). The Impacts of Urban Form on Travel: An Interpretive Review. *Journal of Planning Literature* 15:3-23.
- Dean, T. L and Wellman, M. P. (1991). *Planning and Control*. San Mateo, CA: Morgan Kaufmann.
- Dieleman, F. M., and Burghouwt, G. (2002). Urban Form and Travel Behaviour: Micro-level Household Attributes and Residential Context. *Urban Studies* 39(3), pp. 507-527.
- Dill, J. (2005). *Measuring Network Connectivity for Bicycling and Walking*, School of Urban Studies and Planning, Portland State University (<http://web.pdx.edu/~jdill/ALRbikes>).

- Downs, A. (1992). *Stuck in traffic: Coping with peak hour traffic congestion*. Washington, D.C.: Brookings Institution.
- Druzdel, M. (1996) in *Artificial Intelligence and Simulation of Behaviour Quarterly*, Special Issue of Bayesian Belief Networks, 94, pp.43-54.
- Ewing, R. (1995). *Beyond Density, Mode Choice and Single-Purpose Trips*. *Transportation Quarterly* 49(4), pp. 15-24.
- Ewing, R., and Cervero, R. (2001). *Travel and the Built Environment: A Synthesis*. *Transportation Research Record* 1780, Paper No. 01-3515, pp. 87-122.
- Federal Transit Administration (2003). *Planning Technology*.
- Frank, L., Andersen, A. M., and Schmid, T.L., (2004). *Obesity Relationships with Community Design, Physical Activity, and Time Spent in Cars*. *American Journal of Preventive Medicine*, Vol. 27, No. 2.
- Fried, M., Havens, J., and Thall, M. (1977). *Travel Behavior – A Synthesized Theory*, Final Report, NCHRP, Transportation Research Board, Washington.
- Friedman, N., and Koller, D. (2000). *Being bayesian about network structure*. In C. Boutilier and M. Godszmidt (eds), *Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers, pp. 201-210
- Glasser, E. L., and Shapiro, J. M. (2003). *City Growth. Redefining Urban & Suburban America - Evidence from census 2000*, Kats B and Lang R. E. editors, Washington DC, Brookings. 1 pp. 13-29.
- Giuliano, G. (1995a). *The Weakening Transportation-Land Use Connection*. *Access* 6(Spring 1995) pp. 3-11.
- Giuliano, G. (1995b). *Land Use Impacts of Transportation Investments - Highway and Transit*. *The Geography of Urban Transportation*. S. Hanson, The Guilford Press.
- Glymour, C., and Cooper, G. F. editors (1997). *UAI, Computation, Causation, and Discovery*. 1999, AAAI Press/The MIT Press.
- Gordon, P., and Richardson, H. W. (1997). *Are Compact Cities a Desirable Planning Goal?* *Journal of the American Planning Association* 63(1 Winter):95-106
- Greenwald, M. J. (2003). *The Road Less Traveled – New Urbanist Inducements to Travel Mode Substitution for Nonwork Trips*. *Journal of Planning Education and Research* 23, pp39-57.

- Guo, H., Perry, B. B., Stilson, J. A., Hsu, W. H. (2002). A Genetic Algorithm for Tuning Variable Orderings in Bayesian Network Structure Learning. AAAI02 Student Abstract.
- Haipeng, G., Stilson, J. A., Hsu, W. H. (2002). A Genetic Algorithm for Tuning Variable Orderings in Bayesian Network Structure Learning. AAAI02 Student Abstract.
- Hansen, M. (1995). Do New Highways Generate Traffic? Access 7(Fall 1995): 16-22.
- Handy, S. L. (1992). Regional Versus Local Accessibility: Neotraditional development and its Implications for non-work travel. Built Environment 18(4): 253-267.
- Handy, S. L. (1996). Understanding the link between Urban Form and Nonwork Travel Behaviour. Journal of Planning Education and Research 15 (N3): 183-198.
- Handy, S. L., Clifton K., and Fisher, J (1998). The Effectiveness of Land Use Policies as a Strategy for Reducing Automobile Dependence: A Study of Austin Neighborhoods, Research Report SWUTC/98/467501-1, University of Texas at Austin, October.
- Handy, S., Paterson, R., and Butler, K. (2003). *Planning for Street Connectivity: Getting from Here to There*. Planning Advisory Service Report. Chicago, IL: American Planning Association.
- Handy, S., and Westin, L. (2005a). Driving by choice or necessity. Transportation Research Part A – Policy and Practice 39 (2-3), pp183-203.
- Handy, S. (2005b). Smart Growth and the transportation – Land use connection: What does the research tell us? International Regional Science Review vol 28(2), pp146-167.
- Heckerman, D., Chickering, D. M., Meek, C, Rounthwaite, R, Kadie, C. 2000. Dependency Networks for Inference, Collaborative Filtering, and Data Visualization. Journal of Machine Learning Research vol.1:49-75.
- Heckerman, D., Geiger, D., Chickering, D. M. (1995). Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. Machine Learning 20: 197-243.
- Holtzclaw, J. (1994). Using Residential Patterns and Transit to Decrease Auto Dependence and Costs. Natural Resource Defense Council. Also in Crane, R. (1999) The Impacts of Urban Form on Travel: A Critical Review. Lincoln Institute of Land Policy.
- Hsu, W. H., Haipeng, G, Benjamin, P.B. and Stilson, J. A. (2002). A Permutation Genetic Algorithm for Variable Ordering in Learning Bayesian Networks from Data, Laboratory for Knowledge Discovery in Databases, Kansas State University.

- Irwin, N. (2003). For Urban Developers, A Hard Row to Home, Washington Post. Washington DC: E1.
- Jensen, V. F. (2001). Bayesian Networks and Decision Graphs. Springer.
- Jensen, V. F. (1996). "Bayesian network basics," *AISB Quarterly*, vol. 94, pp. 9-22.
- Janssens, D., Wets, G., Brijs, T., Vanhoof, K. and Timmermans, H.J.P (2003). Identifying Behavioral Principles underlying Activity Patterns by means of Bayesian Networks. Electronic conference proceedings of the 82nd Annual Meeting of the Transportation Research Board (CD-ROM), Washington, D.C.
- Janssens, D., Wets, G., Brijs, T., and Vanhoof, K. (2004) Improving the Performance of a Multi-Agent Rule-Based Model for Activity Pattern Decisions Using Bayesian Networks. Electronic conference proceedings of the 82nd Annual Meeting of the Transportation Research Board (CD-ROM), Washington, D.C.
- Johnston, R. A., Gao, S., Shabazian, D. R., (2003). UPlan: A Versatile Urban Growth Model for Transportation Planning. Transportation Research Board, Washington DC.
- Kelly, E. D. (1994). The Transportation Land-Use Link. *Journal of Planning Literature*, Vol 9, No. 2.
- Kitamura, R., Monkaria, P., and Laidet, L. (1997). Micro-Analysis of Land Use and Travel in Five Neighborhoods in the San Francisco Bay Area. *Transportation* 24: 125-158.
- Knox, P. L., and Marston, S. A. (2001). *Places and Regions in Global Context: Human Geography*. Prentice Hall.
- Kockleman, K. M. (1997). Travel Behavior as a Function of Accessibility, Land Use Mixing, and Land Use Balance: Evidence from San Francisco Bay Area. *Transportation Research Record* 1607, pp116-125
- Krizek, K. J., and Levinson, D. M. (2003). Teaching Integrated Land Use-Transportation Planning: Topics, Readings, and Strategies. Transportation Research Board, Washington DC.
- Krizek, K. J. (2003a). Residential Relocation and Changes in Urban Travel: Does Neighborhood-Scale Urban Form Matter? *Journal of the American Planning Association*, Vol.69, No. 3, pp265-281.
- Krizek, K. J. (2003b). Operationalizing Neighborhood Accessibility for Land Use-Travel Behavior Research and Regional Modeling. *Journal of Planning Education and Research* No. 22 pp.270-287.



- Krizek, K. J. (2003c). Pretest-Posttest Strategies for Researching Neighborhood-Scale Urban Form and Travel. *Transportation Research Record* 1722.
- Krizek, K. and Waddell, P. (2002). Analysis of Lifestyle Choices, Neighborhood Type, Travel Patterns, and Activity Participation.
- Kulash, W, Anglin J, and David Marks (1990). Traditional neighborhood development: Will the traffic work? Paper prepared for Successful Land Development: Quality and Profits Conference, March. American Society of Civil Engineers.
- Kulkarni, A. A. (1996). The Influence of Land Use and Network Structure on Travel Behavior. Master dissertation, University of California, Irvine, Department of Civil and Environmental Engineering. Also in Crane, R. (1999) *The Impacts of Urban Form on Travel: A Critical Review*. Lincoln Institute of Land Policy
- Landis, J. D. (2001). Characterizing Urban Land Capacity: Alternative Approaches and Methodologies. *Land Market Monitoring for Smart Urban Growth*, Lincoln Institute of Land Policy.
- Larrañaga, P. Poza, M., Yurramendi, Y., Murga, R. H., Kujipers, C. M. H. (1996). Structure Learning of Bayesian Networks by Genetic Algorithms: A Performance Analysis of Control Parameters. *IEEE Transactions on Pattern Analysis and Machine Learning* 18(No. 8): pp912-926.
- Leonard, T., and Hsu, J. S. (2001). *Bayesian Methods - An Analysis for Statisticians and Interdisciplinary Researchers*. Cambridge University Press.
- Messenger, T., and Ewing R (1996). Transit-Oriented Development in the Sunbelt. *Transportation Research Record* 1552.
- McGuckin (2004). The Changing Trends in the U.S. – Understanding Travel Behavior for Policy Making. Paper submitted to the Transportation Research Board).
- McNally, M. G., and Ryan, S. (1993) Comparative Assessment of Travel Characteristics for Neotraditional Designs. *Transportation Research Record* 1400, pp 67-77
- McNally, M. G. (1995). How Neighborhood Design Affects Travel. *ITS Review* 19(1-2 Nov. 1995/Feb. 1996): 2-3.
- Miller, H and Han J. editors (2001). *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis
- Miller and Leslie (2003). Cars, trucks now outnumber drivers. <http://hypertextbook.com/facts/2001/MarinaStasenko.shtml> (Last accessed 02/29/2004).

- Mitchell, R. B., and Rapkin, C. (1954). *Urban traffic: A function of land use*. Columbia University Press, New York.
- Moore, T. (1996). Premises and Promises, Putting the Transportation-Land Use relationship in Context. *TR News* 187(November December): 7-8.
- National Cooperative Highway Research Program (1999). *Land Use Impact of Transportation: A Guidebook*. Washington DC, National Academy Press.
- Neapolitan, R.E. (1990). *Probabilistic Reasoning in Expert Systems*. John Wiley and Sons.
- Friedman, N., and Koller, D. (2000). Being Bayesian about Network Structure. *Uncertainty in Artificial Intelligence Proceedings 2000*.
- NTI (2000a) National Transit Institute. Two Schools of Thought in Coordinating Transportation and Land Use. Session 9: Land Use Impacts on Travel Behavior, National Transit Institute, Rutgers State University of New Jersey: 9.1-9.36.
- NTI (2000b) National Transit Institute. Transportation-Land Use Analysis in Coordinating Transportation and Land Use. Session 6: Analysis Tools, National Transit Institute, Rutgers State University of New Jersey: 6.1 6.20.
- NTI (2000c) National Transit Institute. Transportation and Land Use: A Two Way Street in Coordinating Transportation and Land Use. Session 8: Transportation Impacts on Land Use, National Transit Institute, Rutgers State University of New Jersey: 8.3-8.36.
- O'Regan, K. M., and Quigley J. M. (1998). *Accessibility and Economic Opportunity*. Working Paper, Yale and UC Berkeley (February).
- Payne-Maxie Consultants, and Urban and Regional Planners Blayney-Dyett. (1980). *The land use and urban development impacts of beltways: Case studies*. DOT-P-30-80-31. Washington D.C. U.S. Department of Transportation.
- Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge University
- Pearl, J. (1994). A probabilistic calculus of actions, in *Proceedings of the Tenth Conference on Uncertainty in AI (UAI-94)*, San Mateo, CA: Morgan Kaufmann Press.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of plausible Inference*. San Mateo, CA: Morgan Kaufmann Press.
- Pearl, J. (1986). Constraint-propagation approach to probabilistic reasoning. in L. M. Kanal and J. Lemmer (Eds.), *Uncertainty in Artificial Intelligence*, North-Holland, Amsterdam, pp. 357-288.

- Polzin, S. E. (1999). Transportation/Land-Use Relationship: Public Transit's Impact on Land Use. *Journal of Urban Planning and Development*, December (Vol 125 No 4).
- Pushkarev, B. M., and Zupan J. M. (1977). *Public transportation and land use policy*. Bloomington: Indiana University Press.
- Rabiega, W. A. and Howe, D. A. (1994). Shopping Travel Efficiency of Traditional, Neotraditional, and Cul-de-sac Neighborhoods with Clustered and Strip Commercial. Paper presented at the Association of Collegiate Schools of Planning 36th Annual Conference, Tempe.
- Rodriguez, D. A., and Joonwon, J. (2004). The relationship between non-motorized mode choice and the local physical environment. *Transportation Research Part D* vol 9, pp 151-173.
- Rudenstam, K. E., and Newton, R. R. (2001) *Surviving your dissertation*. Sage publishing, 2<sup>nd</sup> edition.
- Schaeffer, K. H., and Sclar E. (1975). *Access for all: Transportation and urban growth*. Baltimore, Md.: Penguin.
- Schlich, R., and Axhausen, K.W. (2003). Habitual Travel Behavior: Evidence from a Six Week Travel Diary. *Transportation* Vol: 30 No: 1 02/00/2003
- Shen, Q. (2001). A Spatial Analysis of Job Openings and Access in a U.S. Metropolitan Area. *Journal of the American Planning Association*, Vol 67, No 1, pp53-68.
- Skinner, R. E. J. (1996). The Transportation-Land Use Interaction. *TR News* 187(November-December): 6.
- Stone, J. R., Foster, M. D. and Johnson, C. E. (1992). *Neo-Traditional Neighborhoods: A Solution to Traffic Congestion? Site Impact Traffic Assessment*. New York: American Society of Civil Engineers.
- Song, Y. and Knaap, G. (2004). "Is Portland less sprawling? A comparative analysis of urban development patterns." ACSP Conference, Portland, OR.
- Taaffe, E. J, and Gauthier, H. L., O'Kelly, M. E. (1996). *Geography of Transportation*, Prentice Hall.
- Targa, F. and Clifton, K. (forthcoming in 2006). Built Environment and Trip Generation for Non-Motorized Travel. Special edition, *Journal of Transportation and Statistics*, January 2006.

- Taylor, M.A.P., Ampt, E.S. (2003). Traveling Smarter Down Under: Policies for voluntary travel behavior change in Australia. *Transport Policy* Vol: 10 No: 3
- Texas Transportation Institute (2000). Transportation Model Improvement Program, <http://ntl.bts.gov/DOCS/467.html>.
- Leonard, T., and Hsu, J. S. J. (2001). *Bayesian Methods - An Analysis for Statisticians and Interdisciplinary Researchers*. Cambridge University Press.
- Torres, F. J, and Huber, M. (2003). Learning a Causal Model from Household Survey Data Using a Bayesian Belief Network. Transportation Research Board Annual Meeting.
- Verna, T. S. (1987). Causal networks: Semantics and expressiveness. In *Proceedings of Third Workshop on Uncertainty in Artificial Intelligence*, pp. 352-359.
- Waddell, P. (2001). Towards a Behavioural Integration of Land Use and Transportation Modelling. 9<sup>th</sup> International Association for Travel Behaviour Research Conference, Queensland, Australia.
- Wegener, M. (1995). *Current and Future Land Use Models*, Technology Share Program, Washington, DC: US Department of Transportation.
- Wingo, L. (1961). *Transportation and urban land*. Washington, D.C: Resources for the Future.
- Young, W. (1986). The Role of Thresholds in Transport choice. *Behavioural Research for Transport Policy*. Ruhl A. editor. Utrecht, the Netherlands, VNU Science Press: 153-170.