

THE APPLICATION OF ADVANCED INVENTORY TECHNIQUES IN URBAN INVENTORY DATA DEVELOPMENT TO EARTHQUAKE RISK MODELING AND MITIGATION IN MID- AMERICA

A Dissertation
Presented to
The Academic Faculty

by

Subrahmanyam Muthukumar

In partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in City and Regional Planning in the
College of Architecture

Georgia Institute of Technology
December, 2008

COPYRIGHT 2008 BY SUBRAHMANYAM MUTHUKUMAR

**THE APPLICATION OF ADVANCED INVENTORY TECHNIQUES
IN URBAN INVENTORY DATA DEVELOPMENT TO
EARTHQUAKE RISK MODELING AND MITIGATION IN MID-
AMERICA**

Approved by:

Dr. Steven P. French, Advisor
College of Architecture
Georgia Institute of Technology

Dr. Patrick McCarthy
School of Economics
Georgia Institute of Technology

Dr. William Drummond
College of Architecture
Institute of Technology

Dr. Barry Goodno
School of Civil and Environmental *Georgia*
Engineering
Georgia Institute of Technology

Dr. Jiawen Yang
College of Architecture
Georgia Institute of Technology

Date Approved: September 4, 2008

To my wife, Suchita

ACKNOWLEDGEMENTS

Developing the research and then writing a dissertation is an arduous task and patently not possible without the personal assistance of numerous people. First and foremost, the author is completely beholden to his wife for her staunch and constant support throughout the entire process. There are really no words that can convey the author's gratefulness towards his wife and daughter for their tireless sacrifice and patience during this period. The author also wishes to express his gratitude to his parents and brother for their unceasing love and encouragement throughout the course of his educational career.

The author would also like to gratefully acknowledge the advice, support, encouragement and inspiration provided by Dr. Steven French throughout his tenure at Georgia Tech. Without his guidance, this dissertation would not have been possible. Thanks are also expressed to the remaining members of the author's dissertation committee – Drs. William Drummond, Barry Goodno, Patrick McCarthy and Jiawen Yang for their time, comments and innumerable educational inputs. Special thanks go to Dr. Linda Thomas-Mobley for her invaluable assistance in the development of the research.

The author also wishes to extend his special thanks to Ning Ai, Sarajoy Crewe and Liora Sahar for their sincere willingness to share their considerable knowledge and their timely assistance and friendship. The author is also grateful to all his friends, too numerous to list here, who have in one way or another encouraged him to persevere and complete the dissertation.

Finally, gratitude is expressed to the Mid-America Earthquake Center – this research was supported by the Mid-America Earthquake Center under National Science Foundation Grant EEC-9701785

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iv
LIST OF TABLES	x
LIST OF FIGURES	xiv
LIST OF ABBREVIATIONS	xvii
SUMMARY	xviii
<u>CHAPTER 1 . INTRODUCTION</u>	<u>1</u>
1.1. BACKGROUND FOR DISASTER MITIGATION.....	3
1.2. THE NEED FOR ACCURATE URBAN INVENTORIES.....	8
1.3. HAZARD MITIGATION IN THE PLANNING PROCESS FRAMEWORK	10
1.4. EXISTING METHODS FOR URBAN INVENTORY DATA COLLECTION AND LIMITATIONS.....	15
1.4.1. Urban Inventory Data Sources.....	16
1.4.2. Classification of the Urban Building Inventory	17
1.4.3. Building Inventory Development in HAZUS MR-3.....	18
1.5. ADVANCED INVENTORY TECHNOLOGIES AND TECHNIQUES FOR DATA COLLECTION.....	24
1.5.1. Remote Sensing Technologies	24
1.5.2. Building Inventory Estimation Methods.....	26
1.5.3.1. Knowledge-based Rules	27
1.5.3.2. Classification Models	28
1.6. THE EARTHQUAKE MODELING PROCESS REQUIREMENTS	29
1.7. DESCRIPTION OF RESEARCH	32
1.7.1. Research Statement.....	32
1.7.2. Research Goals and Objectives	32
1.7.3. Significance of the Research Effort	33
1.7.4. Scope of Research.....	36
1.8. ORGANIZATION OF DISSERTATION	36
<u>CHAPTER 2 . LITERATURE REVIEW</u>	<u>39</u>
2.1. PATTERN RECOGNITION AND THE POTENTIAL FOR AUTOMATION	40
2.2. MULTINOMIAL LOGISTIC REGRESSION FOR CLASSIFICATION	43
2.3. ARTIFICIAL NEURAL NETWORK SOLUTIONS FOR CATEGORICAL DATA ANALYSIS.....	45

2.3.1. A historical perspective on Artificial Neural Networks.....	46
2.3.2. What is an Artificial Neural Network?	47
2.3.3. Transfer Functions.....	52
2.3.4. Applications of Artificial Neural Networks.....	53
2.3.4.1. Function approximation	54
2.3.4.2. Time series analysis and prediction	54
2.3.4.3. Classification	54
2.3.4.4. Data mining	55
2.3.5. Why use Artificial Neural Networks?	56
2.3.6. Artificial Neural Network topologies for classification.....	57
2.3.6.1. Neural Computing for Classification	58
2.3.6.2. Discriminant Functions	58
2.3.7. Conceptual issues in designing and training Artificial Neural Networks	60
2.3.7.1. Error minimization search procedures	61
2.3.7.2. Learning rate	63
2.3.7.3. Learning algorithms.....	64
2.3.7.4. Processing elements in the hidden layer	66
2.3.7.5. Stop criteria	66
2.3.7.6. Performance Measures	68
2.4. SHAPE RECOGNITION BACKGROUND, TECHNIQUES AND APPLICATIONS	69
2.4.1. Definition of a Shape	72
2.4.2. The Process of Shape Analysis.....	73
2.4.2.1. Shape Acquisition	74
2.4.2.2. Shape Representation	76
2.4.2.3. Feature Extraction or Shape Description after Shape Representation	78
2.4.2.4. Invariant Representations	78
2.4.2.5. Statistical and Mathematical Approaches	86
2.4.2.6. Structural and Syntactic Methods	89
2.4.2.7. Syntactic Shape Recognition	96
2.4.2.8. Shape Recognition and Classification.....	97
2.5. GEOMETRY MANIPULATIONS IN THE GIS ENVIRONMENT.....	98
2.5.1. Representation of Points, Lines and Regions in GIS	99
2.5.2. Topological Data Structures	100
2.5.3. Geometric Primitives and Object Hierarchy	101
2.5.4. Manipulating Vector GIS Feature Geometry for Shape Preprocessing	102

2.5.5. Densification of Edges and Polylines	103
2.5.6. Generalization, Polygon Approximation and Line Simplification.....	106
2.5.7. Generalization Routines and Vertex Decimation Strategies	106
2.5.8. Line Simplification	112
2.5.9. Evaluation of Generalization and Simplification Algorithms	115
2.6. BUILDING VALUATION	116
2.6.1. Replacement Costs of Buildings	118
2.6.2.1. Structural and Nonstructural Building Components	120
2.6.2.2. Earthquake-related Damage to Nonstructural Components and Contents	121
2.6.2.3. Content Value of Buildings	126
2.7. CONCLUSION	127
<u>CHAPTER 3 . METHODOLOGY</u>	<u>128</u>
3.1. TAX ASSESSOR’S DATA FOR SHELBY COUNTY	130
3.1.1. Generating Unique Identifiers for Tax Records.....	131
3.1.2. Single-family Residential Building Extraction.....	133
3.1.3. Multi-family and Commercial/Industrial Building Extraction	134
3.1.4. Imputation of Missing Data and Data Refinement.....	138
3.1.5. Spatial Representation of Extracted Buildings	141
3.2. SAMPLE DATA COLLECTION.....	143
3.2.1. Description of Sample Data	145
3.3. STRUCTURE TYPE CLASSIFICATION	149
3.3.1. Multinomial Logistic Regression	150
3.3.2. Design of Neural Network Topology for Classification	152
3.3.2.1. Multilayer Perceptron	153
3.3.2.2. Generalized Feed Forward Network	154
3.3.2.3. Modular Neural Network	155
3.3.2.4. Radial Basis Functions.....	156
3.3.2.5. Support Vector Machines.....	159
3.4. BUILDING FOOTPRINT CLASSIFICATION.....	159
3.4.1. Guidelines for Shape Classification Design Process	160
3.4.2. Preprocessing and Collinear Vertex Decimation	162
3.4.3. Orthogonalization of Polygon Edges by Corner Vertex Adjustment	163
3.4.4. Building Footprint Analysis by Landmark Correspondence	164
3.4.4.1. Computing Circularity Indexes to Eliminate Circular Buildings.....	166
3.4.5. Building Footprint Polygon Simplification	166

3.4.6. Identification of Salient Points.....	169
3.4.7. Derivation of Landmark Sequences by Contour Traversal.....	172
3.4.8. Binary Representation of Landmark Convexity	175
3.4.8.1. Determining Landmark Convexity	176
3.4.9. Building Footprint Classification.....	178
3.5. BUILDING VALUATION	178
3.5.1. Curve Fitting Routines for Model Building Square Foot Costs	180
3.5.1.1. Nomenclature for Model Buildings	182
3.5.2. Estimating Replacement Costs for Buildings	187
3.5.3. Structural and Non-Structural Replacement Costs	188
3.5.3.1. Recording Construction Assembly Costs for Model Buildings	191
3.5.3.2. Estimating Structural, Acceleration- and Drift-Sensitive Nonstructural Costs ..	191
3.6. ESTIMATING CONTENT VALUE	191
<u>CHAPTER 4 . RESULTS AND DISCUSSION.....</u>	<u>192</u>
4.1. STRUCTURE TYPE FROM MULTINOMIAL LOGISTIC REGRESSION.....	192
4.1.1. Multinomial Logistic Regression Model Specification	192
4.1.2. Model Performance	194
4.1.3. Relationships between Inputs and Structure Classes.....	196
4.2. STRUCTURE TYPE FROM NEURAL NETWORKS	201
4.2.1. Model Performance Evaluations.....	203
4.2.1.1. Interpreting the Confusion Matrix	203
4.2.1.2. Comparing Accuracy of Classification.....	207
4.2.1.3. Analysis of Misclassifications in the Models.....	211
4.2.1.4. Analysis of Classification Errors.....	218
4.2.1.5. Consequences of Classification Errors in Loss Estimation and Mitigation.....	219
4.2.1.6. Model Complexity, Sample Size and Model Calibration in Neural Networks ...	221
4.3. COMPARISON OF MULTINOMIAL LOGISTIC REGRESSION AND NEURAL NETWORKS.....	226
4.3.1. Differences and Relative Advantages of Multinomial Logistic Regression and Artificial Neural Networks	228
4.3.2. Using Artificial Neural Networks for Structure Type Classification	231
4.4. RECOGNIZING BUILDING FOOTPRINTS	233
4.4.1. Classification on Manually-digitized Building Footprints	233
4.4.2. Classification of Automatically-extracted Building Footprints.....	237
4.4.3. Notes on the Classification Algorithm	240
4.5. BUILDING VALUATION	241

4.5.1. Replacement Costs for Shelby County	241
4.5.2. Structural and Nonstructural Replacement Costs	244
4.5.3. Content Value	246
4.6. THE SHELBY COUNTY BUILDING INVENTORY DATABASE	246
<u>CHAPTER 5 . CONCLUSION AND VALIDATION.....</u>	<u>247</u>
5.1. VALIDATION OF SHELBY COUNTY BUILDING INVENTORY DATA	248
5.1.1. Validating Residential Data using Dwelling Unit Comparisons.....	248
5.1.2. Validation of General Building Stock Characteristics	252
5.2. APPLICABILITY OF RESEARCH METHODS TO OTHER FIELDS.....	258
5.3. IMPLICATIONS OF THE RESEARCH.....	259
5.3.1. Specific Implications	265
5.3.2. Limitations of the Research and Future Directions	268
5.3.2.1. Limitations in Structure Type Classification Modeling	268
5.3.2.1. Limitations in the Shape Recognition Application	269
5.3.2.1. Limitations in Building Valuation Modeling.....	272
<u>APPENDIX A . TABULATED SUMMARIES AND DESCRIPTIONS OF THE SHELBY COUNTY BUILDING INVENTORY.....</u>	<u>274</u>
<u>APPENDIX B . INFLUENCE OF INPUT VARIABLES ON STRUCTURE TYPE OUTCOME PAIRS IN THE MULTINOMIAL LOGISTIC REGRESSION MODEL</u>	<u>294</u>
<u>REFERENCES.....</u>	<u>300</u>

LIST OF TABLES

Table 1.1 -- Structure type classifications in HAZUS MH MR-3	19
Table 1.2 -- General and specific occupancy classes in HAZUS MH MR-3	20
Table 1.3 -- General occupancy to structure type mapping scheme (Tennessee).....	21
Table 1.4 -- A typical structural classification set for vulnerability modeling.....	27
Table 1.5 -- Building inventory attributes for vulnerability modeling	30
Table 2.1 -- Example raw confusion matrix for three classes (counts).....	69
Table 2.2 -- Example confusion matrix for three classes (percent accurate)	69
Table 2.3 -- Example confusion matrix for three classes (percent misclassified).....	69
Table 2.4 -- HAZUS MH MR-3 2002 Residential Replacement costs (in \$/sq. ft.)	119
Table 2.5 -- HAZUS MH MR-3 division of nonstructural elements and contents	124
Table 2.6 -- Taxonomy of household contents (Saeki et al 2000).....	125
Table 2.7 -- Content value as percentage Replacement cost (HAZUS MH MR-3)	126
Table 3.1 -- Shelby County Tax Assessor's database tables	131
Table 3.2 -- Parcel, Improvement and Building identifiers.....	132
Table 3.3 -- COMAPT extract for Parcel "001001 00026".....	135
Table 3.4 -- COMDAT extract for Parcel "001001 00026".....	135
Table 3.5 -- Cloned COMDAT extract for Parcel "001001 00026"	136
Table 3.6 -- COMINTEXT extract for Parcel "001001 00026"	137
Table 3.7 -- Imputations by occupancy categories.....	139
Table 3.8 -- Multi-family residential imputations of dwelling units by decade.....	140
Table 3.9 -- Multi-family residential imputations for dwelling size by decade.....	141
Table 3.10 -- Sample structure type frequency	146
Table 3.11 -- Sample occupancy type frequency	147
Table 3.12 -- Sample cross tabulation of Structure type by broad occupancy	148
Table 3.13 -- Sample cross tabulation of Structure type by decade.....	149
Table 3.14 -- Sample Structure type and Structural fire rating category	152
Table 3.15 -- Initial and final landmark sequence binary representation.....	176
Table 3.16 -- R. S. Means square foot costs - 1-3 story Apartment building (M.010) ...	180
Table 3.17 -- Curve parameter estimates for a 1-3 story Apartment.....	182
Table 3.18 -- R. S. Means specific occupancy and 3-digit code	183

Table 3.19 -- Standardization of Means External wall categories	185
Table 3.20 -- External walls from Tax Records reconciled with Means categories	186
Table 3.21 -- Generation of Replacement value identifiers	187
Table 3.22 -- Structural and Nonstructural cost breakdowns by Means models	190
Table 4.1 -- Variable specification for the Multinomial Logistic Regression	193
Table 4.2 -- Parameter estimates from the Multinomial Logistic Regression	195
Table 4.3 -- Variable specification for the ANN models.....	202
Table 4.4 -- Training performance evaluation using a confusion matrix (counts).....	205
Table 4.5 -- Testing performance evaluation using a confusion matrix (counts).....	206
Table 4.6 -- ANN training percent accuracy of Structure type classification	208
Table 4.7 -- ANN testing percent accuracy of Structure type classification.....	209
Table 4.8 -- Summary of errors (Structure type not recognized) by ANN model.....	211
Table 4.9 -- ANN training errors (Structure type not recognized) by model.....	212
Table 4.10 -- ANN testing errors (Structure type not recognized) by model	213
Table 4.11 -- Summary of errors (Structure type misclassified) by ANN model.....	215
Table 4.12 -- ANN training errors (Structure type misclassified) by model	216
Table 4.13 -- ANN testing errors (Structure type misclassified) by model.....	217
Table 4.14 -- Comparison of confusion matrices (Logistic vs. ANN).....	226
Table 4.15 -- Classification accuracy (Logistic vs. ANN).....	227
Table 4.16 -- Percent Structure type not recognized (Logistic vs. ANN).....	228
Table 4.17 -- Percent Structure type not predicted (Logistic vs. ANN).....	228
Table 4.18 -- Performance of shape recognition (Manual digitization).....	235
Table 4.19 -- Example classification errors for automatically extracted footprints	238
Table 4.20 -- Replacement cost by Structure type in Shelby County	242
Table 4.21 -- Imputation of Replacement costs (in millions of dollars) by HAZUS MH MR- 3 occupancy category	243
Table 4.22 -- Imputation of Replacement costs (in millions of dollars) by decade	244
Table 4.23 -- Total Nonstructural Replacement costs by Structure type for Shelby County (thousands of dollars)	245
Table 4.24 -- Average Nonstructural Replacement costs by Structure type for Shelby County (thousands of dollars)	245
Table 4.25 -- Content value by Structure type in Shelby County	246
Table 5.1 -- Validation of dwelling units by residential occupancy classes	250
Table 5.2 -- Validation using post-2000 single-family residential construction.....	251

Table 5.3 -- Validation of residential housing units by decade	251
Table 5.4 -- Comparison with NCEER report (Jones and Malik 1997).....	253
Table 5.5 -- Inventory validation by broad occupancy category	254
Table 5.6 -- GBS characteristics from HAZUS MH MR-3 and study inventory	255
Table 5.7 -- Replacement costs per square foot comparisons.....	257
Table A.1 -- Attribute schema for building inventory dataset	276
Table A.2 -- Mapping Tax Record-based specific occupancy to HAZUS MH MR-3 specific occupancy categories	277
Table A.3 -- (cont'd from previous) Mapping Tax Record-based specific occupancy to HAZUS MH MR-3 specific occupancy categories.....	278
Table A.4 -- Mapping HAZUS MH MR-3 occupancy categories to general occupancy classes	279
Table A.5 -- General Structure type (Frequency table).....	280
Table A.6 -- General occupancy classes (Frequency table)	280
Table A.7 -- Building counts and percentages by Structure type and General occupancy	281
Table A.8 -- Building percentages by Structure type and General occupancy.....	282
Table A.9 -- Building counts and percentages by Structure type and Number of stories	283
Table A.10 -- Building counts by Structure type and Year of construction (Decade)	284
Table A.11 -- Building counts by Structure type and Basement class.....	285
Table A.12 -- Building counts by General occupancy and Basement class.....	286
Table A.13 -- Building counts and percentages by Structure type and Area class	287
Table A.14 -- Building counts and percentages by Structure type and Replacement cost categories	288
Table A.15 -- Building Replacement costs (in millions of dollars) by Structure type and General occupancy	289
Table A.16 -- Building counts by Structure type and Content Value category	290
Table A.17 -- Building counts by Structure type and Essential Facility designation.....	291
Table A.18 -- Building counts by General occupancy and Number of Dwellings in structure	292
Table A.19 -- Building counts by Structure type and Number of Dwellings in structure	293
Table B.1 -- Influence of Height (Number of stories) on Factor change in Structure type Odds	294

Table B.2 -- Influence of Area (Square Feet) on Factor change in Structure type Odds	295
Table B.3 -- Influence of Year of Construction on Factor change in Structure type Odds	296
Table B.4 -- Influence of Wholesale Trade, Commercial Office and Bank occupancies on Factor change in Structure type Odds	297
Table B.5 -- Influence of Restaurant, Heavy Industrial and Multi-family residential occupancies on Factor change in Structure type Odds	298
Table B.6 -- Influence of Fire Rating descriptor on Factor change in Structure type Odds	299

LIST OF FIGURES

Figure 1.1 -- Probabilistic risk analysis for hazard management and decision making, as adapted from French and Isaacson (1984).....	12
Figure 1.2 -- Developing a Hazard Mitigation Plan, as adapted from FEMA (2002a)	14
Figure 1.3 -- Schematic process flow for a typical loss estimation model	30
Figure 2.1 -- An abstract artificial neuron with multiple inputs.....	48
Figure 2.2 -- Single layer feed-forward ANN topology – the Perceptron	49
Figure 2.3 -- Multiple layer feed-forward ANN topology	50
Figure 2.4 -- Schematic of ANN training and use.....	52
Figure 2.5 -- Some example transfer functions	53
Figure 2.6 -- General schematic for classifying samples into “p” classes	59
Figure 2.7 -- Weight optimization by minimizing the performance criterion.....	61
Figure 2.8 -- Stopping criterion using the cross-validation data set	67
Figure 2.9 -- General process stages in building footprint shape analysis.....	74
Figure 2.10 -- Problems in acquired building footprint polygons	75
Figure 2.11 -- Landmark or Contour representations of a polygon	77
Figure 2.12 -- Medial axis transformations from Voronoi tessellations.....	93
Figure 2.13 -- Medial axis transformations based on thinning routines.....	94
Figure 2.14 -- Deriving unique shape numbers for specific shape orders.....	96
Figure 2.15 -- Spaghetti data structure for feature representation	99
Figure 2.16 -- Topological data structure for polygon spatial data representation	100
Figure 2.17 -- Hierarchy of primitive part-geometries in a COM-based API	102
Figure 2.18 -- Densification by maximum deviation	104
Figure 2.19 -- Densification by maximum angle.....	105
Figure 2.20 -- Densification of linear features	105
Figure 2.21 -- Eliminating features during generalization.....	108
Figure 2.22 -- Simplifying lines and polygon edges during generalization	108
Figure 2.23 -- Aggregating polygon features during generalization	109
Figure 2.24 -- Exaggerating features for visual clarity during generalization	109
Figure 2.25 -- Collapsing using size or dimensionality reduction for generalization.....	110

Figure 2.26 -- Translating features in conflict resolution during generalization	110
Figure 2.27 -- Removing and moving features in typification during generalization	111
Figure 2.28 -- Refining feature geometry during generalization	111
Figure 2.29 -- Symbolizing lower level into higher groups during generalization	112
Figure 2.30 -- The Douglas-Peucker algorithm for line simplification	114
Figure 2.31 -- Limitations of the Douglas-Peucker algorithm for orthogonal edges	116
Figure 3.1 -- Research methodology described by a schematic process	129
Figure 3.2 -- Extract of parcels and buildings in Central Memphis	142
Figure 3.3 -- Survey sample collection areas in Shelby County	144
Figure 3.4 -- Schematic of MLP network for structure type classification	154
Figure 3.5 -- Schematic of GFF network for structure type classification	155
Figure 3.6 -- Schematic of MNN for structure type classification	156
Figure 3.7 -- Linear combination of RBFs used for approximation	157
Figure 3.8 -- Schematic representation of RBF network	158
Figure 3.9 -- Removal of collinear vertices from polygon edges	162
Figure 3.10 -- Orthogonalizing building edges by adjusting corners	164
Figure 3.11 -- Footprint classification by landmark correspondence	166
Figure 3.12 -- Performance of built-in building simplification tools in GIS	168
Figure 3.13 -- Simplification failure artifacts and desired results	169
Figure 3.14 -- Desired landmark outputs that mimic human judgment	171
Figure 3.15 -- Ambiguity in building footprint polygon classification	172
Figure 3.16 -- Landmark convexity sequences for polygon footprint classes	174
Figure 3.17 -- Determining if a polygon vertex is convex or concave	177
Figure 3.18 -- Parametric curves for a 1-3 story steel frame Apartment	181
Figure 3.19 -- Structural/Nonstructural costs as percent Replacement costs	189
Figure 4.1 -- Influence of covariates on structure type	198
Figure 4.2 -- Influence of fire rating on structure type	199
Figure 4.3 -- Influence of occupancy on structure type (part 1)	199
Figure 4.4 -- Influence of occupancy on structure type (part 2)	200
Figure 4.5 -- Structure type classification accuracy by ANN model type	210
Figure 4.6 -- Structure type classification errors by ANN model type	218
Figure 4.7 -- Structure type sensitivity to input variables (ANNs)	233
Figure 4.8 -- Preprocessing manually-digitized footprints	234
Figure 4.9 -- Example classification errors for manually digitized buildings	236

Figure 4.10 -- Preprocessing automatically extracted building footprints.....237
Figure 4.11 -- Examples of misclassifications for automatically extracted footprints239
Figure 4.12 -- Successful classifications for automatically extracted footprints239
Figure 4.13 -- Examples of edge noise in automatically extracted footprints240
Figure 5.1 -- Residential housing units by decade252

LIST OF ABBREVIATIONS

ANN(s)	Artificial Neural Network(s)
C	Concrete Structure
CBE	Consequence-based Engineering
CRM	Consequence-based Risk Management
DMA 2000	The Disaster Mitigation Act of 2000
FEMA	Federal Emergency Management Association
GBS	General Building Stock
GFF	Generalized Feed Forward Network
GIS	Geographic Information System
HMGP	Hazard Mitigation Grant Program
IFSAR	Interferometric Synthetic Aperture Radar
LIDAR	Light Detection and Ranging
MAEC	Mid-America Earthquake Center
MAT	Medial Axis Transformations
MLP	Multi-Layer Perceptron
MMI	Modified Mercalli Index
MNN	Modular Neural Network
MSE	Mean square error
MTB	Memphis Test Bed
PC1	Precast Concrete Structure
PC2	Concrete Tilt-up Structure
PE	Processing Element
RADAR	Radio Detection and Ranging
RBF	Radial Basis Functions
RM	Reinforced Masonry Structure
S1	Steel Frame
S3	Light Metal Frame
SAR	Synthetic Aperture Radar
SVM	Support Vector Machines
URM	Unreinforced Masonry Structure
W	Wood Frame Structure
W1	Light Wood Frame Structure
W2	Commercial/Industrial Wood Frame Structure

SUMMARY

The process of modeling earthquake hazard risk and vulnerability is a prime component of mitigation planning, but is rife with epistemic, aleatory and factual uncertainty. Reducing uncertainty in such models yields significant benefits, both in terms of extending knowledge and increasing the efficiency and effectiveness of mitigation planning. An accurate description of the built environment as an input into loss estimation would reduce factual uncertainty in the modeling process.

Building attributes for earthquake loss estimation and risk assessment modeling were identified. Three modules for developing the building attributes were proposed, including structure classification, building footprint recognition and building valuation. Data from primary sources and field surveys were collected from Shelby County, Tennessee, for calibration and validation of the structure type models and for estimation of various components of building value. Building footprint libraries were generated for implementation of algorithms to programmatically recognize two-dimensional building configurations. The modules were implemented to produce a building inventory for Shelby County, Tennessee that may be used effectively in loss estimation modeling.

Validation of the building inventory demonstrates effectively that advanced technologies and methods may be effectively and innovatively applied on combinations of primary and derived data and replicated in order to produce a bottom-up, reliable, accurate and cost-effective building inventory.

Chapter 1 . INTRODUCTION

This dissertation forms part of the Mid-America Earthquake Center's [henceforth MAEC] ongoing efforts to create innovative research-based solutions that mitigate the impacts of earthquakes particularly in Mid-America. The MAEC, one of three national earthquake engineering research centers established by the National Science Foundation, has predicated its overall approach on a new engineering paradigm to seismic risk reduction termed Consequence-based Engineering [CBE] that essentially quantifies risk to "societal systems" (Mid-America Earthquake Center 2006) and subsystems on a regional basis, thereby allowing policy-makers to ultimately develop risk reduction strategies and implement mitigation actions. The approach later termed Consequence-based Risk Management [CRM] (Abrams et al. 2002), explicitly includes uncertainty in a framework that facilitates comparisons of mitigation alternatives in terms of their impact on properties and populations at risk from earthquake disasters.

Earthquakes, like all natural hazards have potentially enormous, even catastrophic impacts. These impacts are measured in terms of casualties, direct property losses and losses to other assets, and even indirect economic consequences. Determining the consequences of earthquake events relies on accurate at-risk data, damage models and an understanding of the underlying geophysical processes that lead to their occurrence. The need for accurate risk assessment and mitigation planning tools presents both an enormous challenge and opportunity for the application of advanced technologies – *problematic*, particularly because of the uncertainty rampant throughout the entire risk modeling process, and *beneficial* in terms of information that could potentially guide policy (National Research Council 2006). Thus, the critical

challenge is to **better** understand, anticipate and reduce earthquake risk by integrating the potential consequences into the mainstream planning process. Risk analysis models that demonstrate variations in hazards and resultant damage can prove to be vital and effective in informing policy decisions (French and Isaacson 1984). In addition, planning itself can be particularly effective in mitigating the consequences of natural disasters by guiding the location and design of urban structures (Godschalk et al. 1998) and building vital social capital in terms of a community base that encourages hazard mitigation (Burby and May 1998).

The research efforts and outcomes described and developed in this dissertation are particularly vital because an accurate physical inventory forms a primary factual component in the overall risk analysis process. Increasing precision in the distribution of structures and populations contained in those structures enables effective risk assessment, which is critical to rational decision making, both in emergency preparedness and mitigation planning. Further, if local governments are to play a greater role in reducing community vulnerability, building inventories produced from models calibrated on samples drawn from the local area would allow decision makers to become familiar with the spatial distribution of vulnerable structures and critical assets while increasing overall accuracy. Policies for effective risk reduction and plans for emergency response may then be designed with greater efficiency at the local level.

In specific terms, this dissertation derives three critical building inventory components for risk assessment modeling, including

- classification of buildings by structural type
- classification of buildings by two-dimensional shape configuration, and
- valuation of building components and systems

Structure type and shape configuration will be used to model risk through building behavior under earth-shaking stresses, while values of building components and systems will be used for quantifying potential losses. The structure type distribution is estimated using artificial neural networks, perhaps for the first time in building inventory estimation. Shape recognition is achieved by specific smoothing and classification algorithms implemented by innovatively manipulating building footprint polygon geometry in the GIS environment. Standard construction industry square footage to construction cost ratios are parameterized for building occupancy, area, height, structure type and external wall combinations through curve fitting routines and these equations are used to estimate valuation of building components and assemblies, including structural, nonstructural acceleration- and drift-sensitive and content values.

1.1. Background for disaster mitigation

It has become painfully obvious that there is an urgent and escalating need for developing, validating and implementing accurate and cost-effective methods to identify the vulnerability of the man-made environment in the context of both natural and technological hazards. Natural disasters like Hurricane Katrina in 2005 (Cable News Network 2005), cyclone Nargis (Tun 2008) and the earthquake in China (British Broadcasting Corporation 2008) in 2008 and their terrible death tolls only as serve stark reminders of the vulnerability of life on earth, even in today's technologically advanced world. Our world's resources are increasingly being concentrated, both demographically and economically in natural hazard-prone regions. In the United States, the population in areas exposed to hurricanes has quadrupled since 1970, with over 70 million people in about 439 communities living permanently along hurricane-prone coastlines along the Atlantic Ocean and the Gulf of Mexico. Hurricanes cause over 20 deaths and result in damages of over \$ 5.1 billion annually (Congressional Hazards Caucus 2007b).

Increased development in floodplains that local government has failed to curtail (Burby et al. 1999) coupled with an increase in the frequency of heavy rain events over the last fifty years has resulted in increased flooding-related deaths of about 100 and losses of over \$ 5 billion per year (Congressional Hazards Caucus 2007a). The US also experiences thousands of earthquakes, and about seven annually with magnitudes over 6.0 on the Richter scale. Over 75 million Americans in 39 states face significant risks from earthquakes, and in terms of costs, earthquakes can be genuinely catastrophic – in fact, FEMA’s costs for the 1994 Northridge earthquake was close to \$ 7 billion, more than the combined relief costs of Hurricanes George, Andrew, Floyd and the 1993 Midwest floods (US General Accounting Office 2003; National Science and Technology Council 2005). Earthquakes in the US cost over \$ 5.6 billion annually, with a single event having the potential to cause losses of more than a \$ 100 billion (National Research Council 2006).

In the United States, the responsibility for public health and safety lies with the State Governments, as specified by the Constitution. When disasters occur, first the locally affected jurisdiction attempts to manage the incident(s). If local resources are overwhelmed, then the mayor of the locality will request additional help and resources from the state to combat the disaster. Continuing along the hierarchy, if local and state resources are insufficient to handle the disaster, the governor of the state requests a Presidential Disaster Declaration and federal assistance (Bea 1998). This clearly established hierarchy experienced over so many disasters, has resulted in the public viewing emergency management as a fundamental governmental function. Problems in emergency management are solved by enacting legislation, but have been historically reactive (National Research Council and the Division on Earth and Life Studies 2006), until the Stafford Act of 1988 and the Disaster Mitigation Act of 2000. Refer to Bea

(1998) and Haddow et al (2008) for a listing of historical disaster-related legislation.

Some landmark federal legislations in disaster management include:

- The Flood Control Act of 1934 that allowed the Army Corps of Engineers to design and build flood control projects.
- The National Flood Insurance Act of 1968, motivated by the fiscal losses incurred in Florida and Louisiana following the swaths of devastation caused by Hurricane Betsy in 1965 (Haddow et al. 2008). This act also created the National Flood Insurance Program, a supposedly self-supporting program that was intended to protect owners against flood losses and reduce future losses in the community through floodplain management ordinances. Unfortunately, elements of this program have actively encouraged development in flood prone areas by renters predominantly (federal projects that build dams and levees), have not actively dissuaded development in hazardous areas (through insurance subsidies, disaster relief payments and tax write-offs) and have provided incentives for hazard prone occupation by persons that are least likely to recover from flood losses (Burby et al. 1999).
- The Disaster Relief Acts of 1969 and 1974, following Hurricane Camille in 1969 (Waugh 2000) and flooding related losses in Pennsylvania and New York following Hurricane Agnes in 1972. These acts established a process for Presidential Disaster Declarations and provided relief assistance to local governments and individuals (May 1985).
- Creation of the Federal Emergency Management Agency, FEMA, in 1979, by President Carter, who consolidated the over 100 federal organizations and entities involved in disaster relief (Haddow et al. 2008).

- The Stafford Disaster Relief and Emergency Assistance Act of 1988 (FEMA 2007), which attempted to generate efficiency and order to the process of conducting physical and monetary federal disaster relief aid to state and local governments through FEMA, and for the first time, encouraged the development of mitigation activities before the onset of disasters through the Pre-Disaster Mitigation Grant Program.
- The Disaster Mitigation Act of 2000 (FEMA 2000), which clarifies the special efforts needed to assist disaster-affected states in the process of rendering aid, emergency services and the reconstruction and rehabilitation of distressed areas, and provides funding for promoting public-private partnerships, identifying the community's hazard vulnerabilities and establishing mitigation priorities.

The Disaster Mitigation Act of 2000 requires additional mention. Based on the Robert T. Stafford Disaster Relief and Emergency Assistance Act (Bea 1998; FEMA 2007), the Disaster Mitigation Act of 2000 (FEMA 2000) is the latest legislation to improve the mitigation planning process and was put into motion on October 10, 2000. The Disaster Mitigation Act of 2000 (DMA 2000) reinforces the importance of *mitigation planning* and *emphasizes planning for disasters before they occur*, by establishing a pre-disaster hazard mitigation program and new requirements for the national post-disaster Hazard Mitigation Grant Program (HMGP). It allows HMGP funds to be used for "planning activities", and increases HMGP funding to states that have developed a comprehensive, enhanced mitigation plan prior to a disaster. Mitigation plans are required to demonstrate that their proposed mitigation measures are based on a sound planning process that accounts for the risk to and the capabilities of the individual communities. Based on DMA 2000 requirements, typical mitigation plans contain explanations of the planning process and community involvement, detailed descriptions

of hazards and vulnerability of assets and populations in the jurisdictions, scenarios to quantify consequences, and policy recommendations to reduce the potential impacts of the hazards.

In a recent National Institute of Building Sciences report, researchers found that “money spent on reducing the risk of natural hazards is a sound investment. On average, a dollar spent by FEMA on hazard mitigation (actions to reduce disaster losses) provides the nation about \$4 in future benefits” (Multihazard Mitigation Council 2005a, pp. iii). Thus, mitigation is significantly cost-effective, enough to justify federal funding before disasters and during post-disaster recovery, most successful when systematically executed on a long-term, community-wide and comprehensive basis with better information and institutional commitment, and requires further evaluation for efficient implementation (Multihazard Mitigation Council 2005a, 2005b). In places where mitigation activities are taken seriously, they yield substantial benefits (Burby 1994, 1998) – thus, mitigation can potentially be institutionalized by either standalone programs or by integrating them with the normal planning process and several jurisdictions, particularly in the west coast, have incorporated “Seismic Safety” elements within their comprehensive planning process (Burby 1998; Burby et al. 1999). However, the costs involved in developing mitigation plans at the local level coupled with myopic past federal policies that subsidize development in hazard-prone areas tend to dissuade local jurisdictions from taking a lead role in hazard mitigation policy planning and more importantly, provide disincentives for local jurisdictions to regulate urban development in high risk areas (Burby et al. 1999). In a more recent article in the context of the surprising devastation of hurricane Katrina, Burby (2006) explains that policies of the federal government have substantially increased the potential for catastrophic losses and local governments do not develop policies towards reducing risk and vulnerability.

Measuring and quantifying earthquake consequences (or losses) present serious conceptual and methodological challenges. Major obstacles to describing and analyzing earthquake impacts include (i) the lack of reliable data, (ii) inadequate, poorly specified, inconsistent or undependable models and (iii) the levels of compounding uncertainty rampant throughout the analytical processes – these problems exist at all scales, from international to local levels (National Research Council 2006). This is perhaps another reason why decision makers are extremely reluctant to enact policies based on the relationship between earthquake risks and local development – decisions to counter low-probability high-consequence disasters are perceived to inhibit local economic development and political careers. While most states and jurisdictions satisfy (and more importantly, aim to satisfy) the bureaucratic requirements of DMA 2000 through a separate mitigation planning process, the quality of these mitigation plans have not been sufficiently analyzed. Depending on the DMA 2000 template and the distribution of hazards, these mitigation plans tend to have a boiler-plate appearance and may not be effective enough. While in these plans, hazards are described and located in considerable and accurate detail, most communities do not have the resources to accurately quantify their assets and/or populations, and instead use free or readily available (at coarse resolutions and often inaccurate) data to quantify the vulnerability of their built environment. Leveraging scale economies in integrating mitigation efforts with mainstream planning, and developing low-cost, reliable and accurate accounts of local assets (the primary focus of this research effort) and demographics would certainly alleviate and improve the quality of these mitigation efforts.

1.2. The need for accurate urban inventories

The primary purpose of disaster risk modeling is to use the results from the modeling process to guide plans that reduce vulnerability, mitigate consequences and

respond/recover from disasters. The process of disaster risk modeling is rife with uncertainty originating from several sources. First, the current state of scientific knowledge in terms of disasters and their effects on the built environment is incomplete, leading to epistemic uncertainty (Ellingwood 2007). Second, modeling by its very nature, simplifies and approximates real-world conditions in order to accomplish tractable implementations, leading to uncertainty in the estimates. Again, this source of uncertainty may be classified as epistemic (ibid). Third, there is an element of randomness, both in terms of the areal coverage of the disaster event and the particular behavior of the built environment under stress, that is rarely captured in disaster modeling efforts, leading to aleatoric uncertainty (ibid). Finally, disaster modeling requires factual information about demographics, the natural and built environment at risk (Burby 1998). If these inventories are inaccurate, or arrived at through other modeling processes, estimates produced by the disaster modeling efforts would also be inaccurate –this can be termed factual uncertainty.

Recent research suggests that there are varied and substantial economic benefits in reducing uncertainty in disaster modeling primarily through loss-avoidance regulations and strategies, better engineering design and code enforcement and more effective hazard mitigation planning. While reducing uncertainty could be expensive, the potential benefits would be substantially more than the cost (Multihazard Mitigation Council 2005a; National Research Council 2006). Prior to any mitigation implementation, a primary consideration requires that local communities identify and quantify the population and built asset inventory at risk (FEMA 2001), in order to frame effective policies to redirect growth away from currently vulnerable structures to less hazard-prone areas (Burby 2006).

In the specific context of risk assessment modeling, there are three important reasons why urban building inventories are required. First, different classes of buildings behave differently under dynamic stresses during disasters, and an effective inventory would create key building classes that are uniquely different from one another in terms of disaster response. Second, accurate accounts of the urban building inventory in terms of their counts and replacement costs would reduce the uncertainty inherent in the modeling process and augment the reliability of the risk estimation. Finally, estimation and analysis of injuries, casualties, shelter needs, debris generation and removal, direct losses and indirect economic impacts are based on estimated physical damage to buildings, and an inaccurate inventory would lead to cascading inaccuracies in the downstream aspects of the loss estimation process.

Thus, for both reducing uncertainty in disaster modeling and for effective mitigation planning, a necessary precondition is the availability of an accurately quantified account of the built inventory. This dissertation is primarily concerned with reducing factual uncertainty in the development of urban building inventories, a substantial component of the urban built environment.

1.3. Hazard Mitigation in the Planning Process Framework

There is general agreement that hazard mitigation should influence urban development in order to reduce disaster-related damage and losses, and help the affected communities rebound from the disaster quickly (Burby and May 1998; Godschalk et al. 1998; Burby et al. 1999). Recent federal papers advocate goals aimed at improving data collection and prediction capability along with “the development and widespread use of improved hazard and risk assessment models and their incorporation into decision support tools and systems” (National Science and Technology Council 2003) with the overall objective of reducing disaster vulnerability. Internationally and

nationally, there is consensus on implementing disaster reduction policies aimed at guiding development into less hazard prone areas and enabling communities to be resilient to natural hazards (United Nations 2001; Godschalk and Baxter 2002; United Nations 2003, 2005). In the local context, local jurisdictional disaster policy making tends to be more reactive than proactive and communities that have experienced a disaster are more likely to analyze risk and enact mitigation plans/policies to their constituencies and assets (Berke 1998; Burby and May 1998; Briechle 1999).

While federal policies and a top-down influence on hazard mitigation can ensure attention to mitigation efforts, many researchers argue that risk analysis and mitigation planning should be primarily a bottom-up effort that will account for local awareness and negotiated outcomes from local interests (Reddy 2000; Pearce 2003; Cutter 2005). In fact, Pearce (2003) argues that integrating the disaster management plan with the comprehensive planning process that includes public participation has the highest probability of success, and further, active public participation in the mitigation process is increased substantially when the plan development is broken down into smaller, neighborhood scales (Godschalk et al. 1999). Mitigation is also less expensive when integrated early in the comprehensive planning process, rather than in a standalone process (Godschalk et al. 2003; Pearce 2003). Further, it should be emphasized that hazard mitigation planning, whether incorporated as part of the comprehensive planning process or as a standalone process, requires a “strong factual basis” (Kaiser et al. 1995) in terms of the spatial location of the hazards as well as the spatial distribution of the community’s physical, social, economic and infrastructural assets. Brody (2003) argues that public awareness of hazards is a precondition for participation in any hazard mitigation planning process. This dissertation suggests that vulnerability information may be communicated more effectively within the hazard mitigation planning process if

plan-makers have an accurate accounting of their community’s assets in the context of the hazards.

Public policy making in the context of hazard mitigation and hazard risk analysis is a special case of the “general problem of decision making under uncertainty” (French and Isaacson 1984). They describe a schematic process for developing hazard-related policies in the context of probabilistic earthquake hazards, as outlined in Figure 1.1 below.

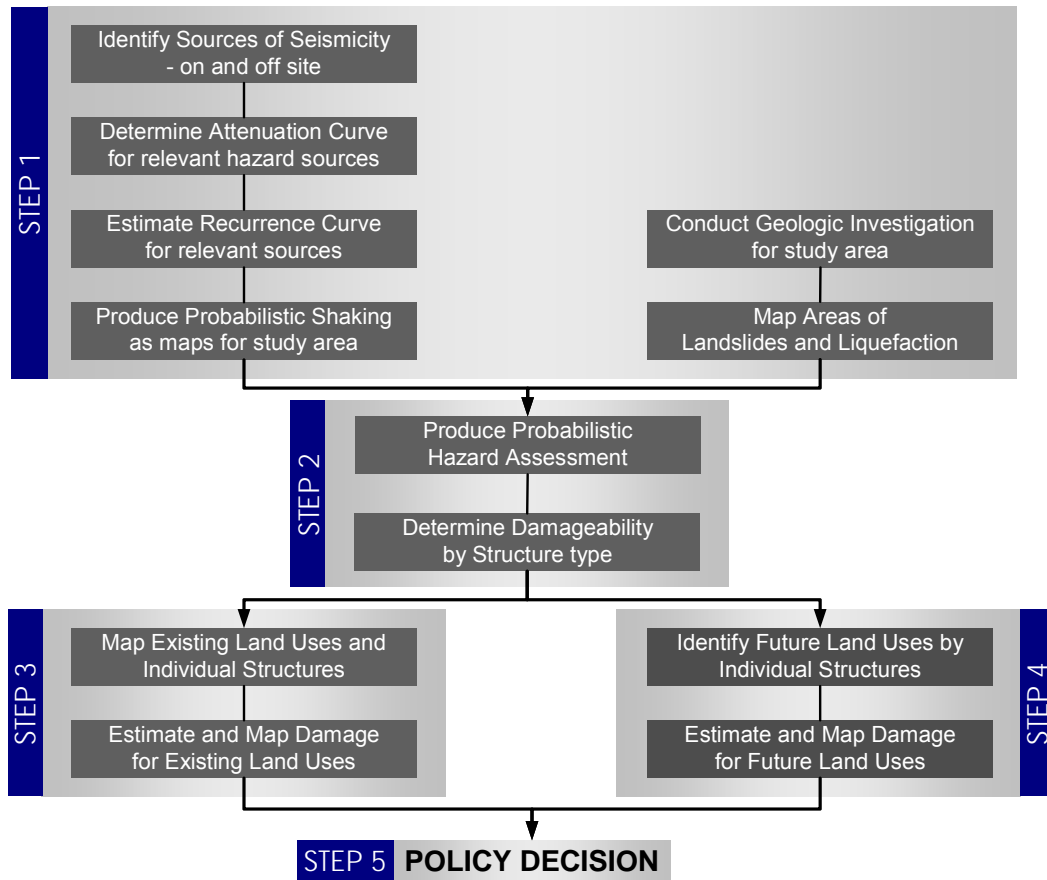


Figure 1.1 -- Probabilistic risk analysis for hazard management and decision making, as adapted from French and Isaacson (1984)

The various sequential steps in the process are recognized as (i) identifying hazard characteristics, (ii) modeling the probabilistic hazard in terms of ground motion effects on specific structure types, (iii) creating an account of the inventory and subjecting it to the hazard to estimate damage, (iv) projecting the inventory to various alternative futures and subjecting them to the hazard to estimate future scenario-based damage and (v) develop policy.

Expanding this framework, prior to hazard mitigation, it is important to analyze and understand the risks posed by the hazard. Hazard risk analysis involves the interaction of the hazard and human activities. In other words, it is the exposure of humans and their activities to the hazard that underlines the risk. After the risk is understood, then mitigation activities that limit human exposure and vulnerability can be conducted.

Typically, mitigation employs several tools to reduce the devastating consequences of disasters, variously classified under structural (Beatley and Berke 1992; Nelson and French 2002), non-structural (Godschalk et al. 1999; Godschalk and Baxter 2002), communicative (Burby et al. 1999; Olshansky 2001; Godschalk et al. 2003) and economic (Berke 1995b, 1995a; Burby et al. 1999) that one may find in the typical comprehensive plan. Additionally, it should be noted that despite its specificity, disaster mitigation planning is a form of planning and mitigation planners at local, state and federal levels need to follow a basic planning framework, including goal development, factual bases, development of alternatives, public participation, implementation, monitoring, evaluation and updating (Kaiser et al. 1995). Accordingly, Figure 1.2 below shows the various typical stages in the normal process of mitigation planning, as adapted from FEMA (2002a).

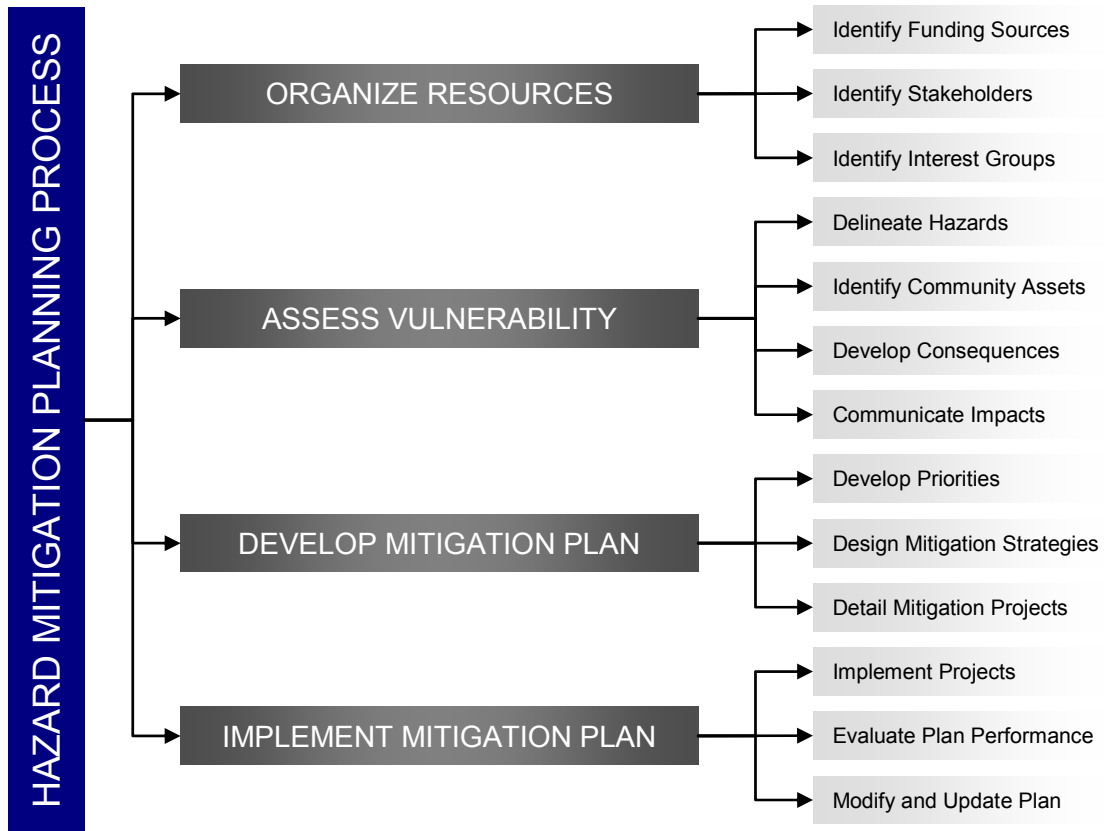


Figure 1.2 -- Developing a Hazard Mitigation Plan, as adapted from FEMA (2002a)

Although Figure 1.2 suggests a clean and linear process, owing to the fact that mitigation planning is not yet a formal or mainstream local government function, in reality the process is more piece-meal and non-linear and tends to develop in sporadic spurts, particularly where either local code enforcement or local comprehensive planning are not required (Burby 2006).

Burby (2006) argues that the surprisingly large devastation in New Orleans and the overall tendency towards more disasters with greater consequences is entirely predictable owing to “well-intentioned, but short-sighted, public policy decisions at all levels of government,” increasing the vulnerability of populations and assets to natural disasters. While federal policies and disaster supplements are unlikely to change, local

government can seize the initiative and design and implement policies that redirect populations away from at-risk urban infrastructure. Further, if federal governments require local governments to include natural hazard mitigation into their comprehensive plan-making process (beyond the bureaucratic requirements of mitigation planning under DMA 2000) and be more financially responsible for consequences in local urban development planning, then local governments would invest more resources into making effective comprehensive plans that enable safe urban growth and development. In other words, recent evidence clearly suggests that hazard risks may be substantially reduced if mitigation planning were to become a part of local government function (Burby et al. 1998; Olshansky 2001; Nelson and French 2002; Burby 2005).

1.4. Existing methods for Urban Inventory Data Collection and Limitations

Inventory databases of the built environment are at best fragmented and contain little information about the structure type of the building, which is a crucial input in risk assessment and modeling (French and Muthukumar 2006). Attribution of the building inventory by structure type and occupancy class would enable risk models to predict damage to the inventory, and subsequently estimate the direct and indirect social and economic losses associated with a particular hazard scenario. Other relevant characteristics related to building inventories in the context of hazard modeling include location, height, value, tenure, area, year of construction, three-dimensional mass distribution and two-dimensional plan configuration. Inventory information could also be used to estimate the extent of damage in an actual event and enable the efficient use of resources for response efforts. Further, accurate inventory information modeled against a historic event could enable loss model calibration, increase our understanding about building behavior under stress and reduce epistemic uncertainty in risk modeling.

Finally, accurate and detailed building inventories could provide valuable input into designing policies and prioritizing projects related to hazard mitigation.

1.4.1. Urban Inventory Data Sources

There are no national databases characterizing the built environment. There are a number of disparate sources of urban inventory information relevant to social or economic analyses such as the US Census, American Housing Survey, County Business Patterns, Woods and Poole Economics, Dun and Bradstreet, local Employment Surveys, County Tax Assessors, etc. In general, these sources do not contain building information in a form fit for disaster risk modeling, and are generally used to infer or derive building inventory data (RMS & CUREe 1993). While the US Census does collect and release information on residential buildings, commercial and industrial building related information are generally unavailable. County Tax Assessors and local governments collect information for taxes and for local development, but such data are often characterized by large gaps (tax-exempt property information is not maintained by the tax assessor). However, tax assessors' data often are rich sources containing at the very least, area, age, use and value of buildings, and may be mined or creatively integrated with other inventory derivation methods or techniques (Jones et al. 1987). For instance, tax assessors' data for each building could be geocoded within a geographic information system (GIS) and combined with remotely-sensed topographic data to derive height of buildings. There are no clear standards on what kind of building information has to be collected at any level, and data collection is performed on an ad-hoc basis. Finally, such building related information rarely contains structural details, which is a primary input in risk modeling.

Collecting building inventory information from various disaggregated and distributed sources would prove to be prohibitively expensive. Therefore, researchers

will have to develop indirect methodologies in order to develop reliable building inventories based on readily available or direct data. Various research efforts have been directed towards quantifying and classifying urban inventories at various levels of geography. Typically, existing and easily available data are collected and used to estimate the general stock of buildings, and tabulating them by structure type and building use. One of the first such tabulations was developed by the Applied Technology Council (ATC-13 1985) for the state of California, using several databases accessed from the Federal Emergency Management Association and Bureau of Economic Analysis repositories. ATC-13's system of cross-classifying buildings by "Social Function Class" or use and "Earthquake Engineering Class" or structure type has generally been adopted for minimum building inventory data.

1.4.2. Classification of the Urban Building Inventory

Studies of the urban building inventory classified by structure and use in different areas show consistent patterns despite considerable differences in demographic and economic structure. Malik (1995) derived building inventory estimates for Memphis-Shelby County and Wichita-Sedgwick County from the tax assessor's records and classified building use under agricultural, commercial and industrial, educational, hospital, institutional, government and residential. Similarly, structure type was classified as wood, light metal, masonry, reinforced concrete, protected steel and unclassified. Despite Shelby County's nearly double demographic count over Sedgwick County, residential proportions accounted for nearly 90% in both cases. Cross-tabulated classifications were found to be within 2% for all categories of use and structure type. The domination of the building inventory by residential structures carried over to the wood structure type, since most residential units are built on wood frames. Malik also indirectly estimated the general building stock (henceforth GBS) counts for Shelby

County using demographic-building stock relationships generated by Jones (1978) for Sedgwick County and compared the counts with tax records. The estimates generated were fairly reliable, but more consistent in square footage than with counts. GBS distributions tended to be more consistent for use than structure type for widely dispersed geographic areas. Savonis' (1985) research also supported building inventory patterns that were dominated by single family residential units, with residential buildings accounting for almost 90% of the total building inventory.

Largely following ATC-13, most building inventories for risk modeling tend to be derived GBS collections classified by structure type and/or occupancy, or detailed building inventories collected by field surveys or inspections of construction documents, often collected for critical facility buildings.

1.4.3. Building Inventory Development in HAZUS MR-3

While it is true that there is no national building inventory, FEMA, along with their loss estimation software HAZUS MR-3 (and in previous versions), deliver “modeled” general and specific building stock databases for the entire continental US. The application extracts the inventory data for a specific study region and converts it into building stock classified by structure type and use, following FEMA Earthquake Hazard Mitigation (FEMA 2002b) conventions. The GBS is classified by general occupancy under agricultural, commercial, educational, government, industrial, religious and residential buildings. The GBS is meant to be used for modeling the probability of damage to all the occupancy types for flood, wind and earthquake hazards. The application also includes default parameters and routines to convert the GBS general occupancy categories to specific occupancy classes and structure types. Table 1.1 shows the structure type classification (FEMA 2002b) in HAZUS MR-3. Table 1.2 shows the 7 general and 33 specific occupancy categories in HAZUS MR-3.

Table 1.1 -- Structure type classifications in HAZUS MH MR-3

S. No.	Code	Description	Height			
			Range		Typical	
			Name	Stories	Stories	Feet
1	W1	Wood, Light Frame (<5,000 sq. ft.)		1 - 2	1	14
2	W2	Wood, Commercial and Industrial (>5,000 sq. ft.)		all	2	24
3	S1L	Steel Moment Frame	Low-Rise	1 - 3	2	24
4	S1M		Mid-Rise	4 - 7	5	60
5	S1H		High-Rise	8+	13	156
6	S2L	Steel Braced Frame	Low-Rise	1 - 3	2	24
7	S2M		Mid-Rise	4 - 7	5	60
8	S2H		High-Rise	8+	13	156
9	S3	Steel Light Frame		all	1	15
10	S4L	Steel Frame with Cast-in-Place Concrete Shear Walls	Low-Rise	1 - 3	2	24
11	S4M		Mid-Rise	4 - 7	5	60
12	S4H		High-Rise	8+	13	156
13	S5L	Steel Frame with Unreinforced Masonry Infill Walls	Low-Rise	1 - 3	2	24
14	S5M		Mid-Rise	4 - 7	5	60
15	S5H		High-Rise	8+	13	156
16	C1L	Concrete Moment Frame	Low-Rise	1 - 3	2	20
17	C1M		Mid-Rise	4 - 7	5	50
18	C1H		High-Rise	8+	12	120
19	C2L	Concrete Shear Walls	Low-Rise	1 - 3	2	20
20	C2M		Mid-Rise	4 - 7	5	50
21	C2H		High-Rise	8+	12	120
22	C3L	Concrete Frame with Unreinforced Masonry Infill Walls	Low-Rise	1 - 3	2	20
23	C3M		Mid-Rise	4 - 7	5	50
24	C3H		High-Rise	8+	12	120
25	PC1	Precast Concrete Tilt-Up Walls			1	15
26	PC2L	Precast Concrete Frames with Concrete Shear Walls	Low-Rise	1 - 3	2	20
27	PC2M		Mid-Rise	4 - 7	5	50
28	PC2H		High-Rise	8+	12	120
29	RM1L	Reinforced Masonry Bearing Walls with Wood or Metal Deck Diaphragms	Low-Rise	1 - 3	2	20
30	RM2M		Mid-Rise	4+	5	50
31	RM2L	Reinforced Masonry Bearing Walls with Precast Concrete Diaphragms	Low-Rise	1 - 3	2	20
32	RM2M		Mid-Rise	4 - 7	5	50
33	RM2H		High-Rise	8+	12	120
34	URML	Unreinforced Masonry Bearing Walls	Low-Rise	1-2	1	15
35	URMM		Mid-Rise	3+	3	35
36	MH	Mobile Homes		all	1	10

Table 1.2 – General and specific occupancy classes in HAZUS MH MR-3

Label	Occupancy Class	Example Descriptions
	RESIDENTIAL	
RES1	Single-family Dwelling	House
RES2	Mobile Home	Mobile Home
RES3	Multi-family Dwelling RES3A Duplex RES3B 3-4 Units RES3C 5-9 Units RES3D 10-19 Units RES3E 20-49 Units RES3F 50+ Units	Apartment/Condominium
RES4	Temporary Lodging	Hotel/Motel
RES5	Institutional Dormitory	Group Housing (dormitory), Jails
RES6	Nursing Home	
	COMMERCIAL	
COM1	Retail Trade	Store
COM2	Wholesale Trade	Warehouse
COM3	Personal and Repair Services	Service Station/Shop
COM4	Professional/Technical Services	Office
COM5	Banks	
COM6	Hospital	
COM7	Medical Office/Clinic	
COM8	Entertainment & Recreation	Restaurants/Bars
COM9	Theaters	Theaters
COM10	Parking	Parking Garages
	INDUSTRIAL	
IND1	Heavy	Factory
IND2	Light	Factory
IND3	Food/Drugs/Chemicals	Factory
IND4	Metals/Minerals Processing	Factory
IND5	High Technology	Factory
IND6	Construction	Office
	AGRICULTURE	
AGR1	Agriculture	
	RELIGIOUS	
REL1	Church/Non-profit	
	GOVERNMENT	
GOV1	General Services	
GOV2	Emergency Response	Police/Fire/EOC
	EDUCATION	
EDU1	Grade Schools	
EDU2	Colleges/Universities	Does not include group housing

The application provides “mapping schemes” for converting general occupancy to specific occupancy and cross-tabulating occupancy classes with basic structure types – a default mapping scheme for Tennessee is seen in Table 1.3.

Table 1.3 -- General occupancy to structure type mapping scheme (Tennessee)

Specific Occupancy Type	General Structure Types					Total
	Wood	Concrete	Steel	Masonry	Manufactured Housing	
RES1	90%	-	-	10%	-	100%
RES2	-	-	-	-	100%	100%
RES3A	75%	-	-	25%	-	100%
RES3B	75%	-	-	25%	-	100%
RES3C	75%	-	-	25%	-	100%
RES3D	75%	-	-	25%	-	100%
RES3E	75%	-	-	25%	-	100%
RES3F	75%	-	-	25%	-	100%
RES4	50%	-	-	50%	-	100%
RES5	20%	45%	-	35%	-	100%
RES6	90%	-	-	10%	-	100%
COM1	30%	10%	30%	30%	-	100%
COM2	10%	30%	30%	30%	-	100%
COM3	30%	10%	30%	30%	-	100%
COM4	30%	10%	30%	30%	-	100%
COM5	30%	10%	30%	30%	-	100%
COM6	-	70%	10%	20%	-	100%
COM7	30%	10%	30%	30%	-	100%
COM8	30%	10%	30%	30%	-	100%
COM9	-	45%	40%	15%	-	100%
COM10	-	70%	30%	-	-	100%
IND1	-	25%	70%	5%	-	100%
IND2	10%	30%	30%	30%	-	100%
IND3	10%	30%	30%	30%	-	100%
IND4	-	25%	70%	5%	-	100%
IND5	10%	30%	30%	30%	-	100%
IND6	30%	10%	30%	30%	-	100%
AGR1	10%	30%	30%	30%	-	100%
REL1	30%	10%	15%	45%	-	100%
GOV1	15%	17%	35%	33%	-	100%
GOV2	14%	16%	24%	46%	-	100%
EDU1	10%	12%	17%	61%	-	100%
EDU2	14%	19%	20%	47%	-	100%

Cross-tabulations of general occupancy and basic structure type are created through mapping schemes that suggest a breakdown by percentage of each specific occupancy class into different basic structure types (wood, concrete, steel and masonry) by state, thus serving as a control for the breakdown of the specific occupancy categories to detailed structure types through another set of mappings.

To serve as input into a risk assessment model, the building inventory needs to be classified into specific sets that represent adequately the average characteristics and behaviors of all the buildings grouped in those sets. In other words, each defined class of building should exhibit substantially different damage behavior and loss characteristics. HAZUS MR-3 defines attributes of these classes using the structural system, height and design level (structural capacity and response parameters), nonstructural acceleration and drift-sensitive building components, specific occupancy (for casualties, business interruption and content damage), regional building practices and aleatoric intra-class variability. The classification is implemented as a cross-tabulation of specific occupancy (see Table 1.2) and detailed structure type (see Table 1.1). General occupancy classes are converted into specific occupancy classes based on the breakdown of specific occupancy floor area ratios by census tract. These floor area breakdowns are based on demographic and housing characteristics for residential buildings and Dun and Bradstreet Inc. business data for non-residential buildings. Using the general occupancy and basic structure type cross-tabulation as a control, the square footages of the various occupancy classes are distributed across the various detailed structure types, based on distributions for specific regions such as the East Coast, West Coast and the Mid-West. Square footage values in the cross tabulations between specific occupancy and structure type are then converted into building counts based on per square foot occupancy ratios. Building counts are then converted into structural,

non-structural and content value costs based on replacement values for the specific buildings.

Note that the classifications thus generated are based on general “mapping schemes” or percentage breakdowns or other parameters, based on manipulations of census, business, energy consumption, proprietary insurance data and expert opinion. The mapping schemes and conversion parameters are crude and based on a coarse geographic resolution and are better suited for large regional loss estimation. In fact, the technical manuals clearly acknowledge the coarseness of the default inventory modeling and suggest its use more as a “guide” to develop building distribution schemes for specific regions of interest (FEMA - DHS 2007, pp. 3-6). FEMA’s intention was to provide government agencies at all levels the opportunity to use a regional loss estimation application at relatively no cost, and therefore distributed the relevant GBS and other inventory data along with the application software. However, the software does contain tools to enhance the quality of the GBS by incorporating more accurate local inventory information.

Of course, in ideal circumstances, information on all relevant variables pertaining to the built inventory would be collected at the finest resolution (that of the individual building) and available for risk modeling. In this dissertation, I propose developing building inventory variables using models calibrated on local data that would eliminate the coarseness of large-area mapping schemes to smaller areas and increase the accuracy of the building inventory accounts. Thus, a typical urban building inventory tries to tabulate buildings into typologies whose behavior under dynamic stresses are similar, and the task is certainly not trivial since these classifications occur along several dimensions such as structure type, occupancy type, height, square footage and design

levels, and perhaps particular building characteristics (symmetry, massing, the number of concavities in the footprint, etc.).

1.5. Advanced Inventory Technologies and Techniques for Data Collection

Advanced technologies in the context of data collection cover a wide range of sophisticated mechanisms including data processing hardware and software, sensors, platforms, data storage, retrieval and analysis instruments (Tralli 2000). Advanced aerial and spaceborne remote sensing technologies are providing higher resolution data at lower costs, and the spatial image-based information thus generated provides numerous opportunities for developing base inventory data, or at the very least, supplements efforts at generating urban inventories. Advances in computing, database and data analysis systems enable faster processing of larger volumes of data and the development of new analytical processing techniques and models for all types of research.

1.5.1. Remote Sensing Technologies

Remote sensing technologies refer to all forms of airborne or spaceborne platforms with active or passive sensors for the capture of details on the earth's surface. Passive sensors capture reflected radiation from the earth's surface (optical and infrared sensors), while active sensors send signals and receive their reflections from the earth's surface. Applying these technologies has several benefits. First, depending on the resolution, fairly detailed urban inventory data could be captured efficiently and cheaply. Second, using such information provides opportunities for the development of automated routines and algorithms for image processing and feature extraction. Third, repeated images of the same area over different periods would enable the characterization of the temporal aspects of urban inventory and help identify growth patterns, and ultimately

inform mitigation and land use planning relative to hazards. Finally, these technologies could vitally assist in describing the hazard potential of large areas, and enable rapid damage assessment, response prioritization, loss estimation and model calibration in the aftermath of a disaster.

Passive sensors typically deliver photogrammetric data in the optical and infrared bands at relatively high resolutions (1 to 5 meters) and are typically used for land cover extraction and classification and the development of elevation models (and building heights) when data is captured in stereo mode (successive image pairs that overlap, enabling relief detection when viewed using stereoscopes). Additionally, by using image-processing applications integrated with feature-based data such as roads, parking lots, water bodies, etc., pixels may be trained and classified into signature-based classes. Building footprint feature extraction could potentially be automated, and the footprints analyzed by height, size and shape in order to generate building inventory by use. Primary advantages of using this technology include support for manual or automated planimetric feature extraction and that it is well-understood (Mollander 2000).

Active sensors send microwave or laser pulses towards the earth's surface and measure the time taken for the signal to be reflected and its intensity. Synthetic Aperture Radar (SAR) and Interferometric Synthetic Aperture Radar (IFSAR) use microwave pulses and record both phase and amplitude information. IFSAR is similar to SAR, except that two antennae are used, and the resulting composite image (two images are formed since the same signal reflection is received at different phases and magnitudes by each antenna) may be processed in order to extract elevation information (Gabriel and Goldstein 1988; Rodrigues and Martin 1992). Light Detection and Ranging (LIDAR) emits laser pulses and receives their reflections (one pulse may be reflected off several features such as building sides and then bare earth and could potentially have

as many as 6 returns). Along with a set of sophisticated instruments including an inertial measurement unit to measure platform velocity, a high quality global positioning system, and a high-precision clock, the time for the returns and their intensities are recorded and processed for elevation and classification data. LIDAR produces extraordinarily large profusions of data points that may be effectively processed and used to determine heights of structures as well aid in automated feature extraction (Fowler 2000).

Typically, the spatial data produced by these advanced technologies may be stored, processed, retrieved, analyzed and visualized using a GIS. When combined with sophisticated relational databases and programming, GIS could enable the design and implementation of specific models and specialized routines to combine large volumes of advanced and traditional data in order to develop reliable, accurate and cost-effective accounts of the built environment.

1.5.2. Building Inventory Estimation Methods

The expense in collected structure type data for an entire region through field surveys or inspections of construction documents is often prohibitive and the speed of structural database construction is too slow relative to the frequent changes in urban buildings (modifications, retro-fits, demolition, etc.). The most effective approach would be to innovatively integrate data from several sources and use them as a basis for structure type inference. The starting point for developing urban building inventory datasets is usually the collection of readily available primary data such as roads, demographics, imagery and tax records, preferably in spatial formats. Since structure type for buildings is not collected, a finite set of structural systems is essential for risk modeling. Structural classification schemes range from as few as four (French and Isaacson 1984) to twelve (ATC-21 1988) or even forty (ATC-13 1985). A typical

structural classification set is listed in Table 1.4. Structure types may be estimated by one of two generic methods, including knowledge-based rules and classification models.

Table 1.4 -- A typical structural classification set for vulnerability modeling

Structure Type Code	General Structure Type
C1	Concrete Moment Resisting Frame
C2	Concrete Frame with Concrete Shear Wall
MH	Manufactured Homes
PC1	Concrete Tilt-up Panels
PC2	Precast Concrete Frame
S1	Steel Frame
S3	Light Metal Frame
RM	Reinforced Masonry
URM	Unreinforced Masonry
W1	Light Wood Frame
W2	Commercial Wood Frame

1.5.3.1. Knowledge-based Rules

Typically based on a structured analyses of primary data and calibration samples, relationships between structure type, building age, occupancy, height and location are derived and statistical correlations, frequencies and cross-tabulation instruments are used in inferring structure types for the rest of the population of buildings (French et al. 1992). While the performance accuracy varies by structure class, databases produced are eminently compatible with risk modeling applications. The knowledge base is therefore a set of conditional rules acting upon known correlations or tabulations generated from the database. Typically, these methods (in fact most methods) do not easily discriminate between wood, masonry and concrete structure types for older, non-residential buildings (ibid). Similar rules reflecting both statistical aspects and construction methods may be applied to larger populations of buildings at larger scales in disparate geographic areas.

1.5.3.2. Classification Models

While there are several advanced classification methods to choose from, typical methods include cluster analysis, self-organized mapping, supervised parametric classification, support vector methods, multinomial logistic regression and artificial neural networks. Cluster analyses, self-organized mapping and supervised classification methods generally examine the data and implement clusters of the input spaces through parametric methods available in several statistical software. Multinomial logistic regression is a maximum likelihood estimation method where a dependent variable consisting of more than two independent alternatives may be identified by a set of explanatory variables (McFadden 1974). Classifications are based on the relative probabilities of each alternative relative to the others generated through a continuous logistic function of the inputs (Aldrich and Nelson 1984). The method is relatively complex and the best models are extremely parsimonious in the choice of independent variables and the dependent alternatives. Based on a calibration sample of structure types, the parameters of the multinomial logistic regression model could be used to predict the structure type for the remaining population of buildings.

Support Vector methods are typically used for classifications where the number of calibration samples is too few relative to the explanatory variables to generate reliable parametric estimates. This method artificially expands the list of explanatory variables by generating interactions and transformed functions from the independents and then discriminating between the dependents at this higher dimension input space. The premise is that any two classes may be linearly classified by transforming the input space into higher dimensions (Vapnik 1999).

Artificial neural networks (ANN in the singular and ANNs in the plural) offer great potential as classification and analytical tools in generating urban building inventories.

ANNs have been used with GIS to model land use change (Li and Yeh 2002), agricultural soil protection (De la Rosa et al. 2004) and landslides (Neaupane and Achet 2004). The number of inputs, their complex interrelationships, the inherent noise and gaps in input data often inhibit traditional parametric modeling efforts. ANNs have the capability to integrate and generalize such wicked inputs and successfully generate classification functions (Principe et al. 2000). ANNs are non-linear, semi-parametric computer models that create parameters through learning mechanisms for successful or desired results based on calibration samples and could replicate the input patterns to classify unseen buildings into the specified structure type classes. ANN classification results are generally robust and forgiving of complex or noisy input data.

1.6. The Earthquake Modeling Process Requirements

In a typical risk modeling and loss assessment model, vulnerability functions for different components of the urban at-risk inventory are applied based on a particular level of hazard. This generates estimates of physical damage to the infrastructure, which are used as inputs to compute direct estimates of damage losses, shelter needs and functionality interruptions. Based on repair and restoration of service parameters applied to the damaged components, business interruption and long term economic losses are computed, and all results are reported and/or visualized. Figure 1.3 details the schematic process flow for such a model. This dissertation is focused on developing the building inventory for input into a risk assessment and loss estimation model. Based on discussions with principal investigators of several other MAEC projects, the key building attribute components that would serve as inputs to earthquake vulnerability models were identified and are listed in Table 1.5, along with their potential sources.

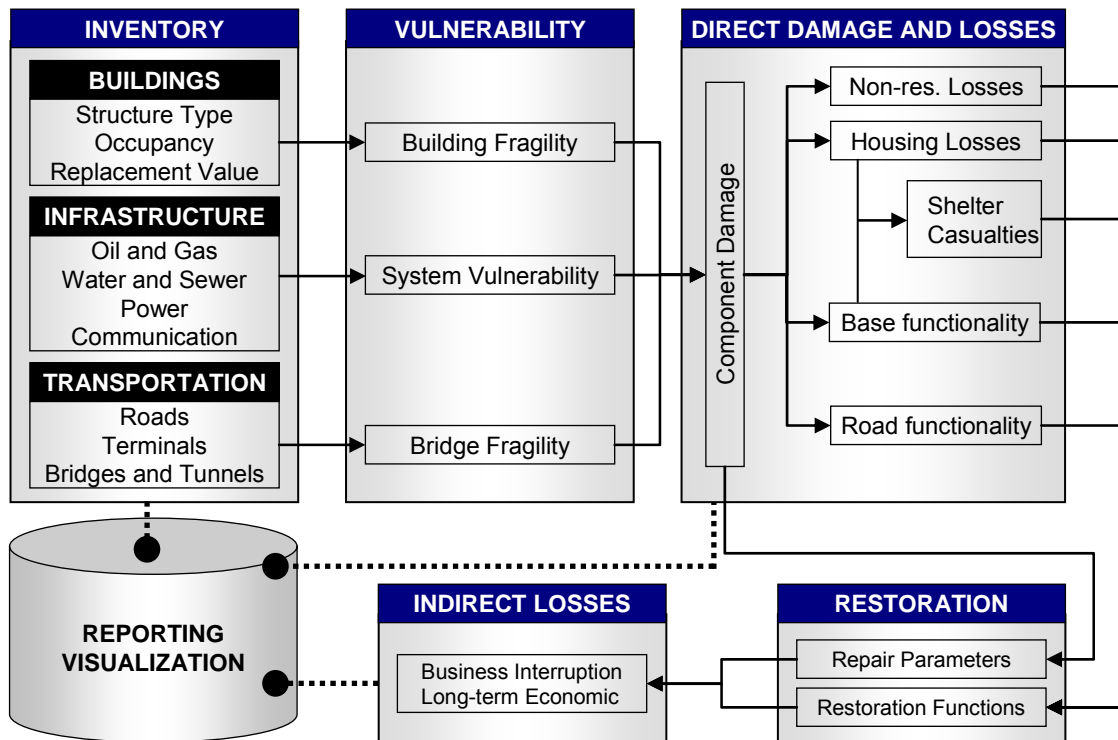


Figure 1.3 -- Schematic process flow for a typical loss estimation model

Table 1.5 -- Building inventory attributes for vulnerability modeling

Attribute	Source
Structure type classification	Estimated through knowledge-based or classification models
Building footprint configuration	Derived through automated GIS-based shape recognition routines
Height	Classified from primary sources (Tax Assessor's Database)
Building floor area	Primary (Tax Assessor's Database)
Year of construction	Primary (Tax Assessor's Database)
Building location	Primary (Tax Assessor's Database)
Building occupancy (use)	Primary (Tax Assessor's Database)
Building replacement value	Parameterized from R.S. Means costs, structure type and height
Structural replacement value	Estimated through occupancy-based component costs
Non-structural replacement value	Estimated through occupancy-based component costs
Content value	Estimated through occupancy-based component costs
Essential facility designation	Primary (Tax Assessor's Database/Internet/Other)

The building footprint configuration attribute requires some special mention here. Apart from structure type, height, location and design level that influence the building's

capacity and response during a hazard event (particularly for wind and earthquake), building behavior is also influenced by its shape (Arnold and Reitherman 1982) and distribution of mass (Murty 2002 a, 2002 b). The shape of the building is essentially the shape of the building footprint polygon that represents it in two dimensions. The distribution of mass is more difficult to extract, since it involves the masses of the unobservable contents inside the building. However, what is observable is the exterior massing of the building, in terms of the distribution of height over the footprint of the building. For instance, an L-shaped building may be 6 stories in height along the short arm of the L and only 2 stories along the long arm. Clearly, the distribution of mass inside the L-shaped footprint is not uniform. Estimating or measuring the distribution of mass is beyond the scope of this dissertation, which will limit itself to deriving automated methods in order to identify the shape of a building from its footprint and reconcile this information with the building inventory database. Based on discussions with principal investigators of other MAEC projects, the various two-dimensional building configurations to be identified in this dissertation include square, rectangular, L-, C-, T-, H-, Z-shaped, octagonal, circular, cruciform (plus-shaped) and irregular.

Structure type, building footprint configuration, height and location identified in this dissertation will be used as inputs for the estimation of direct physical damage and losses. Structure type classifications listed in Table 1.4 were deemed adequate for regional vulnerability modeling, based on discussions with other MAEC project principal investigators. Building floor area, occupancy, replacement and content value estimated in this dissertation will be used for social and economic loss modeling and mitigation decision support.

1.7. Description of Research

In order to estimate the consequences of any disaster, a necessary and required input is a quantitative description of the man-made environment that is exposed to that disaster (Kaiser et al. 1995). In fact, most consequence-based models first estimate the damage to the physical man-made inventory and then translate these estimates in order to estimate the engineering, social and economic consequences (FEMA 2004; FEMA - DHS 2007).

1.7.1. Research Statement

This research focuses primarily on the factual basis requirement of mitigation planning – that is, the development of urban building inventories and their attributes using advanced technologies and methods, including multinomial logistic regression models, artificial neural networks and innovative spatial computing and secondarily demonstrates the application of the inventory in specific earthquake scenarios for regional loss estimation.

1.7.2. Research Goals and Objectives

The overall goal of the research is to utilize advanced technologies and methods in order to identify physical attributes of the built environment that are instrumental in assessing potential earthquake damage, consequences and mitigation strategies. Thus, the research is concerned with the development of new techniques in estimating quantifiable descriptors of the urban building inventory, relying on remote sensing, aerial photography, GIS and other advanced technologies. Inferential techniques will be developed for combining data from such measurements with data from secondary sources that describe parts of the inventory at risk.

The dissertation has several objectives, including:

- development of new techniques and replicable methods for producing inventories of buildings and other facilities over large urban regions
- production of reliable, low-cost inventories of the built environment to make earthquake risk assessment more cost effective
- creation of comprehensive and efficient inventory techniques based on the integration of remote sensing, aerial photography and inferential techniques applied to secondary data
- implementation of the techniques developed in the dissertation to produce the building inventory for a test study in Shelby county, Tennessee, henceforth referred to as the Memphis Test Bed [MTB]
- use the generated building inventory to specified earthquake scenarios in MTB for proof-of-concept, earthquake risk assessment database development and application in regional loss assessment modeling

1.7.3. Significance of the Research Effort

The research will provide a detailed assessment of existing and emerging data mining technologies in the context of developing the factual basis for hazard mitigation planning. Additionally, individual technologies and related efforts currently used in other disciplines will be evaluated for their potential in providing useful information for regional earthquake vulnerability and risk assessment. By developing alternative approaches and integrating multiple methods that use these new technologies and methods, the process of creating and maintaining databases of large-scale urban and regional inventories can be made more reliable and cost effective. The methodologies will be

applied in Shelby County, Tennessee, and may then be replicated over other regions in order to test the generalizability of these techniques. The project can potentially develop new and innovative data collection and inventory modeling techniques in pre- and post-earthquake event periods for risk assessment scientists, structural engineers and decision makers.

With specific reference to the CRM paradigm, the research provides methods for rapid screening and assessment of broad based urban inventory data, including the location, type and function of particularly vulnerable structures. Advanced inventory techniques developed within the project will increase that availability of data for regional damage synthesis and provide inputs for alternative scenarios in consequence minimization and mitigation planning efforts. Improvements in inventory techniques will make damage modeling and analysis more reliable and affordable, and more widely applicable, and further, provide the basis for post-earthquake assessment and recovery planning. With specific reference to artificial neural network techniques, neural networks are very new applications and have wide applicability in planning, particularly where prediction of categories based on other external factors is needed – for instance, in growth models, allocating future uses of land using environmental, transportation, proximity and other developmental factors can be performed using neural networks. Finally, data inventories developed with these new techniques, when combined with visualization techniques, can help decision makers appreciate risk and develop more informed policies for minimizing earthquake-related damage.

Specifically, this research develops methods for modeling and estimating three distinct substantive components of building inventories that are vital to risk assessment modeling. Advanced techniques using artificial neural networks on buildings calibrated at the local level enable the identification of building structure type. Innovative spatial

computing algorithms classify buildings based on their two-dimensional shape configuration. Standard industry-based construction costs are parameterized into equations for detailed estimations of the value of building components and systems based on configurations of building occupancy, area, height, structure type and external wall type.

While the research develops methods that increase the reliability of building inventories, the results of models for structure type classification and building valuation have general and specific mitigation policy implications. The spatial distribution of the building inventory enables the identification of particularly vulnerable structure groupings and the concentration of building asset wealth in the context of hazard-prone areas. The spatial distribution of the building inventory then allows for (i) strategic retrofitting based on life safety (ii) land use planning and regulations for building stock turnover and redevelopment, (iii) design guidelines and code enforcement in the context of improvements to existing structures, (iv) development management for directing new growth and (v) other loss avoidance/minimization guidelines.

In summary, this research effort fulfils a vital requirement of risk assessment modeling that directly informs mitigation policy through the spatial distribution of the building inventory and indirectly, through estimates of potential damage and quantification of vulnerability derived by applying the building inventory in risk assessment models. Additionally, the methods developed in this research will increase the reliability of the building inventory while reducing the cost of inventory development and are eminently replicable, with great potential for automated and semi-automated approaches.

1.7.4. Scope of Research

Based on the previous sections, building-related attributes required for effective earthquake risk assessment and loss estimation modeling include structure type, 2D shape, height, floor area, year of construction, location, occupancy, replacement, structural, non-structural and content value and essential facility designation. This dissertation will develop new and replicable techniques for generating building inventories by integrating primary data with inferential techniques and innovative methods. The dissertation will also demonstrate the application of the techniques in order to generate the building inventory database for the MTB and use the database in loss estimation exercises.

Accordingly, the scope of this dissertation includes the following:

- examining the current literature and state-of-the-art technology for designing the techniques
- estimating the structure type for buildings identified in Table 1.4 for the MTB
- developing spatial computing algorithms and implementing them in the GIS environment in order to identify the building configuration type
- estimate the replacement, structural, non-structural and content value of buildings for the MTB
- identify all essential facility buildings for the MTB

1.8. Organization of Dissertation

This dissertation is organized into five chapters. The introductory chapter details the need for accurate urban inventories in the context of advanced technologies and

techniques that could be used to inform hazard mitigation planning. The introduction also defines the research and outlines the overall scope.

Chapter 2 reviews the current state of the literature for classification models, shape and pattern recognition and building valuation. In particular, the classification section deals with multinomial logistic regression and artificial neural network models. The shape recognition section describes the process of shape analysis, and brief explanations of techniques categorized by shape representation or recognition methods. The literature also covers geometric manipulations within a GIS framework for methods that exhibit potential in spatial computing. The chapter concludes with a review of prevalent methods in building valuation.

Based largely on the literature review, Chapter 3 describes the methodological approach and design of particular aspects of the study for each of the three modules outlined in the scope. The chapter begins with a description of the available primary data and the details of a field survey for calibration and validation exercises. Classification methods for determining structure type of buildings are then discussed, including multinomial logistic regression and ANN approaches. In order to programmatically identify building footprint configurations, the chapter proceeds to describe various algorithms that enable preprocessing the footprint to serve as input into a shape recognition module. The overall design of the process is emphasized. Finally, the methodology for estimating the replacement costs, the nonstructural acceleration- and drift-sensitive component costs and the content value of the building is described.

Chapter 4 describes the results of the structure type classification models, the shape recognition routines and the models for estimation of building values. This chapter also discusses various aspects of the results, including where they may have policy implications.

Chapter 5 concludes the dissertation by validating the integrated building inventory data produced by the application of the methods designed in the research. This chapter also discusses the applicability of the methods to other substantive areas, limitations of this research and future areas for directed study.

The dissertation also includes two appendices that describe (a) the Shelby County demonstration building inventory produced by the research in the form of tabulated summaries, and (b) the influence of explanatory variables used in the multinomial logistic regression on structure type outcome pairs in the form of changes in odds.

Chapter 2 . LITERATURE REVIEW

This dissertation is primarily concerned with estimating the structure type of buildings using existing and primary data, identifying the shape of building footprint polygons and estimating the value of buildings using advanced technologies and methods described in Section 1.5 earlier. While the individual elements themselves are quite disparate and distinct from one another, what unifies them is that they are all components of the urban building inventory that will be used for various elements of earthquake risk assessment and loss estimation. Identifying the structure type and the shape of the building both require methods of pattern recognition, while building valuation involves statistical curve-identification routines. The discrete aspects of the variables require distinct approaches and this is reflected in the organization of this chapter and indeed, the dissertation itself.

This section begins with a brief description of classification as related to pattern recognition, in the contexts of typical classification and shape-based classification. A statistical multinomial logistic regression approach to classification is described in the next section. A relatively recent and advanced approach to classification using artificial neural networks is described in Section 2.3. The literature for identifying building configuration is surveyed in the following section. Aspects of geometric manipulation in the GIS environment particularly as they relate to configuration identification are described in Section 2.5. The use of ANNs has been somewhat limited in developing building inventories and building configurations have not yet been used in large-scale regional loss estimation. Additionally, the audiences interested in building inventory development have little access to the evolution and application of such techniques. Consequently, the material presented in these sections is comprehensive in nature. The

section concludes with some background in building valuation and the components of a building that are sensitive to acceleration and drift aspects of earthshaking hazards.

Most of the literature for structure type classification and building shape recognition is adapted from pattern analysis research, particularly in the context of digital recognition and machine intelligence. Specializations include fuzzy computing, pattern recognition, image processing, medical imaging, multimedia, signal processing, neural networks, computer graphics, robotics and artificial intelligence, etc. (Costa and Cesar 2001b).

2.1. Pattern Recognition and the Potential for Automation

One part of this dissertation is aimed at classifying buildings to specific structure types, based on some primary attributes of the buildings, such as building age, function, height, size, etc., while another attempts to identify and organize buildings by their two-dimensional representation. These are classification problems often encountered in human activity. In cases where critical or high-precision decisions have to be made especially in the context of repetitive or tedious visual inspection, humans are prone to fatigue or error (Fabel 1997). For instance, Transportation Security Administrations have the extraordinarily high responsibility of achieving a 100% success rate in terms of spotting security threats like weapons or explosives. The screening job is not hard (“How hard can it be?” was a frequent response when I discussed this with my peers) but tedious and repetitive – how long would a TSA screener be able to stare at a monitor without dozing or having blurred vision or suffering some lapse in concentration? TSA screeners in tests at 15 airports missed 90% of security threats during covert tests (Sherman 2007). Such situations provide the need and opportunity for automating decision-making in order to reduce errors or to increase precision and consistency in classification.

The term classification includes any context where a judgment is made to assign an object to one of several classes, based on existing information. The classification procedure is therefore an application to analyze a set of observed attributes for one sample among many and then assign that sample to one of a set of pre-defined classes. Creating this classification procedure from a sample set for which the correct classes are known has also been coined “pattern recognition, discrimination, or supervised learning” (Michie et al. 1994). Humans are often able, without conscious thought, to identify patterns and classify objects well. However, in today’s world, there is intense pressure to develop systems or machines that can perform the same classification task with higher accuracy, or greater speed, or greater economy, or simply to release humans from repetitive effort (Devijver and Kittler 1982). Thus, classification procedures are aimed at mimicking or exceeding human judgment with the added benefits of consistency, explanatory power and generalization (Duda and Hart 1973; Devijver and Kittler 1982; Michie et al. 1994).

Michie et al (1994) identify three historical research traditions for classification problems including statistical, machine learning and neural networks.

Statistical approaches (the oldest methods to identify structure from samples) for classification are based on discriminant functions or joint distributions of sample attributes within each class, and usually have explicit underlying probability distributions. This approach relies on some degree of human intervention for attribute choice, measurement and transformation. See Anderson (1984b) or McLachlan (1992) for standard textbooks that deal with statistical approaches to pattern recognition. See also Jain et al. (2000) for an excellent review on modern statistical pattern recognition. Statistical models for classification also include multinomial logistic regressions, extended from binary dependent variable regressions. Statistical approaches often

require assumptions about underlying population distributions that may not be valid and the nature of parameterization makes the modeling process somewhat inflexible. Nevertheless, statistical models have had wide applicability, and have consistent parameters that may be used for explanation. In addition, statistical models have clear measures of uncertainty that would be useful to know in a loss estimation process that has uncertainty stemming from multiple sources.

Machine learning, which emerged from research groups in artificial intelligence and computer science (Russell and Norvig 1995), also attempts to identify classification procedures, usually by learning from binary labels and known examples (Langley 1996). Machine learning is often implemented through decision-trees, where classification is achieved from hierarchical binary paths, though other advanced procedures such as genetic algorithms (Luger 2002; Association for the Advancement of Artificial Intelligence 2007) and inductive logic procedures (Srinivasan 2001) are not uncommon. Machine learning is expected to generate classification mechanisms that are explicit and provide comprehensible explanation for human understanding (Michie et al. 1994). After initial development, machine learning does not require human intervention. See Nilson's draft textbook (1996) for an excellent survey on machine learning, statistical learning, neural networks and inductive logic programming. Typical classification algorithms implemented in this tradition include k-nearest neighbors, decision trees and support vector machines.

ANNs are increasingly becoming common in tasks that involve functional approximation and classification, as evident from the number of ANN-based research papers in pattern recognition, medical statistics and other applied disciplines, particularly in the last two decades (Dreiseitl and Ohno-Machado 2002). Just as in statistics and

machine learning, ANNs also require the presentation of pattern examples showing the desired results that serve as calibration or training data.

ANNs will be discussed in greater detail in the following sections, and this section concludes with Michie's comment (1994) that neural networks are performance-based, and integrate statistical complexity with the machine learning aim of mimicking human judgment without the necessity for explicit explanation (Anderson and McNeil 1992).

2.2. Multinomial Logistic Regression for Classification

A binary logistic regression is a form of regression where the dependent variable is dichotomous. When the dependent variable is polytomous (has more than two categories), the multinomial logistic regression model is used. The multinomial logistic regression is different from an ordinal logistic regression in that the categorical dependent variable in the ordinal logistic regression is ordered. For instance, if the dependent variable analyzes a set of preferential responses, where 1 is "Excellent", 2 is "Good" and 3 is "Average", you would use an ordinal logistic regression. If the dependent variable lists clear categories, such as 1 is "Republican", 2 is "Independent" and 3 is "Democrat" (though some would argue that this categorical variable is also ordinal!), you would use a multinomial logistic specification. The ordinal logistic regression is a specific case of multinomial logistic regression (Anderson 1984a), where the model performance is definitely better if the discrete dependent variables are indeed ordered (Campbell and Donner 1989).

Identifying the appropriateness of a category (the dependent variable) for a particular combination of inputs is a classification exercise. The multinomial logit model has its earliest applications in transportation to classify individual mode choice, based on the respective utility functions for each transportation mode (McFadden 1974). Refer to

Aldrich and Nelson (1984) or Greene (Hensher et al. 2005; Greene 2008) for introductory texts and applications of logistic regression. Also, see J. Scott Long (Long 1997; Long and Freese 2006) for text and examples of working with categorical dependent variables using Stata, a very useful source.

Logistic regressions are often the preferred models for classification tasks. Consider 'n' observations, (y_i, \mathbf{X}_i) , where y_i are conditionally independent (J+1) categorical dependent variables, and \mathbf{X}_i are the covariates or independent variables. In the multinomial logistic regression, each outcome is modeled, or a set of parameters are identified for each category of y . Thus if there were three discrete categories of y (represented by P, Q and R), then the conditional probability for outcome P is given by

$$\Pr(y = P) = \frac{e^{X\beta^{(p)}}}{e^{X\beta^{(p)}} + e^{X\beta^{(q)}} + e^{X\beta^{(r)}}$$

To identify this model, one of the outcomes is set as the reference category by arbitrarily setting the estimated coefficients to 0. The same conditional probability for outcome P is

$$\Pr(y = P) = \frac{1}{1 + e^{X\beta^{(q)}} + e^{X\beta^{(r)}}$$

The general model is specified by comparing the J different outcomes to the reference category $J = 0$ and is given by

$$\Pr(y = J | X_i) = \frac{e^{X\beta^{(j)}}}{1 + \sum_{k=1}^J e^{X\beta^{(k)}}}, \text{ where } j = 0, 1, 2, \dots, J \text{ and } \beta_{(j)} = 0 \text{ when } j = 0$$

The relative probability of $y = k$ to the reference outcome $y = 0$ is

$$\left(\frac{\Pr(y = k)}{\Pr(y = 0)}\right) = e^{X\beta^{(k)}}$$

Performance measures for multinomial logistic regression models include a Wald statistic or the likelihood ratio, a pseudo R-squared statistic and finally, each coefficient

may be associated with a confidence interval. Model performance may also be evaluated by creating confusion matrices of observed versus predicted classifications – confusion matrices are described later in Section 2.3.7.6.

Based on the particular combination of independent variables (factors and covariates), a model calibrated on some training data may be used for classifying unseen data. The model calculates a probability score for each of the outcomes, and class assignment is implemented by choosing that class with the highest probability. Logistic regressions that include only the original set of variables is called “main effects models” while “interaction effects models” include combinatory effects between the independent variables. Although higher flexibility is generally better, the interaction effects models may overtrain the data (or begin to memorize patterns) and therefore, might not be generalizable to unseen data. Prudent selection of independent covariates and factor levels can help in preventing overtraining (Dreiseitl and Ohno-Machado 2002). Additionally, practical experience suggests that parsimonious models are easier to interpret and explain – even with adequate numbers of samples, if the number of independent variables increases, or if the number of classes in categorical independent variables increases, the estimated parameters may be so great in number that interpreting the model becomes cognitively difficult.

2.3. Artificial Neural Network Solutions for Categorical Data Analysis

This section defines ANNs and briefly describes their evolution, along with examples of successful modern ANN applications. The section concludes with concepts and theoretical aspects that are relevant from a methodological point of view.

2.3.1. A historical perspective on Artificial Neural Networks

Fukunaga (1990) argues that human decision-making is strongly related to the recognition of patterns, and that the overall goal of pattern recognition is to clarify this process and automate the same using computers. Since many humans perform classification by pattern recognition, often better than any machine, there has been tremendous interest in understanding the process of human decision-making among computer scientists, engineers, psychologists and physiologists.

During the 1940s, McCulloch and Pitts (1943) introduced the first mathematical model of the neuron and demonstrated that networks of neurons with simple outputs could, in principle, compute any arithmetic or logical function. Hebb (1949) then proposed a learning law that rigorously described learning at the cellular level. In the 1950s, Frank Rosenblatt (1958) demonstrated the first neural network that was able to perform pattern recognition using the “perceptron network” and the associated learning rule. Widrow and Hoff (1960) demonstrated a new learning algorithm and used it to train adaptive linear neural networks, similar to the perceptron. In this model, the network processes inputs into desired output categories, calculates the error between network and desired outputs and then adjusts input weights using a gradient descent method that minimized the least mean square error [MSE]. The Widrow-Hoff learning rule is still in use today. This led to great enthusiasm (and extremely inflated expectations!) in the field of machine learning as influenced by mathematics, psychology and biology. The balloon was quickly punctured by Minsky and Papert (1969) who rigorously determined what a perceptron network was capable of learning, and demonstrated their limitations so pessimistically and effectively that it caused a research and funding drought in neural computing for several years.

With advances in computing technology and processing power, and two key conceptual breakthroughs, the field experienced a renewal during the 1980s. The first concept used statistical mechanics to explain the associative memory properties of specific recurrent networks (Hopfield 1982). The second key milestone was the development of the backpropagation algorithm to train multilayer perceptron networks (Werbos 1974; Hinton and Sejnowski 1986; Rumelhart et al. 1986; Rumelhart and McClelland 1986) by several researchers that successfully refuted earlier criticisms and resurrected the field. Over the last two decades, thousands of papers have been written with successful applications of artificial neural networks in many different fields. This is only a brief account of the fitful and dramatic progress of knowledge in neural computing, and the interested reader is referred to Anderson and Rosenfeld (1990) for an excellent review of the history, evolution and theoretical perspectives of the leading exponents of neural networks.

2.3.2. What is an Artificial Neural Network?

As you read this sentence, you are using a complex and intricate neural network comprising of over 10^{10} neurons (StatSoft 2003) with on average, about 10,000 inter-neuron connections. In general, all biological functions, including memory, are stored in neurons and in inter-neuron connections – learning is conceptualized as the generation of new connections or the modification of existing ones (Hagan et al. 1996). Following this conceptualization, Rojas (1995) defines the fundamental problem of an information processing system as the transmission of information, since data storage can be transformed into a recurrent transmission of information between two points. While the biological neural network is extremely complex in terms of structure and connectivity (and therefore are extremely powerful processing units), artificial neurons are simple abstractions of biological neurons and arranged in some interconnected sequence as an

artificial neural network. Such artificial neural networks may be trained to perform some specific and useful functions.

The simplest conceptual definition of an artificial neural network is a model whose output is some linear or non-linear combination of the inputs. These models are based on numeric inputs and outputs, which may therefore require some preprocessing of input data. A biological neuron has dendrites that receive information at the contact points (synapses) between neurons, a cell body that produces energy consumed by the other components of the cell, and an axon to transmit an output signal. This structure is abstracted and represented in Figure 2.1 as an artificial neuron, along an input channel (analogous to a Dendron), a weight (corresponding to a synapse), a summation and transfer function (the resource for firing a signal, from the cell body) and an output channel (the axon).

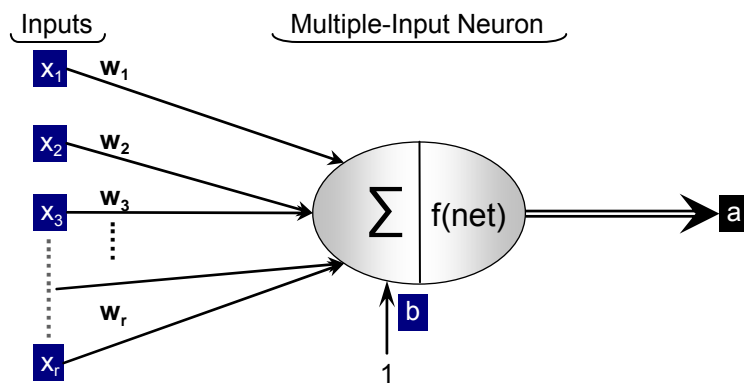


Figure 2.1 -- An abstract artificial neuron with multiple inputs

Each artificial neuron may be conceptualized as a simple processing element [PE] carrying unidirectional communication channels, operating only on the local data that they receive through their connections. Thus, each input received $[x_1, x_2, x_3, \dots, x_r]$ is weighted by the corresponding weight elements $[w_1, w_2, w_3, \dots, w_r]$ and along with the bias, transmitted to the summation operator. The summation operator adds the bias and

the products of the inputs and weights and transmits the result 'n', to the transfer function. Thus, $n = \sum_{i=1}^r w_i * x_i + b$. The transfer function **f(net)** processes the result of the summation operator and, depending on whether the computational result is above a threshold, fires the output signal.

Typically, one neuron with multiple inputs may not be sufficient to solve the problem. Several neurons, operating in parallel, form a layer, whose PEs are connected locally to all the inputs. Figure 2.2 shows such a "single layer feed-forward" artificial neural network architecture (Hagan et al. 1996). Note that each of the four input attributes for the sample object is weighted and connected to each of the processing elements and the weights form a matrix, whose rows correspond to the number of PEs and columns to the number of input attributes.

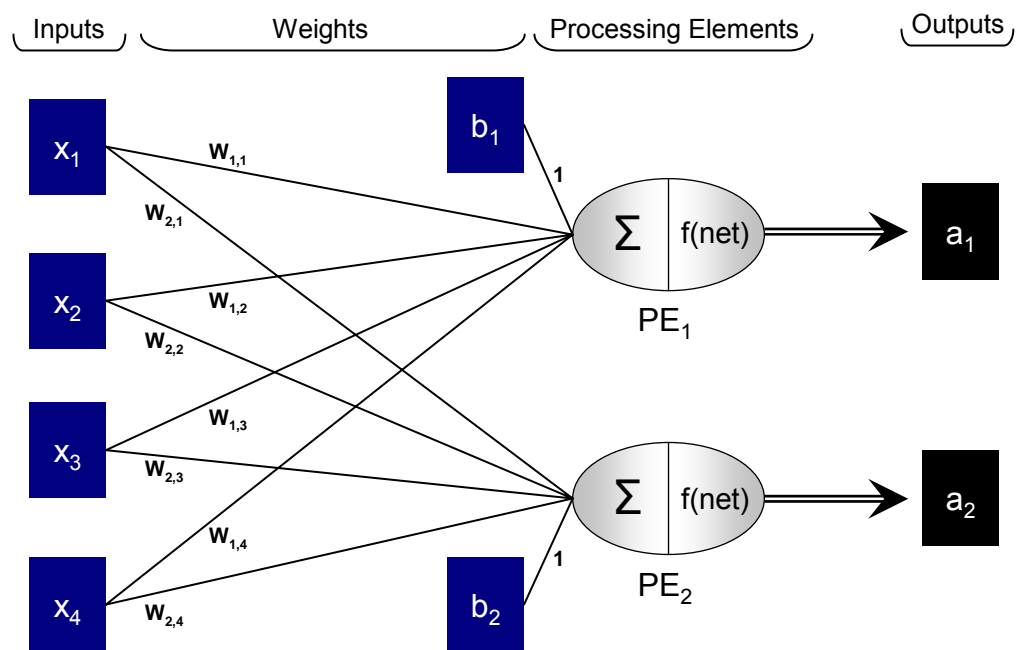


Figure 2.2 -- Single layer feed-forward ANN topology – the Perceptron

Thus, summations for PE₁, $n_1 = \sum_{i=1}^4 w_{1,i} * x_i + b_1$ and for PE₂, $n_2 = \sum_{i=1}^4 w_{2,i} * x_i + b_2$

Each PE's transfer function **f(net)** processes the result of the summation operator and, depending on whether the computational result is above a threshold, fires an output signal. Extending the single layer to several other "hidden" layers, Figure 2.3 shows a multiple layer feed-forward network architecture. Here, each of the three layers has its own weight matrix and bias vector. Layers may have different numbers of processing elements. The outputs of each layer serve as inputs for the succeeding layer. A layer whose output is the network output is called the output layer.

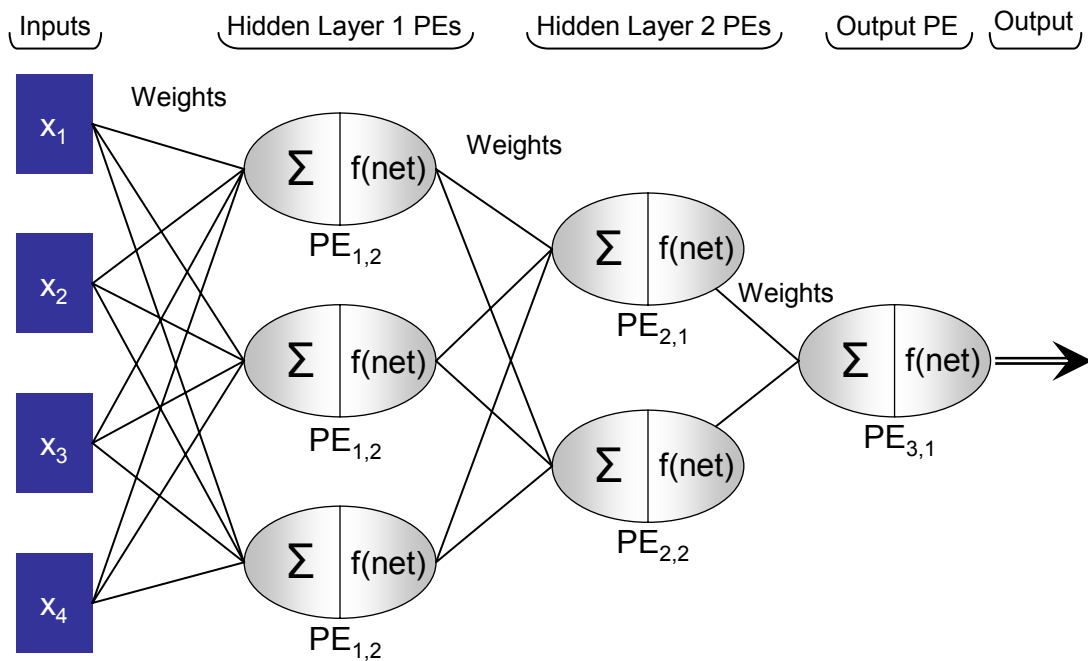


Figure 2.3 -- Multiple layer feed-forward ANN topology

In Figure 2.3, there are two hidden layers and one output layer. Hidden layer 1 has 3 PEs and a weight matrix of order [3 x 4], corresponding to the 3 PEs and 4 input

attributes. It has 3 outputs that serve as inputs to Hidden layer 2 that has 2 PEs and a weight matrix of order $[2 \times 3]$, and so on.

Thus, a partial definition of ANNs would be *networks of many simple processors connected by unidirectional communication channels that carry numeric data -- these simple processing units operate in parallel and act only on the local data inputs they receive along their communication channels.*

Human pattern recognition however, is behavior learnt through training, or detecting structure through example (Ripley 1996). In many cases, while we recognize patterns, we may be unable to describe the explicit rules by which we make judgments (and this is often the case with ANNs also!). A common mode of learning between humans and machines involves the presentations of input features with known class examples. With the addition of this additional mechanism, our working definition would be complete – the learning rule. The learning rule essentially adjusts the weights of the various connections by comparing the network output to the desired pattern (or known class example). Thus, the ANN learns from examples, by calculating the error and adjusting the connection weights so that this error is minimized. Figure 2.4 shows the general process of training the ANN, a schematic of the process of weights adjustment through error minimization. Once the weights have been adjusted so that the error is at a minimum, the weights are frozen, or the ANN has been “trained.” Then, new data may be presented to the ANN, and the network computes an output based on the optimum weights determined during training.

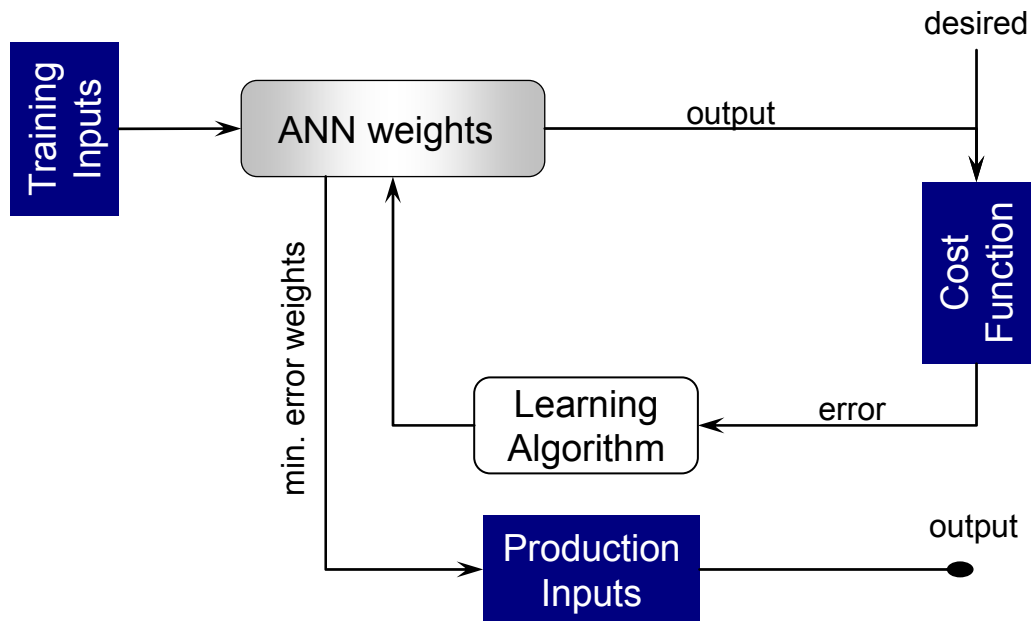


Figure 2.4 -- Schematic of ANN training and use

Thus, ANNs are “distributed, adaptive, generally non-linear learning machines” (Principe et al. 2000) built from many PEs, whose interconnectivity defines the topology of the network. Signals flowing across these connections are scaled by adjustable parameters or weights, one for every connection. In a mathematical sense, ANNs build discriminant functions from their PEs, with the topology determining their number and shape. Since the discriminant functions change with the topological specifications, ANNs are regarded as semi-parametric classifiers.

2.3.3. Transfer Functions

The transfer function of the PE is an important concept – the output of the summation operator is processed by the transfer function for conversion into some real output of the PE. The transfer function is therefore an algorithm that transforms the

output of the summation into a zero, or one, or negative one, or some other number (Haykin 1994; Hagan et al. 1996). The transfer function may also scale the output. There are several transfer functions commonly supported by most neural software applications as seen in Figure 2.5 below. The combination of layered topology and transfer functions in a neural network is what enables non-linear approximation.

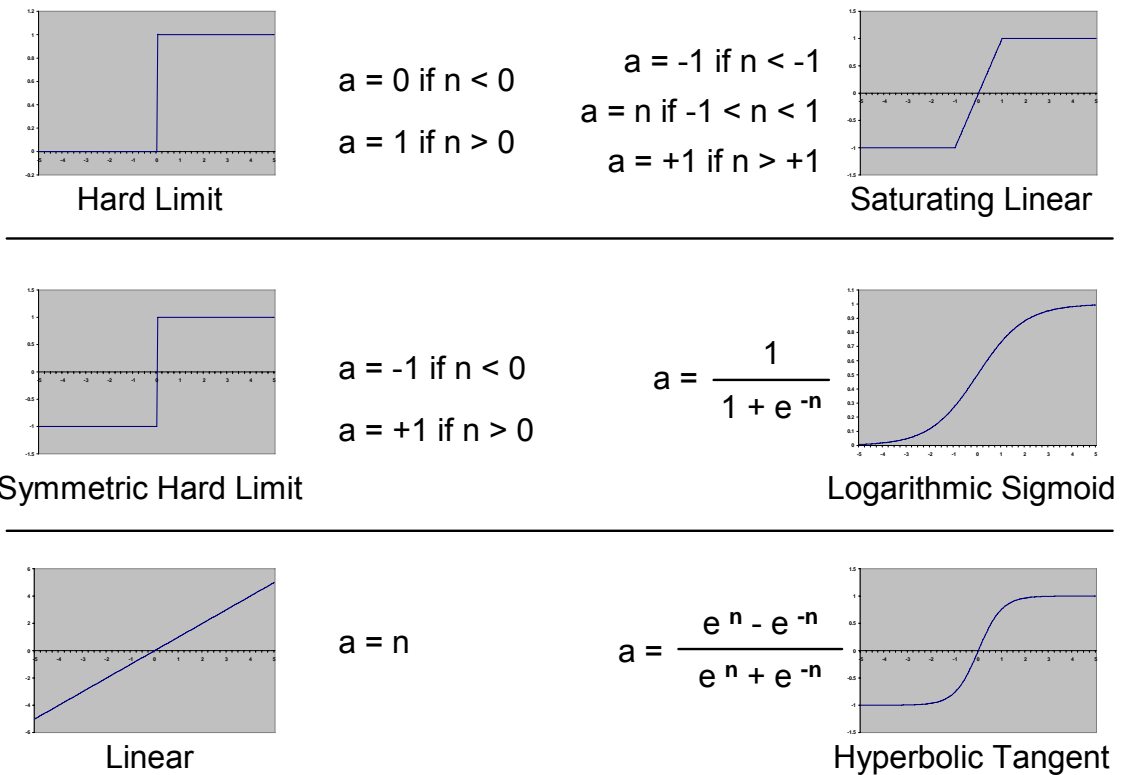


Figure 2.5 -- Some example transfer functions

2.3.4. Applications of Artificial Neural Networks

ANNs have been successfully applied in a variety of contexts and applications in the last decade, and the applicability has been dramatically increasing (StatSoft 2003; California Scientific 2007; Makhfi 2007). There are indeed a plethora of examples where ANNs have been used and this section highlights only a few. ANNs are used for the following topical areas:

2.3.4.1. Function approximation

These models are used for modeling processes, process control, data modeling, machine diagnostics, regression, etc. ANNs have been used in engine management to analyze signals from engine sensors for controlling functional parameters in order to achieve specific goals such as minimizing fuel consumption. In real estate appraisal, ANNs use sales and multiple listing data to appraise properties with high accuracy in Pennsylvania and New York. ANNs are increasingly being used for weather forecasting -- the Fort Worth National Weather Service uses neural network models to analyze weather data and predict rainfall with over 85% accuracy.

2.3.4.2. Time series analysis and prediction

ANNs are being used by many technical analysts to make predictions about stock prices based upon a large number of factors such as past performance of other stocks and various other economic indicators. Over 80% of Fortune 500 companies currently have and actively use ANNs (Makhfi 2007). LBS Capital Management, Inc uses ANNs to predict the S&P 500 one day ahead and one week ahead with higher reliability than other existing methods (California Scientific 2007). Several utility companies, including Northern Natural Gas, predict gas price fluctuation with over 95% accuracy. Banks apply ANNs to predict bankruptcy rates regularly.

2.3.4.3. Classification

ANNs have been increasingly used for pattern recognition and classification, and by far seems to dominate the application areas. In medicine, ANNs have been used to classify malignant cancer cells and predict the functional recovery time for hospital patients, which insurance companies are observing with great interest. In finance, banks and lending institutions use ANNs to establish the credit-worthiness of applicants by

analyzing attributes such as age, vocation, education, past financial history, etc. to classify credit risk (StatSoft 2003). ANNs are also increasingly being used to analyze performance of machinery to differentiate between false alarms and real problems, and in particular, to predict the imminent failure of machines or machine parts. Intel uses ANNs in computer chip manufacturing and quality control to identify patterns in chip failures (California Scientific 2007). ANNs are actively being used in voice and word recognition for phone systems and voice to digital conversion. Several companies use ANNs for optical character recognition for identification of text, characters, symbol and map features and for conversion to digital format.

Of particular interest to us, in the field of GIS, ANNs have performed better than traditional methods for classification of remote-sensed data from images by integrating texture with color values (Bischof et al. 1992). Other researchers have demonstrated the viability of ANNs in all stages of a GIS system, ranging from data preparation to analysis and modeling (Kavzoglu et al. 2000; Kavzoglu and Mather 2000). Currently, we are also exploring the potential to use ANNs for automated shape recognition of building footprints for earthquake risk inventory, by analyzing vector digital GIS building polygons.

2.3.4.4. Data mining

ANNs are widely used in identifying patterns from raw data, data extraction and visualization, general data warehousing and mining applications. ANNs have proven to save resources and time in emergency room testing logistics by predicting test types based on symptoms and demographic information. Pharmaceutical companies use ANNs to analyze sales of their products by mining ancillary data such as sales frequency, demographics, transportation logistics, pharmacy locations, etc.

2.3.5. Why use Artificial Neural Networks?

ANNs have often been criticized since they lack good parametric measures of performance. ANNs, particularly in classification exercises, are evaluated primarily based on their performance. Further, in most instances, their inner workings are a mystery to even experienced users -- they have been severely criticized for their black-box approach, and their lack of explicit explanatory power (Anderson and Rosenfeld 1990; Anderson and McNeil 1992). Nevertheless, ANNs are gaining in popularity and, as shown in the previous section, are used in an extraordinary variety of disciplines. ANNs are being used wherever there are needs for classification, prediction, signal identification or control. There are several reasons for the current increase in popularity.

First, ANNs clearly provide sophisticated and cutting-edge techniques capable of modeling very complex functions. Traditional modeling relies heavily on linear techniques, because several optimization routines exist for linear solutions. However, linear techniques are not universally applicable, and where applied in non-linear situations, modeling results are often poor. Speech recognition is a typical area where traditional linear solutions offer very poor performance. Traditional non-linear solutions further require almost prohibitive amounts of data, while ANNs, with their iterative techniques, control the dimensionality problem to some extent (Makhfi 2007). Secondly, ANNs learn by example, where the network trains on representative, known examples using learning algorithms to identify input data patterns. Most practitioners readily admit that ANN users require some heuristic knowledge for variable selection, data preparation, network topological design and interpretation of diagnostics and results (Nilsson 1996; Patterson 1996; Principe et al. 2000). Nevertheless, the level of user knowledge is substantially lower than traditional non-linear statistical techniques, particularly when performance is the key (Anderson and McNeil 1992).

Third, ANNs are intuitively appealing, since they are based on some level of similarity of biological systems. Fourth, since they are semi-parametric, they can rely on learning complex patterns in the data directly, without user intervention. Indeed, in many situations, the dimensionality of the problem may overwhelm human analysis (StatSoft 2003). Thus, ANNs may be self-organizing and can adapt themselves continuously to newer data. Fifth, ANNs may be specified to include some level of fault tolerance and still perform well – faulty or incomplete input data severely inhibits traditional statistical approaches (Rojas 1995). Sixth, considering the advances in computing technology, particularly multiple processors and thread-based routines, ANNs can be designed for optimization by parallel processing of inputs. This would greatly enhance speed of training and prediction of response (Rumelhart et al. 1986; Rumelhart and McClelland 1986). Finally, the results generated by a neural network may be generalized and applied to new or unseen data with relatively high performance.

ANNs are not suited for all applications, particularly in well-specified problems. For instance, inventory accounting and data maintenance are applications where traditional computing approaches would be better. Thus, ANNs offer a new approach to solving problems and identifying patterns, by providing tools that learn by themselves, without the necessity of experts or specialized computer programming.

2.3.6. Artificial Neural Network topologies for classification

There are several theoretical and practical aspects to the design and training of artificial neural networks that directly influence classification performance. These include topologies or neural network specifications for classification, efficiency and control of learning, error criterion, control of training for validity and generalized classification performance.

2.3.6.1. Neural Computing for Classification

ANNs are used both for function approximation (as in regression-type fitting hyperplanes to input points) and for classification. In general, ANNs that approximate functions may not be used to separate items into classes. The function approximation problem is aimed at capturing the relationship between the input points and the desired response. In classification, we acknowledge that different mechanisms generate input data, and the goal is to separate the input space into one of several classes that are arbitrarily labeled. Since participation in a class implies non-participation in other classes, a good classifier is characterized by a non-linear separating mechanism, such as an all-or-nothing switch (Principe et al. 2000).

Any class assignment is not error-free. In creating a threshold to separate two classes, the tails of the likelihoods of the two classes overlap, creating the error region. Calculating the Bayesian threshold (that maximizes the unknown, but computable a posteriori probability) minimizes the error probability region (Fukunaga 1990). While separability is a function of the mean and variance of each class, the computation of the posterior probability is not trivial in higher dimensional spaces. Both statistical classification and ANNs use discriminant functions to separate inputs among classes (Michie et al. 1994).

2.3.6.2. Discriminant Functions

Consider a case where we have “k” samples with “d” input attributes for each sample. Each sample may then be viewed as a point in d-dimensional space, or as a vector \mathbf{x}_k with “d” components. By Bayes’ rule, class assignment is based on the comparison of likelihoods scaled by the corresponding a priori probability (generally a

simple proportion of sample cases belonging to a class). Any sample \mathbf{x}_k will be assigned to a class “i” if

$$g_i(\mathbf{x}_k) > g_j(\mathbf{x}_k) \text{ for all } j \neq i$$

Each scaled likelihood is then regarded as a discriminant function $g(\mathbf{x})$ that assigns a score to every sample in input space. Each class has its own scoring function that produces higher values for samples belonging to it. Discriminant functions intersect in the d-dimensional space, creating “decision surfaces” – in other words, decision surfaces partition the input space into volumes where one of the discriminants has a higher value than all the others. Thus, ANNs used for classification attempt to produce mechanisms that compare discriminant functions and assign the sample to the class that provides the largest discriminant value for the sample. Figure 2.6 shows an ANN schematic for a general classifier for “p” classes.

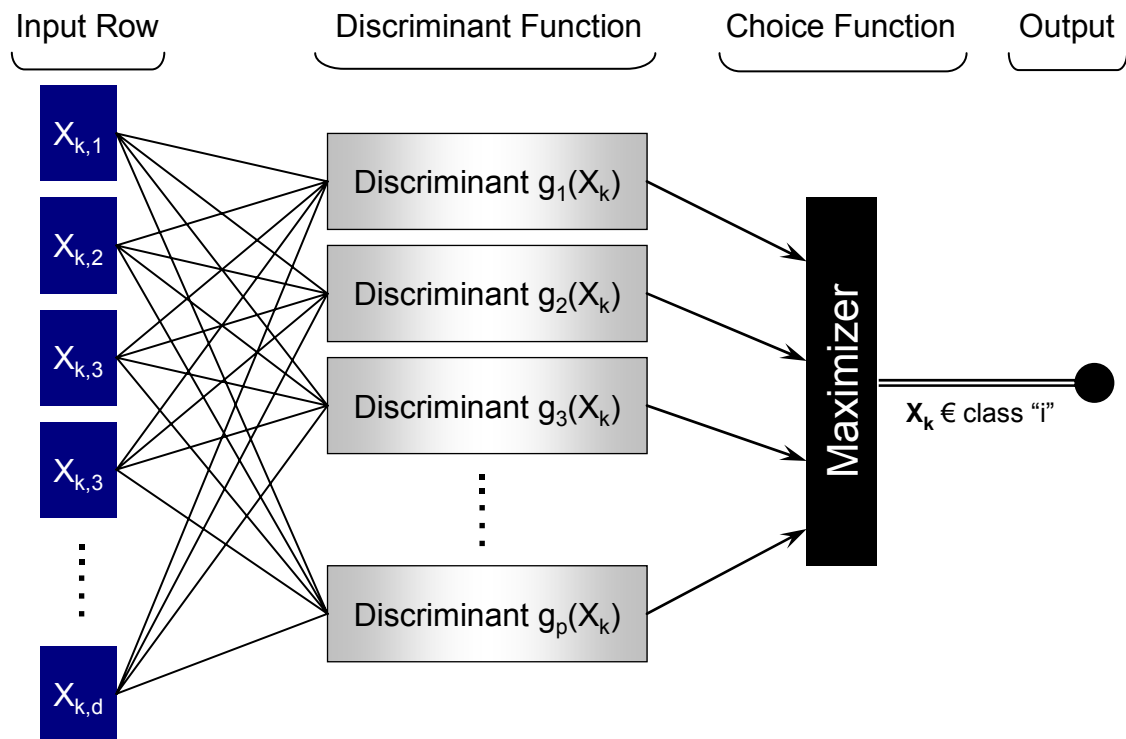


Figure 2.6 -- General schematic for classifying samples into “p” classes

When discriminant functions have a well-defined functional form in terms of parameters (for instance, mean and variance for a Gaussian), the resulting classifier is termed a parametric classifier. It is possible to train ANNs on non-parametric classifiers that do not assume any underlying functional form, but estimate the discriminant solely on the data (Fukunaga 1990). However, such classifiers require a large number of samples for acceptable performance. For a classifier built parametrically on the Gaussian distribution, Fukunaga (ibid) showed that the optimal classifier is always a quadratic. Parametric discriminant functions are also sensitive to the number of samples – even though they may retain their overall shape, classification performance is lower with fewer samples, particularly in higher dimension input space (Principe et al. 2000). It is also important to note that linear discriminant classifiers are less powerful than quadratic discriminants because the former rely primarily on differences in means.

2.3.7. Conceptual issues in designing and training Artificial Neural Networks

ANNs adapt connection weights iteratively by comparing network outputs with known examples. The comparison between outputs and desired results produces an error measure – the goal of the network is to adjust weights so that the error measure is minimized. This process is called “learning (Haykin 1994; Rojas 1995; Patterson 1996). If learning is inadequate, the weights will not be optimal and performance will be affected. While systematic procedures exist to search the performance surface, the search process has to be controlled heuristically. Note that the search process will not yield the best results if the amount of data is inadequate or if the sample data is not representative of the true process being modeled. The user directly influences learning by selecting the search techniques, learning algorithms, specifying the learning step sizes, the size of the topology and the number of learning cycles (Carling 1992; Hagan et al. 1996).

2.3.7.1. Error minimization search procedures

There are several measures of error including absolute cost, quadratic cost, polynomial error functions. The cost criterion is generally a positive quantity that is sensitive to the network output, and should be chosen such that it approaches zero as the network outputs approach the desired response. The most commonly used error cost function is the mean square error often termed “J.” In the one dimensional case, since the output of the network is a function of the connection weights, the mean square error is quadratic on the weights and is a parabola facing upwards, as seen in Figure 2.7.

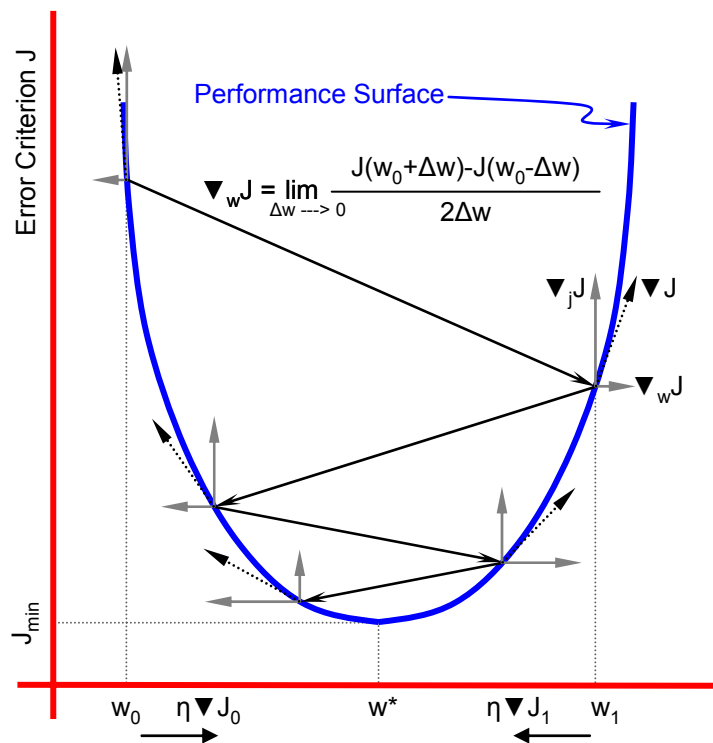


Figure 2.7 -- Weight optimization by minimizing the performance criterion

This error function graph shown in Figure 2.7, is called the performance surface (adapted from pp. 24, Principe et al. 2000). For more complex problems, particularly

classification, the performance surface may be more complex, with several local minima. However, we can use the one dimension case to illustrate error analysis concepts.

The gradient of the performance surface at any point is a vector of “w” dimensions in the steepest upward direction at that point, with larger magnitudes for steeper slopes. Minimizing the performance criterion is usually performed by methods based on gradient computation. Since the overall objective is to minimize the error criterion, one can search along the performance in the opposite direction of the gradient – this is the popular gradient descent method. As Figure 2.7 indicates, the steps involved in the gradient descent method include the following:

- initialize the search at w_0 , an arbitrary initial weight
- compute the gradient of the performance surface at w_0
- modify the initial weight proportionately to the negative of the computed gradient at w_0 , changing the new weight to w_1 , and repeat the steps

Thus, $w(k+1) = w(k) - \eta \nabla J(k)$, where η is a small constant, called “step size” or “learning rate” (Rojas 1995; Patterson 1996; Principe et al. 2000), which ensures that the new operating point is identified not too far along the performance surface and $\nabla J(k)$ is the gradient of the performance surface at the k^{th} iteration. Where $w > w^*$, w is decreased to find the new operating point and vice versa. If η is small, the optimum weight w^* will be found (Rojas 1995; Patterson 1996; Principe et al. 2000). Widrow and others (Widrow and Hoff 1960; Widrow and Sterns 1985) proposed a gradient estimate method termed the least mean square algorithm that uses the error from a single sample rather than the previous methods that had the larger overhead of summing the error for each point in the data set. This algorithm suggests that the instantaneous estimate of

the gradient at iteration k is simply the product of the current input and the current error. The estimate is noisy, but gets filtered out over several iterations.

Search procedures may be analytic or iterative. Analytic solutions require all the data beforehand in order to compute optimum values. However, in a practical sense, if the data is not representative, or if autocorrelation matrices are ill-conditioned, analytic solutions may not be accurate. Further, analytic solutions require a great deal of time. ANNs use iterative searches because solutions may need to be implemented on a sample-by-sample basis. Additionally, very efficient gradient search algorithms have been created, and optimization is achieved in linear time, faster than analytic approaches. Finally, iterative searches may be extended to non-linear systems with several minima, for which analytic solutions may not exist.

2.3.7.2. Learning rate

As seen in Figure 2.7 and in the previous section, the rate of error decrease is proportional to the step size or learning rate. Larger step sizes will need fewer iterations to reach the vicinity of the minimum. However, too large a step size will create a divergent iterative process. If the step size is small, learning takes a long time. Even if the step size is constant, as advocated by Haykin (1994), the adjustments to the weights reduce in magnitude as the search progresses towards the minimum, because the slope of the quadratic performance surface correspondingly decreases. In some cases, particularly close to the minimum, the iterative process begins to wander in the vicinity of the minimum without ever reaching the minimum, a phenomenon termed “rattling” (Principe et al. 2000). The iterations may have to be stopped externally, leading to a sub-optimal solution. Again, a smaller step size can avoid this misadjustment at the cost of longer learning times. Several neural software applications avoid the rattling problem by scheduling a large step size at the beginning of the training to move quickly to the

neighborhood of the minimum and then decreasing it near the performance surface minimum, using linear, geometric or logarithmic functions.

2.3.7.3. Learning algorithms

We had previously seen the least mean square algorithm and the weight modification routine, where the instantaneous estimate of the gradient was the simple product of the current error at that weight and the current input value for that iteration.

Thus,

$$\nabla J_i = \varepsilon_i * x_i$$

The same algorithm may be reached by the “delta rule” using partial differentials. Since the error cost J, was defined as the Mean Squared Error between the desired value and network output

$$J = \frac{1}{2} \sum_i (d_i - y_i)^2 = \sum_i J_i, \text{ and}$$

$y_i = w * x_i$, since the output is a function of the weight and the input value

$$\frac{\partial J_i}{\partial w} = \frac{\partial J}{\partial y_i} * \frac{\partial y_i}{\partial w} = -(d_i - y_i) * x_i = -\varepsilon_i * x_i$$

Extending the least mean square concept to the perceptron or MLP, whose sigmoid threshold function defines a non-linear system is relatively straightforward. First, the partial derivative of the output with respect to the transfer function is computed. Then, compute the partial derivative of the transfer function with respect to the weights. The product of these two terms determines the sensitivity of the output to the weights. The same rule is extended to hidden layers, because the chain rule may be applied as many times as necessary (Carling 1992; Rojas 1995; Patterson 1996).

In the context of learning, all the weights are adjusted in order to minimize the error, but using the generalized delta rule, the adjustments are distributed in proportion to the sensitivity of the output to the weight – this rule is also termed backpropagation (Werbos 1974; Hinton and Sejnowski 1986). Note that in the case of non-linear systems, the performance surface becomes more complex, characterized by several local minima and a global minimum or by flat regions where the gradient is zero. The noisy local estimates described earlier become useful, because the natural perturbation increases the chances of escaping from flat spots or local minima (Principe et al. 2000).

In terms of curvature, for complex, non-linear performance surfaces, the local and global minima may be identical, causing the search to stall. Alternately, since the weights change very little if the performance surface is flat, users may confuse this with the end of the training. Momentum learning is one such robust method where the magnitude of a previous increment is used to speed up and stabilize the convergence routine, thereby preventing the search from getting trapped in local valleys (Haykin 1994; Patterson 1996). Another method commonly used is the delta-bar-delta rule, which essentially looks at the magnitude of the previous weight change and adapts the learning rate continuously during training. More stable non-linear variations of this method include Fahlman's quickprop and Almeida's adaptive step methods (Fahlman 1989; Silva and Almeida 1990).

Several other methods such as the conjugate-gradient, pseudo-Newton, Levenberg-Marquardt methods have been applied in neural network applications and the user is directed to Fletcher (1987) or Luenberger (1984) for a review of these techniques.

2.3.7.4. Processing elements in the hidden layer

Setting the number of hidden PEs is an important issue in specifying the network topology. If the number of PEs is more than necessary, training times are longer, and while correct classifications in the training set increase, the solution does not perform well with unseen data. In other words, the network memorizes the training data patterns and does not generalize well to unseen data. If there are too few PEs, the network will randomly change weights in order to reduce the MSE. The classifier will attempt to place discriminant functions to correctly classify the majority of the samples first, before proceeding to sparse regions. Performance will be better than if too many PEs were specified, but the solutions weights are sub-optimal and not as good as a network with the correct number of hidden PEs. Again, this is a heuristic determination on the part of the user.

2.3.7.5. Stop criteria

Training may be stopped based on a specific number of iterations, or based on the output mean squared error, or based on generalization. Stopping training based on the number of iterations offers no guarantee that the classifier has generated optimal weights. In terms of MSE, one might choose an acceptable error level and stop training when the MSE threshold is reached. Alternately, training may be stopped when the incremental change in MSE falls below a specified threshold. As mentioned before, if the classifier is training in flat regions of the performance surface, training may stop prematurely.

At this stage, we should examine the concept of generalization – how well does the system perform on data samples that it has not been trained on? Researchers have demonstrated that after a critical point, the system will continue to do better in the

training set, but deteriorate in the testing data set. In other words, the system begins to memorize the data patterns in the training data set (Rojas 1995; Patterson 1996; Vapnik 1999). Given the current training data set and the network architecture, an accepted method of maximum generalization potential is to stop training at the point of minimum error in the testing or cross-validation dataset. See Figure 2.8 for a schematic of the cross-validation criterion for stopping training.

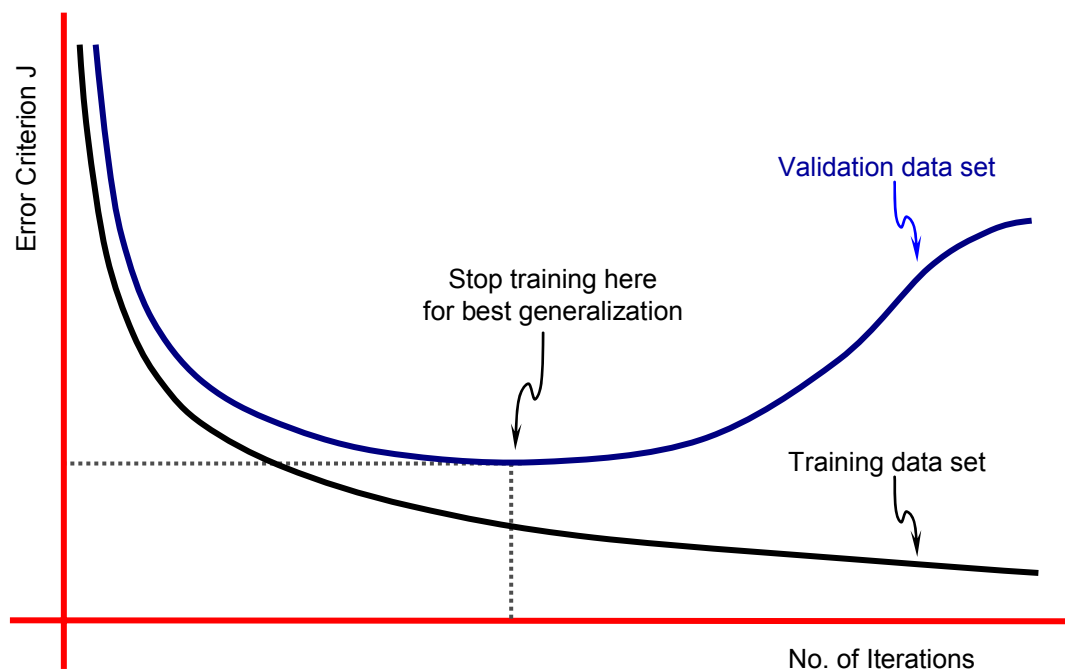


Figure 2.8 -- Stopping criterion using the cross-validation data set

The training data set should be split in order to reserve about 10% for the cross-validation data set. After every few iterations, the current weights at that iteration are tested against the cross-validation data set. Training should ideally stop just when the error criterion for the cross-validation data set begins to rise.

2.3.7.6. Performance Measures

In real world applications, models usually have some measure of performance. ANNs have been criticized for their lack of explicit explanatory and parametric performance measures. While the MSE is an indirect measure in function approximation, there is no precise relationship between classification performance and MSE. Typically, the performance of a classifier is measured in terms of true and false classifications, represented by a confusion matrix (Rojas 1995; Principe et al. 2000). The confusion matrix is a table that compares the classifier output in columns with the known class in rows. Thus, perfect classification would result in a confusion matrix whose diagonal elements are populated and all others are zero. Overall classification error is the ratio of the sum of the off-diagonal values and the total number of samples. The confusion matrix also allows the analysis of where classification had difficulties, since some classes produce greater errors than others. Table 2.1 shows an example confusion matrix for 3 classes. Elements along the diagonal (intersections between predicted and desired, highlighted in blue) are correct classifications. Elements along the rows show the number of correct classifications and the number of each row class misclassified as one of the other classes. Table 2.2 deconstructs the raw confusion matrix to show only the correct predictions. Note that 82%, 64% and 74% of Species A, Species B and Species C respectively have been correctly classified, with an overall model accuracy of 73%. Table 2.3 deconstructs the table to show the number of correct predictions in each class, and the number of misclassifications for each class relative to the other two classes. Of the 49 exemplars predicted as Species A, 41 (or 84%) were correctly classified, 6 (or 12%) were misclassified as Species B and 2 (or 4%) were misclassified as Species C. Similarly, from the second row, 12% of Species B was misclassified as Species A and 22% misclassified as Species C. Finally, 6% of Species

C was misclassified as Species A, and 23% were misclassified as Species B. From the decomposition, it is obvious that the model performs best for predicting Species A, and that it has some difficulty discriminating between Species B and C, based on the particular input combinations.

Table 2.1 -- Example raw confusion matrix for three classes (counts)

Desired	Class Name	Predicted			Total Predicted
		Species A	Species B	Species C	
	Species A	41	6	2	49
	Species B	6	32	11	49
	Species C	3	12	37	52
	<i>Number of Samples</i>	50	50	50	150

Table 2.2 -- Example confusion matrix for three classes (percent accurate)

Desired	Class Name	Predicted			Overall Accuracy
		Species A	Species B	Species C	
	Species A	82%	-	-	-
	Species B	-	64%	-	-
	Species C	-	-	74%	-
	<i>Percent of Samples</i>	100%	100%	100%	73%

Table 2.3 -- Example confusion matrix for three classes (percent misclassified)

Desired	Class Name	Predicted			Total Predicted
		Species A	Species B	Species C	
	Species A	84%	12%	4%	100%
	Species B	12%	65%	22%	100%
	Species C	6%	23%	71%	100%

2.4. Shape Recognition Background, Techniques and Applications

Apart from soil, site, intensity, structure type and building capacity, the behavior of the building under earthquake stresses is also influenced by its shape (Arnold and Reitherman 1982). When earth-shaking motions are transferred to the building, additional, torsion stresses are created when the stiffness center of the building is

displaced from the center of gravity. In simple, symmetrical buildings, the centers of stiffness and gravity tend to be coincident, particularly if the massing is uniform – for asymmetrical or irregular buildings, the particular shape configuration and the building mass distribution determine the different locations of the centers of stiffness and gravity, and therefore the torsional forces (Murty 2002 a).

Thus, irregular buildings exhibit inappropriate dynamic behavior when subject to horizontal earthquake stresses. From a building occupancy perspective, irregular shapes for buildings provide convenient solutions for environmental and human design considerations. Concomitantly, from a structural point-of-view, these irregular structures are less desirable than simple, regular and symmetric structures because the former require significant engineering effort to reach an acceptable level of seismic performance (Lopez and Raven 1999). While the behavior of the building under earthquake stresses is dependent on its overall three-dimensional (3D) configuration (Murty 2002 b), the scope of this research limits itself to the identification of the two-dimensional (2D) configuration in plan for different buildings. Thus, 2D building shape types in the research include square, rectangle, L-, C-, T-, H-, Z-, octagonal, circular, cruciform and irregular.

While most jurisdictions, at least in the United States have, or are in the process of developing cadastral and planimetric databases, building shape information is usually not captured. In addition, the process of building footprint capture mainly relies on individuals “drafting” or “digitizing” the building from aerial photographs and is extremely inconsistent at best.

Further, cities and regions have several thousand structures, and it is cost-prohibitive to identify and code each building’s shape on a per-building basis. Several researchers have been working on the problem of automated building extraction from

aerial photographs (Lee et al. 2003; Wei et al. 2004; Jin and Davis 2005; Sohn et al. 2005). Another component of this MAEC project builds on Sahar and Krupnik's (1999) work to automate the detection and extraction of buildings from aerial images and forms the subject of Liora Sahar's ongoing Ph.D. dissertation at the Georgia Institute of Technology. This research assumes that building outlines have been captured either by automatic extraction from aerial photographs or other remotely sensed sources as raster footprints or digitized into vector format polygons. Accordingly, this chapter attempts to develop an automated process to identify the footprint configuration of all such presented buildings.

In the GIS field, shape analysis is somewhat limited, and more often than not, restricted to generalization and simplification methodologies. I have not come across any application that analyzes shapes of buildings with the emphasis on automated database development. However, considerable research has been conducted on image recognition and classification in Geosciences and Remote Sensing (refer to journals from Institute of Electrical and Electronics Engineers 2008). Much of the other material related to shape recognition and pattern classification comes from pattern recognition, image processing, medical imaging, robotics and artificial intelligence traditions.

There are several frameworks for classifying shape analysis approaches. Veltkamp and Hagedoorn (1999) classify image comparison methods very broadly as color and texture-based or shape geometry-based. Ashbrook and Thacker (1998) use methods for shape representation as a classification framework for visual recognition – thus they classify shape analysis research by invariant representations, template matching, skeletonization, moment invariants, log-polar mapping, geometric feature descriptors, boundary profiles, Fourier transformations, dynamic shape modeling, 2D projection invariants and pairwise correspondence. Chang et al (1991) separate shape

representation aspects from shape recognition, and organize shape representation under Fourier descriptors, moment invariants, autoregressive modeling, polar mapping and syntactic approaches, while shape recognition is achieved through statistical or syntactic methods. Loncaric (1998) elaborately classifies shape analysis by several frameworks including a) boundary or global, b) numeric or non-numeric and c) information preserving or information non-preserving methods and presents a comprehensive survey of several published papers under these categories. In all these surveys however, there is considerable overlap between methods of shape description and analytical approaches. An exhaustive review of several literature surveys in shape analysis reveals that most publications can fit in two or more classification frameworks.

2.4.1. Definition of a Shape

From both scientific and technological perspectives, the human sense of vision will provide the basis for considerable research effort in the future. Over 50% of our daily activities are involved in the processing and analysis of visual input, with over 30 distinct areas of the brain participating (Ramachandran and Blakeslee 1998). Thus, our cognitive abilities and processes of learning are extraordinarily related with vision, and vision is more than an identification or navigational system. With the sharp increase in the development and use of digital systems, the potential for computers vision systems to substitute for human ones has tremendously increased (Fabel 1997). Many tasks that require aspects of human vision will be performed by computers for which designing and deploying effective computer vision systems becomes essential.

Generally, humans perceive a shape through its properties – similarly, even for artificial visual systems, shapes tend to be defined in the context of its attributes. In the current literature for automated shape recognition and in the field of shape analysis, definitions of shapes are predominantly based on those properties of an object that are

invariant to geometric transformations such as translation, scale or rotation (Bookstein 1991; Dryden and Mardia 1993; Small 1996; Dryden and Mardia 1998). Thus Dryden and Mardia (1993) borrow Kendall's (1984) definition of shape as "all the geometrical information that remains when location, scale and rotational effects are removed." Other researchers more specifically use the invariant geometrical properties of the relative distances among a set of static spatial features of an object to define shape (Ansari and Delp 1990). Most of these definitions of shape deal primarily with specific attributes of human perception without specifying the underlying shape originator – in other words, humans tend to perceive shapes informally, in terms of similarities and metaphors. In addition, shapes may be skewed or deformed, or occluded or noisy, and yet be recognized by humans. Thus, any definition of shape should address the attributes of the represented object and its equivalence under a set of transformations – a shape could therefore be defined as "a single visual entity comprising of any connected set of points" (Costa and Cesar 2001a). In the context of this research, especially in a spatial vector format, we modify the definition of the building footprint shape as a polygonal area distinguished from the surrounding area by a connected and closed set of line segments. Note that the Costa and Cesar definition of shape is a subset of our definition, since a set of connected line segments may also be represented as a connected set of points.

2.4.2. The Process of Shape Analysis

The general steps in the process of shape analysis include shape acquisition, shape representation, feature extraction, and shape classification (Loncaric 1998; Costa and Cesar 2001a). Numerous variations of this approach are represented in the literature, but agree generically on characterization of shapes through their attributes and then analyzing them for retrieval, comparison or recognition (Grenander 1996; Dryden and Mardia 1998; Belongie et al. 2002; Adamek and O'Connor 2003; Acharya

and Ray 2005; Golland et al. 2005; Salih et al. 2006; Pratt 2007). Figure 2.9 details such a schematic approach to building footprint shape analysis.

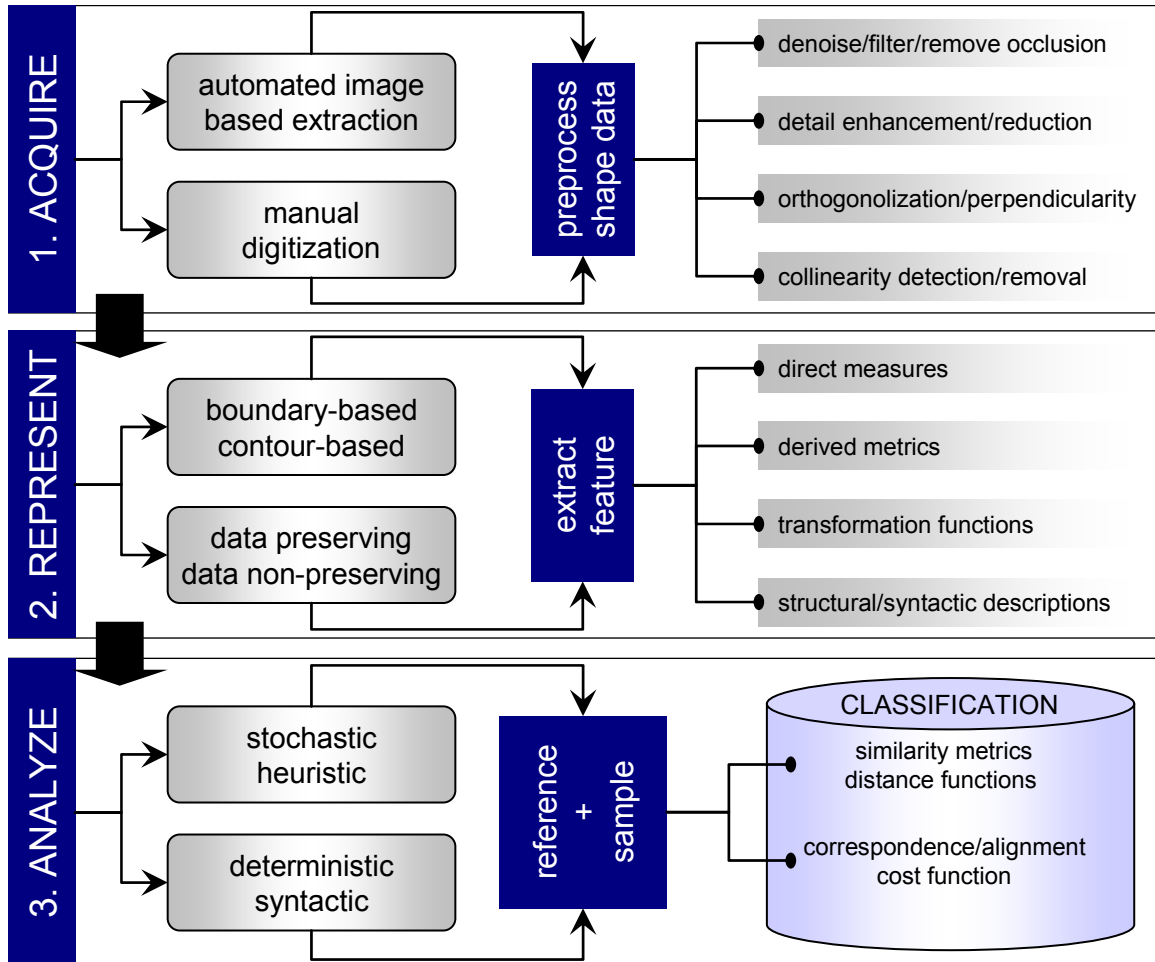


Figure 2.9 -- General process stages in building footprint shape analysis

2.4.2.1. Shape Acquisition

The process begins with identifying and separating the shape of interest from its surrounding and acquiring its digital representation. In the context of this research, the process of building footprint extraction relies typically on human drafting or digitizing from aerial photographs. For the purposes of this research, we assume that there exists a digital spatial dataset of polygonal building footprints in GIS vector format. As

mentioned before, such derived building footprint spatial databases are extremely inconsistent. The digital representations of the buildings may then have to be pre-processed in order to remove noise and other distortions. Typical problems include capturing the roof outline of the structure rather than the building area, extraneous detail in the captured footprint, non-orthogonal angles in captured outlines, deficient outlines owing to occlusion, collinear vertices, protrusions and intrusions as artifacts of automated building feature extraction, etc. as seen in Figure 2.10. In Figure 2.10, building A was digitized manually and building B was extracted through automated feature recognition routines from aerial images. Note the extraneous detail and collinear vertices in building footprint A and the protrusion and intrusion artifacts in building B.

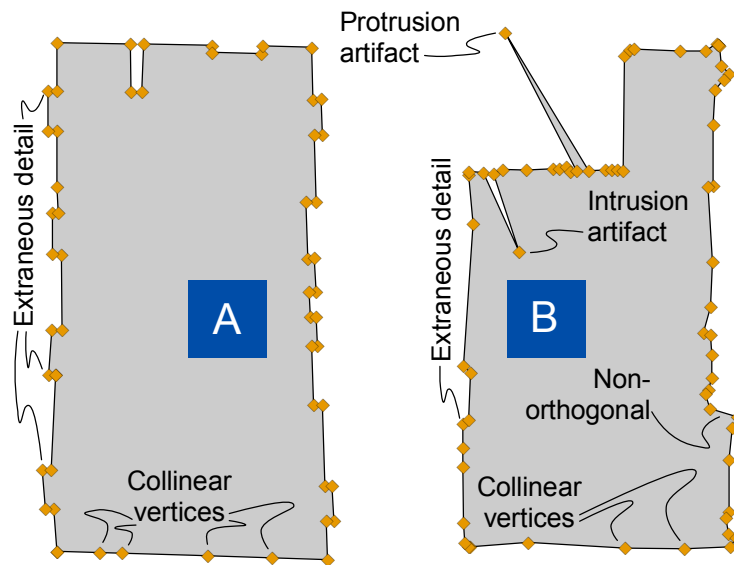


Figure 2.10 -- Problems in acquired building footprint polygons

Depending on the design of the feature extraction stage, details extraneous to the shape may have to be removed, or the perimeter contour vertices may have to be densified or decimated, or perimeter contour segments may have to be orthogonalized. Again, depending on the methodology, the building footprints may have to be normalized

with respect to selected transformation parameters such as translation, scale and rotation.

2.4.2.2. Shape Representation

After the shape has been acquired and pre-processed, it has to be represented in a manner appropriate for the task at hand. The representation of a shape or the method of computing its shape signature (Acharya and Ray 2005) is fundamental to the design of the shape analysis system (Costa and Cesar 2001a). The literature abounds with variations of shape representation typologies. Pavlidis (1978) suggests a shape reconstruction-based classification as information-preserving and information non-preserving. Others have classified shapes by thickness, or as is seen in more recent articles, boundary- or contour-based (thin shapes) or region-based (thick shapes). Contour-based approaches can represent the shape in the form of a stream of one-dimensional signals and may often be computationally less expensive than region-based two-dimensional signals (Dryden and Mardia 1998; Costa and Cesar 2001a).

Landmark-based shape representations of shape, though derived from contour-based approaches, deserve some special mention. Refer to an excellent survey of landmark point types for morphological characterization by Bookstein (1991). A landmark, in the context of shape analysis, is defined as “a point of correspondence on each object that matches that matches between and within populations” (Dryden and Mardia 1993, pp. 460) . Landmark points typically include nodes (end points and intersections) as well as salient points along parametric curves. The usage of nodes as landmark points is self-evident and does not require explanation. For polygonal (straight-line) segments, landmark points are usually natural features, particularly at points of inflection that typically allow complete reconstruction of the original shape. However, for parametric curves that consist of infinite sets of points, several methods

exist for choosing a sample of salient points that enable a reasonable reconstruction of the original shape (Fischler and Wolf 1994; Salih et al. 2006). Parametric curves require choice of landmark points that enable approximate reconstruction of the original shape – the amount of deviation from the original curve depends on both the choice and the linear density of landmark points (Bookstein 1991; Costa and Cesar 2001a). The choice of appropriate landmark points is often difficult and involves trade-offs between accuracy and processing speed and the application it is designed for, and is challenging to automate. Commonly used techniques to generate landmark points include the salience of points on the curvature of the curve (Fischler and Wolf 1994; Cesar and Costa 1995, 1996), or random sampling of the contour, or sampling strategies based on a specified number of points or minimum distances between points (Loncaric 1998). Note in Figure 2.11, a polygon may be represented as a sequence of contour points or alternatively, as a sequence of contour segments. Note also, that the first and last points, PT_1 and PT_9 are coincident in the GIS polygon geometry.

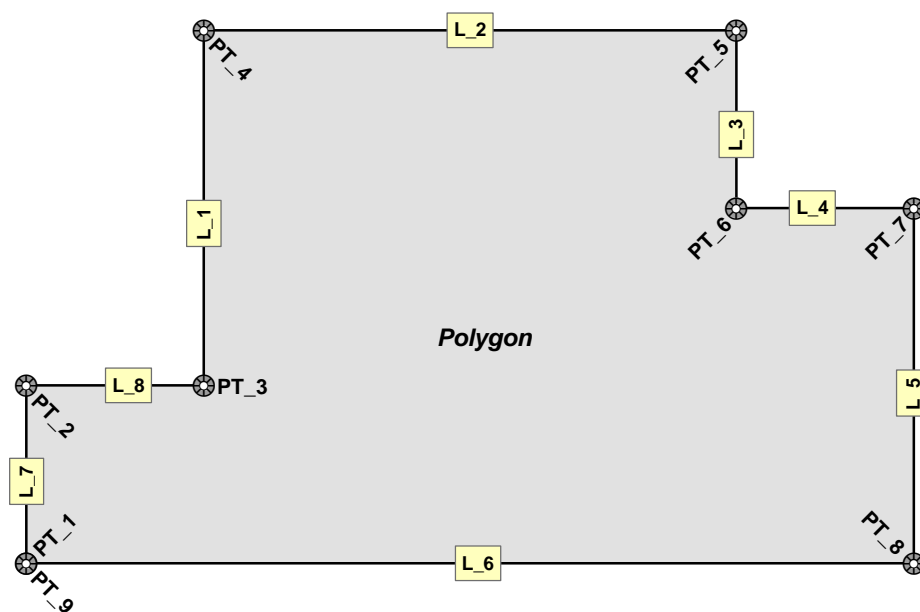


Figure 2.11 -- Landmark or Contour representations of a polygon

2.4.2.3. Feature Extraction or Shape Description after Shape Representation

Subsequent to shape representation, the feature extraction phase defines a set of techniques that extracts specific attribute characteristics about the shape. These attributes or features of the shape enable the description of the shape in the context of a specific task as well as develop measures that may be used for classification – often some particular subset of aspects of a shape will be more important than others in the context of the shape analysis or classification system. In addition, some shape aspects of a shape may be more important for a particular class of shapes – in this building footprint shape analysis research, straight line segments of null curvature and their orthogonal interconnections at corners are particularly important. In fact, researchers (see Attneave 1954) have identified corners as particularly significant in human shape analysis and corner aspects are formally parameterized in shape analysis systems as critical points of curvature (Super 2004) or as key landmark points (Ansari and Delp 1990) etc. In other cases, specific feature measures may be more important than the shape itself, particularly where size matching is an important aspect (Costa and Cesar 2001a). For instance, distance measures of the shape may be compared with a range of bounding polygon dimensions, and higher probabilities of successful classification arise when those distance measures are between the maximum and minimum threshold bounds.

2.4.2.4. Invariant Representations

The basic concept here is to extract certain intrinsic attributes from the object such that the properties are consistent over a wide variety of perspectives. These intrinsic attributes or feature descriptors may be directly measured (area, perimeter, orientation) or computed (circularity, elongation, concavity, etc.) from the contour geometry or derived from the entire shape (moments). In other approaches, the

geometric information is transformed into a different set of signals through log-polar or arc-tangent or arc-height mapping or Fourier sequences that are geometry-invariant. Combinations of discriminative invariant feature descriptors are compared between the sample and reference objects and usually, shape recognition is achieved through similarity metrics.

Direct Measurements of Shape Geometry

Specific shape characteristics are directly measured from the actual geometry of the shape, and include aspects such as *area, perimeter, Euler number, number of corners or segments, length of principal axes, length measurements at specified intervals* along major and minor axes, etc.

The location of the *centroid* of the shape is often an important parameter and estimated as the average values of the contour point coordinates. An often used shape aspect is the *shape diameter* that is the largest distance between any two points.

Shape Array Measures and Derived Measurements of Shape

The shape could be represented as an *array* of contour point coordinates. Alternately, based on a set of contour points and the centroid, various other array measures may be extracted. The shape could be characterized by computations on arrays of distances of the contour points to the centroid (Chang et al. 1991), or angles subtended at the centroid between successive contour points, or ratios of distance between successive vertices and the angle subtended by the two vertices at the centroid (Veltkamp and Hagedoorn 1999). Measures based on these arrays could include mean centroid-to-contour-point distance or mean boundary segment length. Ratios may be used when scale independence is required. The original arrays themselves could be modified through division by means or minima for the purposes of scale invariance

before transformation techniques may be applied. The shape complexity measure, which depicts the ratio between area and the square of the centroid-boundary mean distance, is an additional derived feature. The *Euclidean*, *Root-Mean-Square*, *Mean* and *Centroid Norms* are derived shape “size” measures that are invariant to translation and rotation (Costa and Cesar 2001a). Typically, these measures involve normalizing the squared distances of boundary points from the centroid (Kendall 1984; Dryden and Mardia 1998).

Thus, the 2n Euclidean Norm of a shape with ‘n’ boundary points and known centroid is given by

$$\|S\| = \sqrt{\left(\sum_{i=0}^{n-1} (P_{x,i} - P_{x,c})^2 + \sum_{i=0}^{n-1} (P_{y,i} - P_{y,c})^2 \right)}$$

where $P_{x,i}$ and $P_{y,i}$ are the ‘x’ and ‘y’ coordinates of the ‘i’th point, and

$P_{x,c}$ and $P_{y,c}$ are the ‘x’ and ‘y’ coordinates of the center of gravity of the shape

Similarly, the Centroid Size of a shape with ‘n’ boundary points and known centroid is given by

$$\|S_c\| = \sqrt{\frac{\left(\sum_{i=0}^{n-1} (P_{x,i} - P_{x,c})^2 + \sum_{i=0}^{n-1} (P_{y,i} - P_{y,c})^2 \right)}{n}}$$

where $P_{x,i}$ and $P_{y,i}$ are the ‘x’ and ‘y’ coordinates of the ‘i’th point, and

$P_{x,c}$ and $P_{y,c}$ are the ‘x’ and ‘y’ coordinates of the center of gravity of the shape

Measures of circularity and compactness relate the shape area to the square of its perimeter. Measures of rectangularity and concavity, relate the shape area to the area of the minimum bounding rectangle of the shape or the convex hull area of the

points along the contour. *Concavity* may also be expressed as a ratio between the perimeter of the shape and the perimeter of the minimum bounding rectangle. Thus,

- *Circularity*, $F_{cir} = (4\pi * A/P^2)$
- *Compactness*, $F_{com} = (16 * A/P^2)$
- *Rectangularity*, $F_{rect} = (A/A_{mbr})$
- *Area-based Concavity*, $F_{hole1} = (A/A_{chull})$
- *Perimeter-based Concavity*, $F_{hole2} = (P/P_{mbr})$

where A is the shape area, P is the shape perimeter, A_{mbr} is the area of the shape's minimum bounding rectangle, A_{chull} is the area of the shape's convex hull, and P_{mbr} is the perimeter of the shape's convex hull.

Elongation measures the ratio of the linear dimensions of the shape along the principal axes. Similarly, eccentricity measures the ratio between the longest chord of the shape and the largest chord length perpendicular to the longest chord. Other measures could include curvature and bending energy (refer to Costa and Cesar 2001a; Acharya and Ray 2005 for a thorough survey of scalar feature measures).

In order to compute or derive invariant representations, such as elongation or others that require the minimum bounding rectangle, often the *principal axes* of the shape have to be determined. A number of approaches have appeared in the literature, ranging from numerical searches and numerical analysis (Niu et al. 2002) to least moments of inertia (Gil-Jimenez et al. 2005) to least squares methods (Chaudhuri and Samal 2007). Chaudhuri and Samal (ibid) use a least squares approach to determine the slope of the principal axis using simple coordinate geometry and trigonometric

methods – the squared distance of the boundary points is minimized with respect to the centroid of the shape. Once the major axis is determined, it becomes a simple matter to determine the distance of the boundary point furthest above and below the principal axis, thus creating both the minor axis and the minimum bounding rectangle.

Direct and Computed Geometric Feature in Shape Analysis

Geometric feature measures as descriptors have been used in several applications, such as in the pharmaceutical, food, and chemical industries, where particle size and behavior (flow or compressibility or absorption) are strongly related to its shape (Realpe and Velázquez 2006). The process of identifying such features with high discriminatory power is difficult, and further, these feature descriptors are sensitive to noise. Finally, several different objects may have the same feature values or the same type of object may exhibit a wide variation in the feature values (Kashyap and Chellappa 1981; Zhang et al. 2003). Consequently, applications that use only direct and computed geometric measures of shape as feature descriptors for shape recognition or classification are rare, and often, geometry-based feature measures will be combined with other descriptors such as moment invariants in order to implement shape classification.

Moment Invariants

Another class of features that deserves special mention, because it is frequently encountered in shape analysis literature, consists of shape moments (Hu 1962; Bookstein 1991; Jiang and Bunke 1991; Wood 1991; Li 1992; Safaee-Rad et al. 1992; Trier et al. 1996; Loncaric 1998; Costa and Cesar 2001a). Generally, moments are based on the region, though vector-based calculations from point coordinates along the

boundary are not uncommon (Jiang and Bunke 1991; Adamek and O'Connor 2003).

Two-dimensional moments of a digital $M \times N$ image are given by

$$m_{p,q} = \sum_x \sum_y^{M-1, N-1} (x)^p \cdot (y)^q f(x, y)$$

where $f(x, y) = 1$, if the pixel belongs to the shape,

$f(x, y) = 0$, otherwise, and

$p, q = 0, 1, 2, 3 \dots$

For the same image, the central moments, or moments about the center of gravity are given by

$$m_{p,q} = \sum_x \sum_y^{M-1, N-1} (x)^p \cdot (y)^q f(x, y)$$

where $f(x, y) = 1$, if the pixel belongs to the shape,

$f(x, y) = 0$, otherwise,

x_c and y_c are the coordinates of the center of gravity of the shape

$p, q = 0, 1, 2, 3 \dots$

Note that $x_c = \frac{m_{1,0}}{m_{0,0}}$ and $y_c = \frac{m_{0,1}}{m_{0,0}}$

Central moments are invariant to translation and rotation. Size invariance is achieved by

$$\eta_{p,q} = \frac{\mu^{p,q}}{\mu_{0,0}^\gamma}, \text{ where } \gamma = [(p+q)/2]+1$$

In particular, Hu (1962) defines seven transformation-invariant functions, computed by normalizing central moments through order three, that are invariant to scale, position, and orientation. These seven functions are commonly referred to in pattern recognition literature as Hu's moments (whose moments? Oh, his moments. No, no, Hu's moments!). Thus, any shape may be uniquely characterized by a set of values of the moment invariants. Moment invariants, based on the entire shape, are globally scalar and represent a fairly fundamental and comprehensive set of information-preserving shape descriptors and therefore, figure consistently in standard pattern recognition texts (Li 1992; Duda et al. 2001).

Dudani et al (1977) generated moment invariant values for aircraft silhouettes and used them in an application that automated the process of aircraft identification. In a comparative study, Blumenkrans (1991) implemented Hu's moments to recognize simple objects by their moment invariant representations. More recently, Realpe and Velazquez (2006) characterized pharmaceutical powders by morphology and size by implementing moment variants realized from 640x480 pixel images of pharmaceutical powders. They calculated the seven invariant moment values for each particle in the image and compared the values with reference particles. Recognition rates were as high as 88%, with 1984 particles being recognized in 22 seconds, and the authors proposed the moment invariant recognition algorithm as an in-line production monitoring tool to classify granules by size and shape.

Moment invariant methods are mathematically concise and theoretically pleasing. However, the methods do have disadvantages -- higher order moments are extremely sensitive to noise, making it difficult to correlate shape features with higher order moments. Hu's moments are not orthogonal and therefore contain a high order of redundancy, but this disadvantage is easily circumvented by kernel-based

transformations of the original moments to yield orthogonal polynomials, such as the Legendre, Zernike, pseudo-Zernike polynomials or Chebychev moments which have minimum redundancy (Rothe et al. 1996; Zhang et al. 2003). For instance, Li (1992) used several higher order moments based on Hu's formulations to identify particular characters and found that half the integral variants were not used in the construction of moment invariant values, and concluded that such traditional moment invariant functions contain highly redundant information. Using various normalization transformations, Rothe et al (1996) modified Hu's moments and several other descriptors to alternate representations of Legendre and Zernike descriptors that are invariant to both geometric and affine transformations. The natural orthogonality in Zernike and Chebychev moments results in minimum information redundancy and lower sensitivity to noise, particularly in the higher order moments (Teague 1980). Finally, as with most scalar transformations, local shape information, particularly in high-curvature areas, is not captured adequately by Hu's moments (Loncaric 1998).

If the shape descriptor is transformed into a set of one-dimensional signals, such as normalized distance between boundary points and centroid, the method of moments may be easily modified to develop a classifier that is computationally less expensive and has the added advantage of being generated from the contour boundary (Gupta and Srinath 1987).

Representations based on Shape Transformations

In these methods, first, the object is *described* by one of the representation modes described earlier and then *transformed* into another set of signals that serve as an alternate representation of the shape. The transformation function determines whether the image can be reconstructed exactly (information-preserving) or approximately (information non-preserving). In some cases, the shape analysis

application may be concerned only with classification and not with image reconstruction, so the transformation function might be designed to have high discriminatory power, but allows very approximate image reconstruction.

Two-dimensional shape information may be converted to a stream of one-dimensional signals through many methods, including tangent angle versus arc length (Zahn and Roskies 1972; Bennett and MacDonald 1975; Arkin et al. 1991), complex functions made periodic by repeating contour arc lengths (Richard and Hemami 1974; Persoon and Fu 1977), centroid-based signals of distances or angle sequences from contour boundary points (Gupta and Srinath 1987; Chang et al. 1991), partitioned sequences of boundary segments (Liu and Srinath 1990; Wang et al. 1994; Cesar and Costa 1995), arrays of distance and angle, etc.

Once the signals have been generated, the shape signature is recomputed using the discrete Fourier transform with a specified number of coefficients as appropriate for the application (Zahn and Roskies 1972; Kiryati and Maydan 1989; Ashbrook and Thacker 1998), or normalized Fourier transform for two-dimensional signal streams (Rothe et al. 1996), or wavelet transforms (Acharya and Ray 2005) or the Gabor and the Karhunen-Loève transforms (ibid), or conversion to bending energy representations (Young et al. 1974; Morse 2007).

2.4.2.5. Statistical and Mathematical Approaches

In the statistical approach to shape analysis, as the name implies, shape patterns are assumed to be generated by a probabilistic process, and can range in application from very simple Bayesian approaches to support vectors and neural network-based classifiers. Since the early 1980s, statistical approaches that treated shape signals (after appropriate transformations) as periodic functions used time series and

autoregressive modeling concepts to analyze shapes (Kashyap and Chellappa 1981; Kartikeyan and Sarkar 1989; Ansari and Delp 1990; Das et al. 1990).

The main disadvantage with autoregressive modeling is that the relatively small number of parameters may not be sufficient for complete shape description, particularly where the edge is complex (Loncaric 1998). Further, the specification for the order of the autoregressive model is not always straightforward, since many papers in this genre attempt specifications with different lags and choose one with the best performance. The autoregressive approaches were predominant in the early 1980s to the mid 1990s.

More recently, several statistics-based applications have extracted features from one- and two-dimensional signals and transformed these features into higher dimensional space to enable the use of linear classifiers (Leventon et al. 2000; Golland et al. 2005).

Recent developments in statistical shape theory represent objects as points in higher dimensional shape space, termed a "manifold" (Kendall 1984), such that all potential poses of an object caused by translation, rotation or scaling correspond to a single point in that shape space. Recognition and classification may be achieved by computing the geodesic distance between a sample and reference object. Thus, if the sample was generated by a geometric transformation (translation, rotation and/or scaling), the geodesic distance between the sample and the reference will be zero.

Since the mid 1990s, mathematical approaches that redefine coordinate systems, which eliminate standard transformations or describe objects as points in higher dimensional space are becoming increasingly common (Kendall 1984; Bookstein 1991; Dryden and Mardia 1993; Grenander 1996; Dryden and Mardia 1998; Comaniciu

and Meer 2002). The reader is also asked to refer to Schalkoff (1992) or Webb (Webb 2002) for a detailed introduction to statistical pattern recognition.

Dryden and Mardia (1993) suggest a mathematical approach where the traditional location is removed from the description of a shape through matrix decompositions, till the shape is described by a *hypersphere of unit radius in a higher dimension non-Euclidean space*. They suggest approximation of the hypersphere by a tangential hyperplane in local space when the variations in a dataset are small. They implement their shape space approach by analyzing skulls of macaques through 7 landmarks and determine if the skulls of male macaques are different in mean shape from females. The paper also suggests similarity metrics based on non-Euclidean distances such as the Procrustes distance or the Riemannian distance. The *Procrustes distance* represents the closest chord on the hypersphere between two transformed shapes, and the *Riemannian distance* is the closest great circle distance along the hypersphere between two transformed shapes (Kendall 1984).

The *shape space approach* provides a comprehensive representation of the object that is invariant to any standard transformation. The comprehensive representation also makes the recognition process less sensitive to noise or occlusion. Additionally, representation in higher dimension shape space enables greater classification efficiency, in the sense that higher dimensionality permits the use of effective linear classifiers. Finally, well known statistical pattern recognition techniques may be extended into non-Euclidean space. However, the shape space methods are mathematically dense and the theory for such descriptions is still being developed. Implementing classification schemes based on the shape space approach are computationally burdensome. Finally, the problem of classifying building footprint polygons is legitimately a trivial problem for implementing a shape space methodology.

2.4.2.6. Structural and Syntactic Methods

Interest in structural pattern recognition corresponded with the realization that using only invariant features or their transformations might not be enough to efficiently recognize or classify shapes. Analytical limitations in existing methods required the representation of shape components through symbols and their spatial relationships (Fu 1982). Research was beginning to get directed to the structure of the shape towards the mid 1970s, emphasizing relationships between and among parts and between parts of the shape and the whole shape (Pavlidis 2003). Syntactic pattern recognition also deals with parts of the shape and their interrelationships, but emphasizes that the process follows syntactic rules of composition (Bunke and Sanfeliu 1990). Structural pattern recognition relies on the extraction of features that are attributes of parts of shapes or attributes of relationships between parts of shapes (Pavlidis 1972). In fact, Pavlidis argues that structural approaches have little theoretical bases and are more philosophical than methodological, and that there are no general methodologies available for direct application (Pavlidis 2003). Syntactic approaches however, have a strong theoretical basis because the theory of formal languages is well developed, as can be seen in the post 1970 period (Fu 1982; Bunke and Sanfeliu 1990). However, for shape analysis based on descriptions of shape component relationships, structural representations such as string contexts, trees and graphs began to be increasingly used (Pavlidis 1972). The overlap between structural and syntactic approaches to shape analyses prompted the unification of the two fields since the 1980s -- today, the two fields are generally viewed as one, largely based on Fu's work (1982, 1986), where shapes are described and analyzed by their components and the interrelationships between components.

Structural and syntactic methods examine shapes through their component relationships in more complex terms than is allowed by view-invariant or statistical methods. Structural methods typically describe shapes through graphs and topological concepts; hierarchical formulations are common in such representations (Pavlidis 1972, 1979). Syntactic approaches represent shape through strings according to rules specified in a formal language.

Structural or syntactic methods have also been successfully used in shape analyses for classification (Fu 1982). The main advantage with both structural and syntactic methods is that in addition to successful classification, the methods include intrinsic descriptions about the objects, and how the original shape may be reconstructed accurately. These methods are used in analyzing complex shapes by breaking down the overall pattern into a series of sub-patterns, each of which is described by a sequence of primitives based on a specified syntax (Jain and Dubes 1988). Shapes may be represented using topological concepts, as parts and connections or relationships. Typically, these relationships are described using graphs, trees or strings (Zhu and Yuille 1996; Chen et al. 1998; Gdalyahu and Weinshall 1999; Latecki and Lakamper 2000), and shape analysis methods based on these representations achieve high classification or reconstruction accuracy despite tremendous shape variability. Recognition is achieved by minimizing the costs of transforming one shape descriptor (graph or string) to another (Wu and Wang 1999; Kaygin and Bulut 2002).

Pavlidis (2003) suggests that syntactic methods have not found universal applicability despite their sound theoretical foundations and shape recognition potential because they currently do not have good algorithms for inference and that rules based on formal language do not provide good bases for how components are integrated into

the whole shape. In addition, they are often difficult to automate and computationally expensive. On the other hand, structural approaches based on correspondences between shape components have gained recent popularity because similar shapes have similar primitive arrangements or component sequences that can be matched or aligned at significantly lower cost than scale-space or other mathematical approaches (Super 2004).

A Note on Structural and Syntactic Shape Analysis

In general, structural and syntactic shape analysis is based on the premise that a shape comprises of simple components that are composed of even simpler components or primitives. The structure and relationships between the primitives are analogous to the theory of formal languages – a sentence is likened to a shape, the words to its simple components and the alphabet to its primitives (Basu et al. 2005). The meaning of a sentence depends on a sequence of individual words strung together using a grammatical framework based on linguistic forms that reflect thinking and rules that establish consistency (Bellone et al. 2004). In addition to syntax-based classification, the rules provide a composition methodology to derive the whole shape from its primitives (Pentland 1987). Syntactic approaches therefore require the selection of an appropriate grammar, the use of a descriptive method (topological trees, planar graphs or string symbols), the choice of an optimal set of primitives (too many primitives may make the approach too cumbersome to implement, while too few may result in poor discrimination), inferential techniques to learn the syntactic rules from sample objects and parsing methods to decompose shapes into simpler components with a view to ascertain if the components follow the specified grammar (Basu et al. 2005).

In some cases, the decomposed shapes cannot be suitably expressed in the context of a grammar – here, components are represented through symbols or string

data structures or trees or graphs, usually by hierarchical sets of prototypes (Pavlidis 1979; Chen et al. 1998). Recognition is achieved when a component pattern expressed as a string matches or resembles the string of a reference shape (Gonzalez and Thomason 1978). Graphs or topological trees may be used for the same purpose, and are generally more descriptive than strings, but are computationally resource-hungry and difficult to automate (Bicego et al. 2006). There are several examples of shape analysis applications that circumvent this problem by introducing constraints or contextual information (Belongie et al. 2002) or using sub-optimal methods or implementing heuristic approaches (Gdalyahu and Weinshall 1999). Recognition is usually based on dynamic programming techniques, clustering methodologies or similarity metrics based on edit distances. Edit distances are simply the costs or weighted costs associated with transforming one string (or graph, or tree) into another through elementary editing operations such as substitutions, additions and deletions (Kaygin and Bulut 2002).

Structural Analysis – The Medial Axis Transform

Among the most researched region-based structural representations is the family of “medial axis transforms” or MAT, a term first coined by Blum (1967). In concept, the shape is represented using a linear graph, a stick-like skeleton (Loncaric 1998), derived from a transformation of the entire shape. Terms synonymous with MAT include shock graphs, symmetric axis transform, skeleton transform or skeletonization (Torsello and Hancock 2004). Conceptually, MAT are based on the premise that most of the information about a shape is contained within its topology (Trier et al. 1996; Sebastian et al. 2001).

Several methods exist to generate the MAT from a polygonal shape. One approach is based on Voronoi tessellations (Ogniewicz 1993; Skiena 1997; August et al.

1999) generated around equally-spaced points on the contour of the shape, as depicted in Figure 2.12.

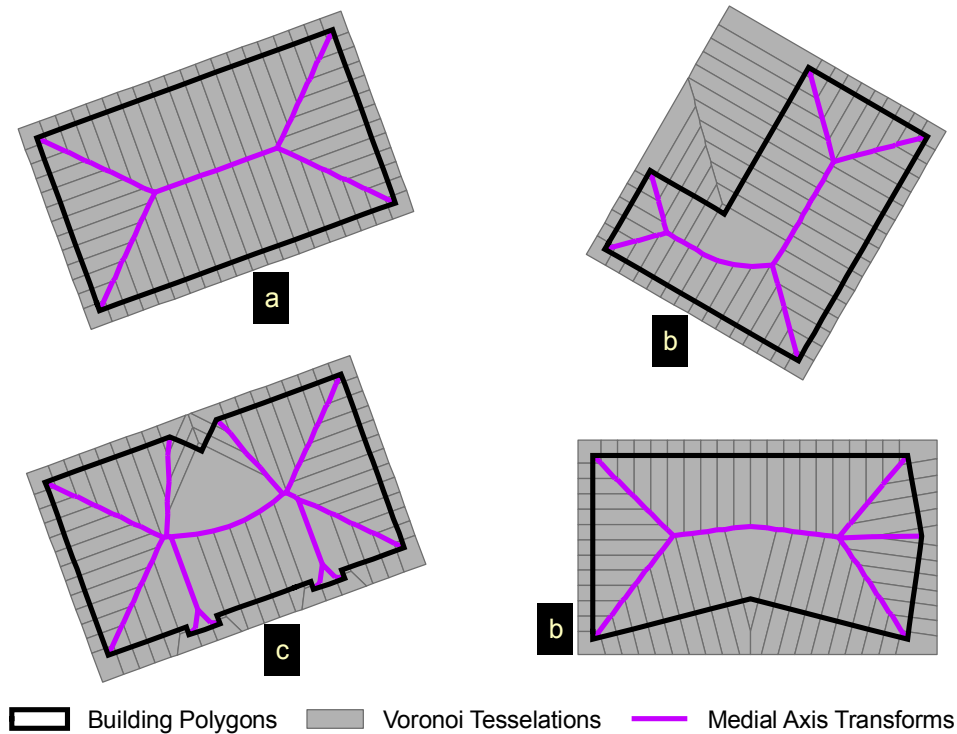


Figure 2.12 -- Medial axis transformations from Voronoi tessellations

Another approach, shown in Figure 2.13, is based on generating small polygonal buffers internal to the original polygonal buffers, or thinning the original polygon until what remains is a linear graph feature – similar “thinning” algorithms exist for raster or pixel structures (see Lam et al. 1992 for an excellent survey on thinning techniques). Other approaches are based on drawing lines inwards from convex landmark points that connect centers of circles that are tangent to at least two points on the boundary of the shape (Torsello and Hancock 2004), and even based on electrostatic field approaches (Grigorishin et al. 1998)!

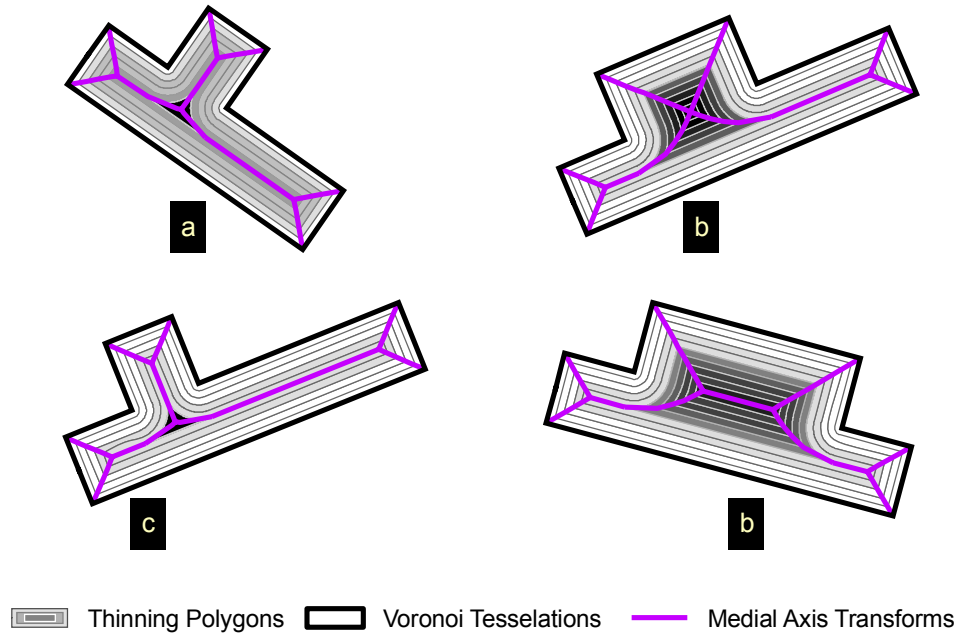


Figure 2.13 -- Medial axis transformations based on thinning routines

However the shock graphs or skeletons or MATs are created, they are then described in structural or topological terms (Arcelli and Baja 1985), or as ratios of change in boundary length to distance along medial axis (Sebastian et al. 2001; Torsello and Hancock 2004). Classification or recognition of a shape occurs when the sample object's topology matches that of a known reference (August et al. 1999). Recognition is typically achieved through dynamic programming or computing edit transformation costs associated with changing the input representation to the reference description (Fu 1982). MAT are very sensitive to local perturbations of the boundary or holes within the region, and representation methods to compute transformations that are less sensitive to region changes are fairly difficult to automate (Ashbrook and Thacker 1998). Recently, Katz and Pizer (2003) criticize the extreme sensitivity of MAT and the inherent difficulty in decomposing the MAT into a set of connected line primitives that reflect an intuitive parts hierarchy.

Structural Shape Analysis – Derivation of Shape Numbers

Another interesting manner of representation describes the shape in terms of connected curves or corners. It is a one-dimensional notational representation that is independent of geometric transformations, and generates unique sequences of coded numbers based on the convexity or concavity or collinearity of the connections between boundary curves (Bribiesca 1981). The coded numbers are rotated till the minimum number is reached, making the shape invariant to rotation. The method incorporates some degree of fuzziness in the representation by first converting the shape into a grid – the accuracy of the representation depends on the resolution of the grid size. Using Freeman chains (specific numbers for movement along the cardinal direction, where W=1, N=2, E=3 and S=4), Bribiesca describes the gridded shape as a sequence of numbers, and further, uses Freeman Chain corner derivatives (specific numbers for corner types, where convex corner = 1, straight corner = 2 and concave corner = 3) to describe the gridded shape as another sequence of numbers. This normalized differential chain code is termed the “shape number” (Bribiesca and Guzman 1980; Morse 2007). See Figure 2.14 for a diagrammatic representation of how shape numbers are derived for an arbitrary shape.

Shape numbers are very sensitive to image extraction artifacts, so this approach usually generalizes the contour edge and creates grids of various resolutions that are orthogonal to the principal axes of the shape. The number of grid edges making up the boundary of the gridded polygon specifies the “order” of the shape number. While the order clearly depends on the resolution of the grid, for a given order, the shape number is unique (Bribiesca and Guzman 1980).

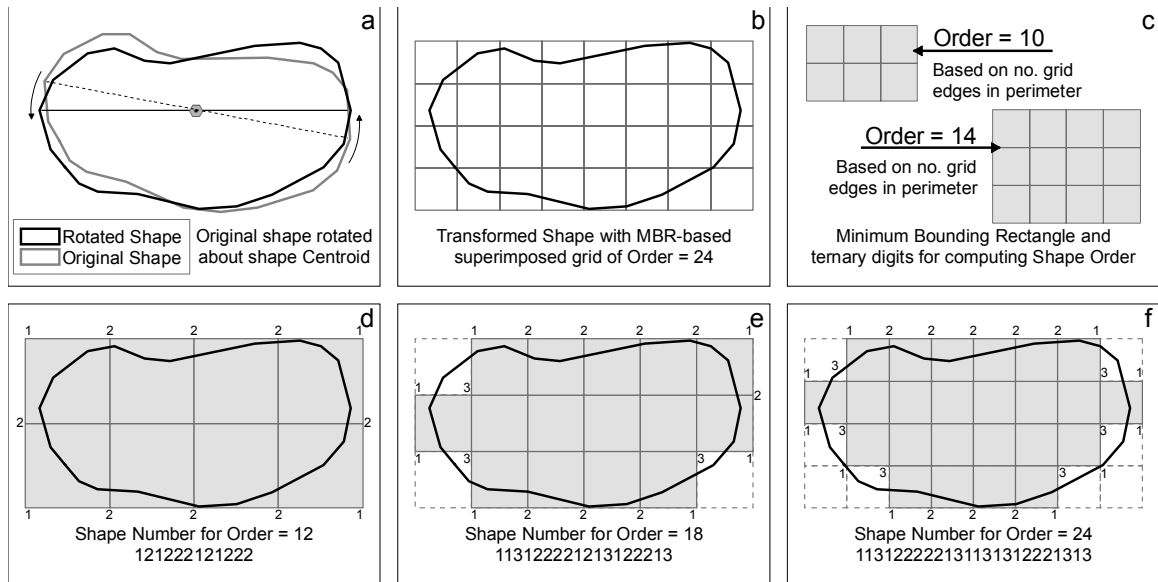


Figure 2.14 -- Deriving unique shape numbers for specific shape orders

2.4.2.7. Syntactic Shape Recognition

Applications that require comparative analyses of objects (shapes or texts) and that can be represented as sequences or strings of elements may be implemented through string matching. String components may be *symbolic* or *attributed* – symbolic strings are composed of a determinate set of discrete building blocks or alphabets that are combined in accordance with a set of syntactical rules, while attributed strings are associated with quantitative measures that correspond with the semantic or contextual characteristics of the components (Yang and Pavlidis 1990).

In string matching, one attempts to identify all incidents of a pattern within a superset, where both the pattern and the superset are composed of the same primitive components. In a typical pattern-spotting implementation, a *sliding aperture* of the same width as the pattern moves sequentially along the superset till a match is found. Other approaches, typically seen in optimization contexts, include *dynamic programming*, *elastic or spring matching*, *dynamic time warping* or *edit-operation based*

transformations (Chen et al. 1998). *String transformations* generally convert one string to another through additions, substitutions and deletions, each with associated costs, and the total cost of transformation is the sum of all the edit costs involved in the transformation. While there are several methods of transforming one string to another, the preferred transformation would be that one which incurs the least total cost. In addition, it must be noted that strings as definitions of objects are not unique, and circular shifts on a string may lead to completely different results.

2.4.2.8. Shape Recognition and Classification

In the final step, the shapes that have been processed and represented are analyzed and classified. In most applications, classes are specified a priori, and the analysis recognizes an input shape as belonging to a class – this method of classification is termed supervised classification. In other applications, the classes are not predefined; rather, classes are created during the analysis phase and subsequent input shapes are assigned to classes that they are like or new classes created. This type of analysis is called unsupervised classification. However, in both methods, input shapes are compared to previously created classes and measured as to how similar they are (Duda et al. 2001; Acharya and Ray 2005).

Classification algorithms therefore depend on computing indexes of shape similarity that are essentially objective and quantifiable measures of how close or similar one input shape is to another (unsupervised classification) or how similar an input shape is to shape representatives of predefined classes (supervised classification). It is now obvious why the shape description step is crucial for the overall shape analysis application – the representation provides the discriminative basis for quantitatively measuring similarity and therefore potential membership to a class. However, despite considerable investment in research in the field of pattern recognition, there are no

general methods to identifying the best set of features or for creating the perfect classifier (Costa and Cesar 2001a). In a very general sense, an optimal classifier puts objects that share some attributes in the same class while other objects with distinctly different properties are placed in other distinct classes.

The literature varies significantly in terms of using the feature descriptors of the shape in developing classifier mechanisms that use similarity computations such as Manhattan, Euclidean, Minkowski (Veltkamp 2001; Black 2004a, 2004e, 2004d, 2004c; Shahrokni et al. 2004; Barile 2008) or Mahalanobis (Jain and Dubes 1988; Dwinnell 2006) distances to determine class membership. Typically, the shape feature vector or its transformation is compared with all the reference shape classes it can potentially belong to, through the similarity measure – the shape will then be assigned to the class that it is closest to. Structural methods achieve recognition or classification typically through correspondence between component primitives (Liu and Srinath 1990; Loncaric 1998; Latecki and Lakamper 1999, 2000; Belongie et al. 2002). Syntactic methods implement classification through dynamic correspondence between component parts based on edit costs or Levenshtein distances (Chen et al. 1998; Kaygin and Bulut 2002; Black 2004b).

2.5. Geometry Manipulations in the GIS Environment

This section describes the representation of polygons within the GIS and some techniques for pre-processing the building footprint polygons, based on methodological aspects uncovered in the literature review. Data structures for spatial data representation in GIS vector formats vary in different software application. Even though all the pre-processing and computational geometry routines were executed in the ESRI® ArcGIS 9.x system [henceforth ArcGIS], the following sections will attempt to explain the representation and pre-processing in generic terms within the ArcGIS environment.

2.5.1. Representation of Points, Lines and Regions in GIS

Spatial data is represented primarily in two architectures, the vector and raster formats respectively in the context of a spatial reference that combines a projection method with a coordinate system (Antenucci et al. 1991). Vector architectures abstract real world information and represent them explicitly as points, lines or polygons, with their spatial relationships represented implicitly. Real world phenomena that show locations with little dimensional information are represented as points and described in terms of x,y coordinate locations. Regions in the real world that have appreciable length and width are represented as polygons and typically described as an ordered sequence of x,y coordinate locations that define the closed polygon edge. Depending on the scale of representation therefore, a city could be represented as a point or a polygon. Real world features that are much longer than broad lend themselves to representation as lines (or polylines), described as sequences of x,y coordinate locations (Demers 1999). The “spaghetti” data structure that encapsulates points, lines and polygons as strings of coordinate locations is depicted in Figure 2.15 (Lakhan 1996).

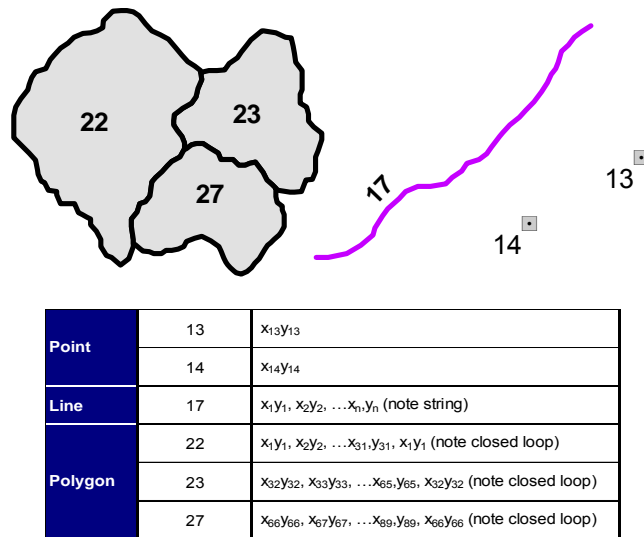


Figure 2.15 -- Spaghetti data structure for feature representation

2.5.2. Topological Data Structures

More commonly used is the *topological data structure*, where spatial relationships are explicitly referenced in sets of relational tables (Environmental Systems Research Institute 2007) and depicted in Figure 2.16.

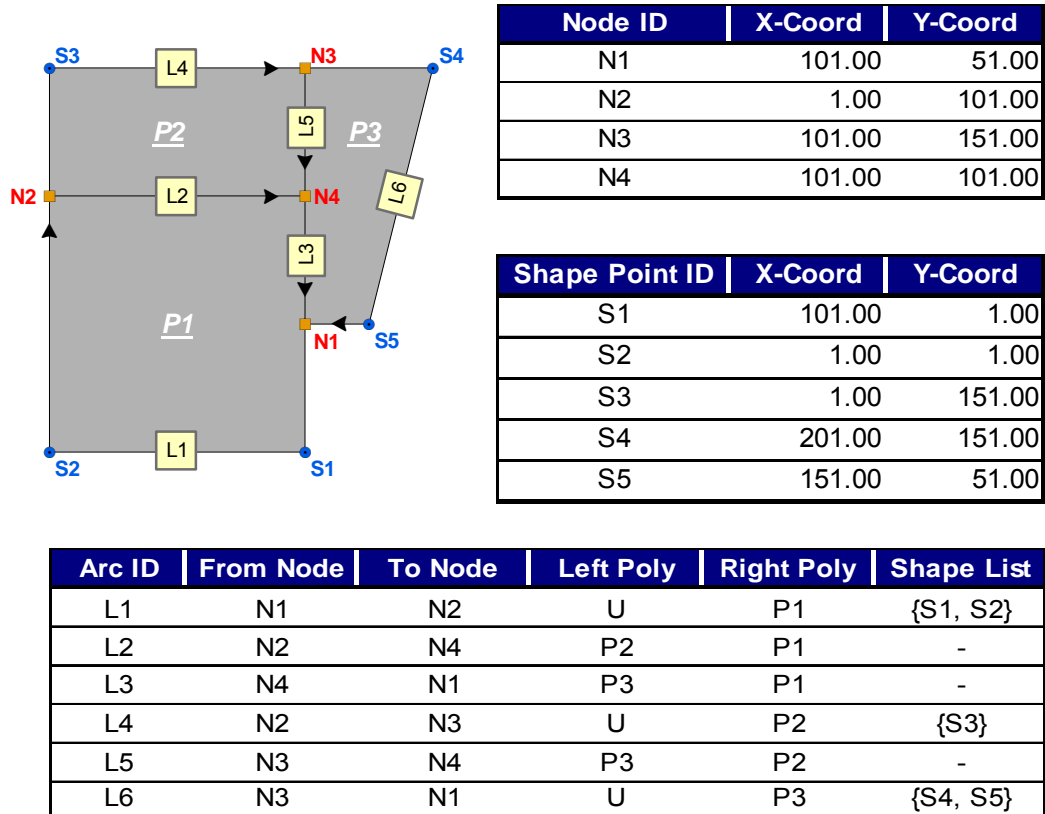


Figure 2.16 -- Topological data structure for polygon spatial data representation

As before, points are represented by their x,y coordinates and uniquely labeled. Points are classified as shape points or nodes – shape points are connection points between two line segment primitives that give shape to the complete line representation, while nodes are beginning or termination points, or junctions of three or more line features. Lines are also uniquely named and represented by labels for starting node, ending node and shape point lists, and therefore have explicit directionality. The line

table also has polygon labels for the polygon on the left and the polygon on the right of the line segments (since the lines have explicit directionality, left and right sides to the line are defined), and therefore incorporates contiguity explicitly. Coordinate locations for the lines are fetched when required by relating the appropriate point label with the node or shape point table. Polygons are uniquely named and represented by ordered sequences of line labels that constitute their edges. When required, polygons fetch lines for their edges through successive relationships with the line table and then from the line table to the point table.

2.5.3. Geometric Primitives and Object Hierarchy

In the context of working and manipulating the geometric information for vector datasets within the ArcGIS ESRI™ application, all features are composed of objects that follow the Component Object Model [henceforth COM] and have an Application Programming Interface [API]. These objects are organized in various libraries and have properties and behaviors that developers can programmatically use for specific applications that are not part of the software graphic user interface. Specifically, objects for manipulating feature geometries are available from the geometry API library. Higher level geometries (like polygons and polylines) may be generated from primitive part-geometries that are schematically shown in Figure 2.17, in a hierarchy derived from the complexity of the primitive.

Polylines are used to characterize real world linear features (such as roads, rivers, etc.) and may be represented as a sequence of point primitives (or vertices) or segment primitives or paths. A segment is a function (straight line, part of a curve or ellipse, etc.) that describes a curve between two points, and consists of a pair of point primitives if it is a straight line. A path is a sequence of connected segments, or a sequence of point primitives. Polygons are used to symbolize real world features that

occupy real space at the scale of representation and are usually denoted by their linear edges. The edge of a polygon may be represented as a sequence of point, polyline, path, ring or segment primitives. The only new definition here is a ring, which is a closed path (begin and end points coincide), and may comprise of a sequence of point, polyline, path or segment primitives. Understanding the geometric hierarchy or the alternate programmatic digital representation of features is critical for writing and implementing spatial computing algorithms.

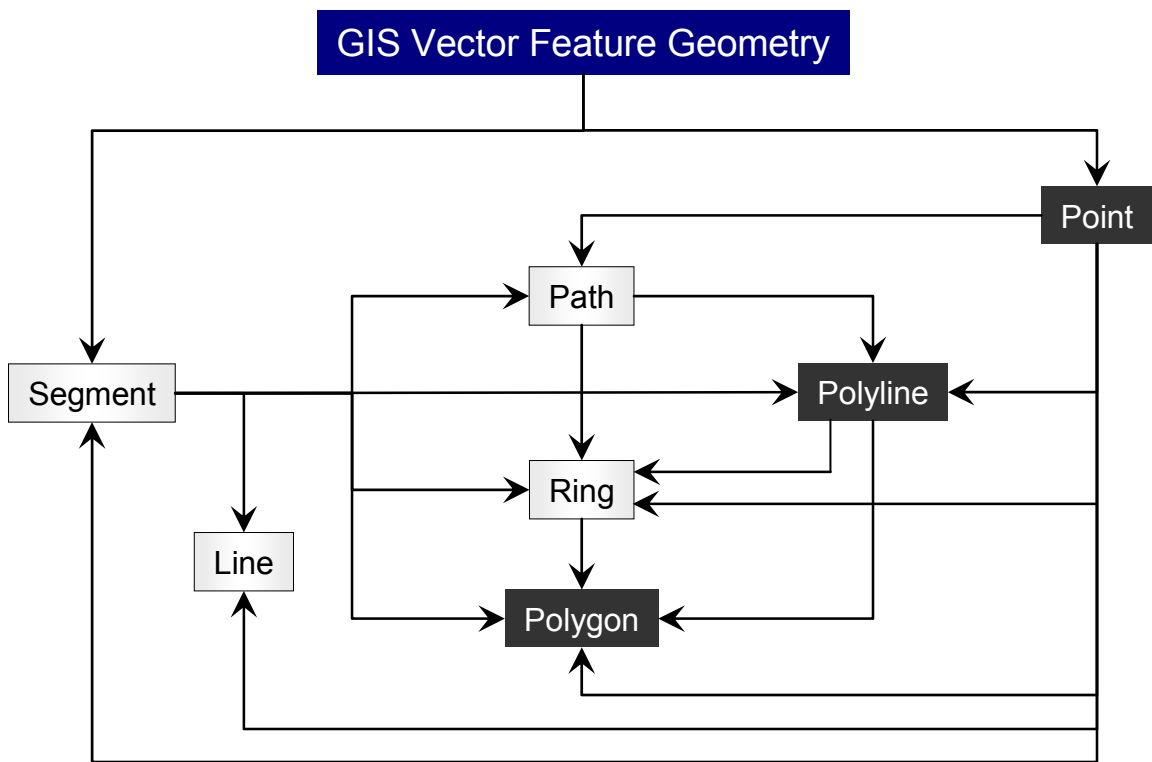


Figure 2.17 – Hierarchy of primitive part-geometries in a COM-based API

2.5.4. Manipulating Vector GIS Feature Geometry for Shape Preprocessing

In the preprocessing stage, often, the geometries of the input shapes will require some level of manipulation. For instance, in order to extract array-based feature measures of shape geometry (centroid to contour vertex distance), the input GIS

polygon shape will require several points or vertices along its edge. If a rectangular polygon is defined by just the corners, the vertex count may not be enough to permit further analysis. Therefore, the number of vertices defining the rectangular polygon has to be increased such that all the new vertices lie on the edge of the rectangular polygon. In cases of log-polar descriptors where the form requires distances of contour vertices from the centroid such that vectors from centroid to vertex subtend equal angles at the centroid, existing vertices must be moved and new vertices introduced at the edges, conforming to the equal centroid-vertex requirement. These examples require “densification” of the edges of the polygon. In some cases, it may be necessary to reduce the complexity of the edge through “generalization” and remove vertices such that the general shape of the edge is retained, but the number of vertices describing the edge is far less than the original shape. In still other cases, particularly for landmark point identification of 2D polygons, input GIS polygons may need to be processed to remove extraneous vertices and retain only those that match the landmark criteria. This example requires a customized vertex decimation strategy that could include aspects from generalization, densification and other coordinate geometry routines. When input shapes are derived from vector GIS polygons, some level of preprocessing may be necessary to manipulate the shape for downstream shape analysis.

2.5.5. Densification of Edges and Polylines

Densification of polylines (and polygon edges) is relatively straightforward in the context of a GIS. Densification routines are usually executed in order to generate a larger sample of point signals from an existing curve or polygon edge. Densification is also used in cases where segments that are described parametrically (such as three points on a circular arc of specified radius) have to be approximated as a series of linear segments.

Densification routines are executed as one of two types, both of which require a maximum vertex distance argument. The maximum vertex distance argument specifies the maximum length of the approximating line segments, or the maximum Euclidean distance between successive vertices. The first densification algorithm, called densification by maximum deviation, is based on the maximum perpendicular distance between the original segment and the approximating segments. In other words, the maximum deviation specified the maximum Euclidean distance any approximating line segment may be from the original polyline. Densification by maximum deviation is depicted in Figure 2.18.

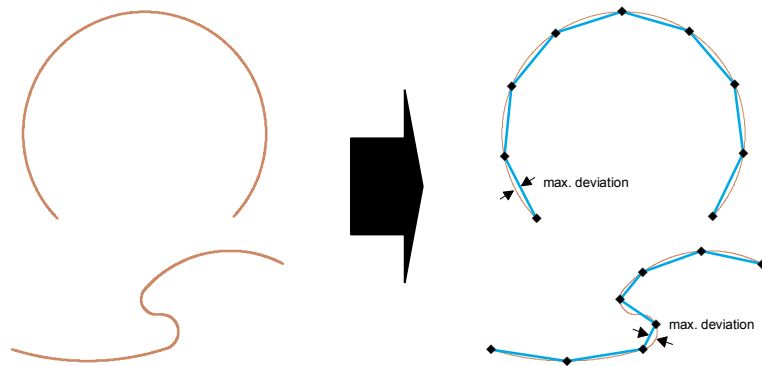


Figure 2.18 -- Densification by maximum deviation

An alternate densification routine, termed densification by maximum angle, specifies the maximum angle that any approximating polyline may be relative to the original segment. See Figure 2.19 for a diagram of densification by maximum angle.

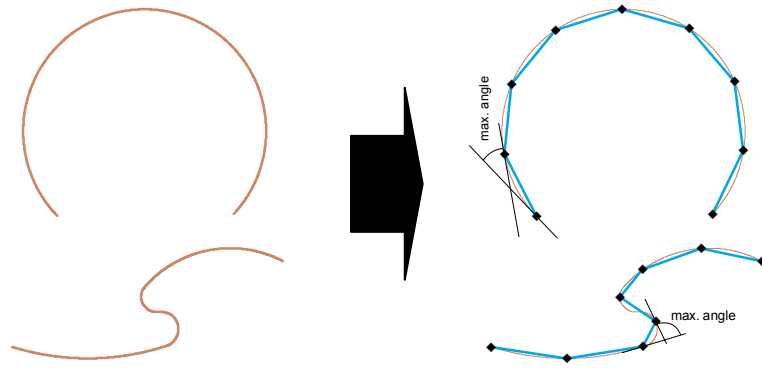


Figure 2.19 -- Densification by maximum angle

Note that both densification algorithms produce approximating line segments that are less than the maximum vertex distance. For polylines that are linear segments, the maximum deviation or maximum angle has no effect, and the only argument relevant is the maximum vertex distance. In other words, straight line segments are densified by adding vertices such that the length of the smaller linear segments generated in the output are less than the maximum vertex distance. Figure 2.20 demonstrates a densification routine applied on the linear edges of a polygon.

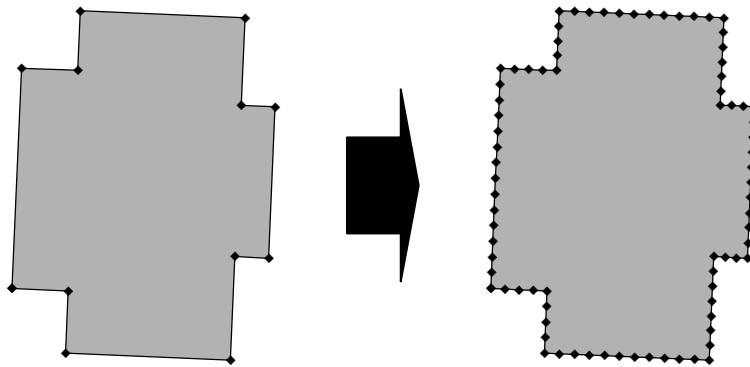


Figure 2.20 -- Densification of linear features

2.5.6. Generalization, Polygon Approximation and Line Simplification

Generalization is “a process which realizes transitions between different models representing a portion of the real world at decreasing detail, while maximizing information content with respect to a given application” (Weibel and Dutton 1999). In other words, generalization routines coarsely represent the real world while attempting to maintain the maximum possible validity in geometric and semantic correspondence.

Line generalization, also termed line simplification, attempts to represent input polylines by approximate output polylines with fewer vertices, while maintaining the topological connectivity among the polylines and preserving as much of the initial morphology as possible. Polygon approximation attempts to represent input polygons by approximate output polygons, primarily by generalizing the linear edges of the input polygons while attempting to maintain the topological character of the dataset. Maintaining the original topology is a difficult problem, and in most cases, new vertices and lines may be created and old ones deleted (Johnston et al. 1999).

2.5.7. Generalization Routines and Vertex Decimation Strategies

Considerable research has been directed towards automated generalization, especially in the context of GIS-based technologies (McMaster and Shea 1992; Baelia et al. 1995; Joao 1998). A number of these approaches attempt to produce maps at different scales by generalizing graphics from a back-end spatial database (McKeown et al. 1999). Several solutions have been integrated with GIS-based applications (Lee 2003).

Generalization has two motivations, one being map-based or cartographer driven (visual) and the other database or model-based (conceptual) (Muller et al. 1995a; Muller et al. 1995b; Weibel and Jones 1998). Map-based generalization is primarily driven by

the use of maps as communicative devices, where the emphasis is on abstracted and reduced representation based on geometric and semantic feature relevance with visual clarity and legibility. In other words, the cartographer makes decisions on the necessity of feature inclusion or generalization based on the scale of the map, and the contextual relevance of the features, while ensuring that the symbology of different features do not interfere with each other. Model or database generalization, while serving a visual function, adjusts feature geometry based on scale levels, and technically produces multiple generalized manifestations of features for continually varying scale ranges. Object-oriented data structures and technologies lend themselves aptly for such continuous generalization functions (Yang and Gold 1997). In fact, a number of internet-based mapping services require feature representation at several scales, and research is being focused on dynamically altering/generalizing the features based on client requests as the need arises (Oosterom 1995; Cecconi et al. 2002), a delivery approach that is scale-less (Muller et al. 1995b). See Cecconi et al (2002) for an extensive survey of generalization operations that automate delivery for web mapping.

Database generalization may require altering features or even eliminating them and is performed by several major operations based on the geometry and the meaningful context of the features' relationships with other features and the particular relevance of the feature's display. These operations include feature selection, elimination, simplification, aggregation, exaggeration, collapse, displacement, typification, symbolization and refinement (McMaster and Shea 1989; Oosterom 1995; ESRI 1996) as shown below:

Elimination – features that are not semantically relevant (such as ramps for a set of highway features) or geometrically insignificant (small dangles or tiny polygons) are progressively eliminated, as depicted in Figure 2.21.

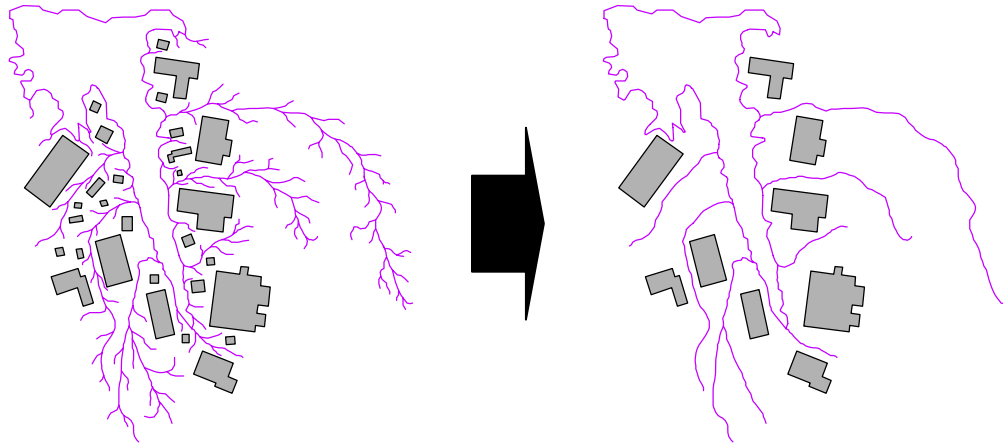


Figure 2.21 -- Eliminating features during generalization

Simplification or Line Generalization – where spikes, irrelevant detail and contour fluctuations are removed, usually by vertex decimation or translation, without compromising on the intrinsic shape, as seen in Figure 2.22.

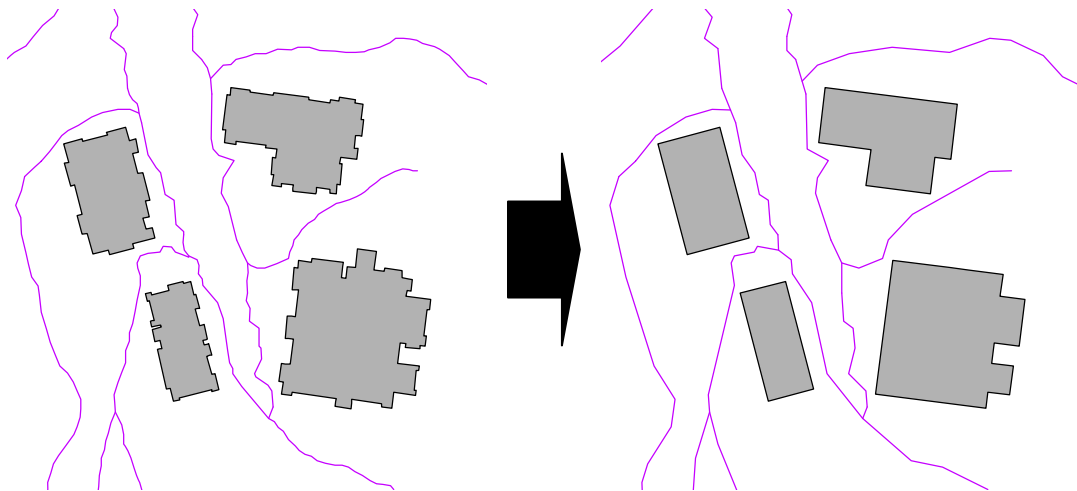


Figure 2.22 -- Simplifying lines and polygon edges during generalization

Aggregation – features that are adjacent or very close to one another are merged into single features (for instance, merging distinct agricultural polygons into a larger crop patch, or converting a cluster of points into a region feature when the points may not be distinctly seen at a certain scale) – see Figure 2.23.

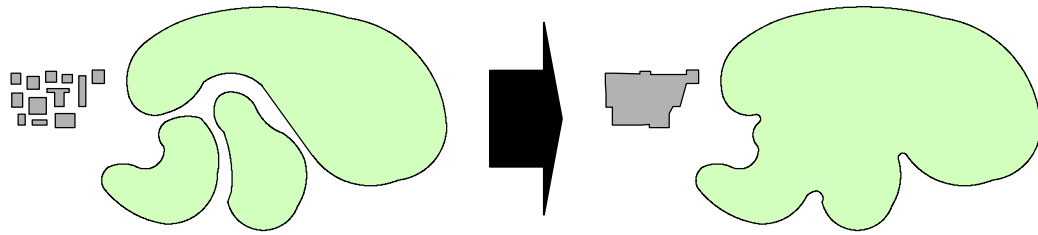


Figure 2.23 -- Aggregating polygon features during generalization

Exaggeration – increasing the size of particular features for semantic purposes (a wetland polygon that needs to be emphasized) or for clarity and legibility, as shown in Figure 2.24.

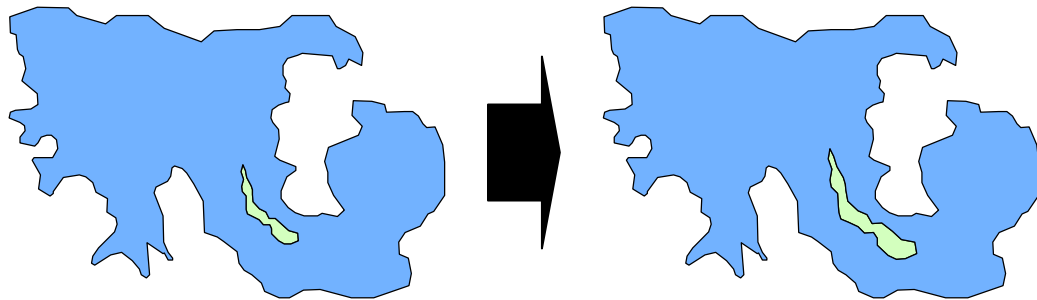


Figure 2.24 -- Exaggerating features for visual clarity during generalization

Collapse – reducing the dimensionality or size for legibility or semantic significance (such as converting high tension power infrastructure polygons into line features, or changing a polygon feature into a point if its area is less than a specified threshold), as denoted in Figure 2.25.

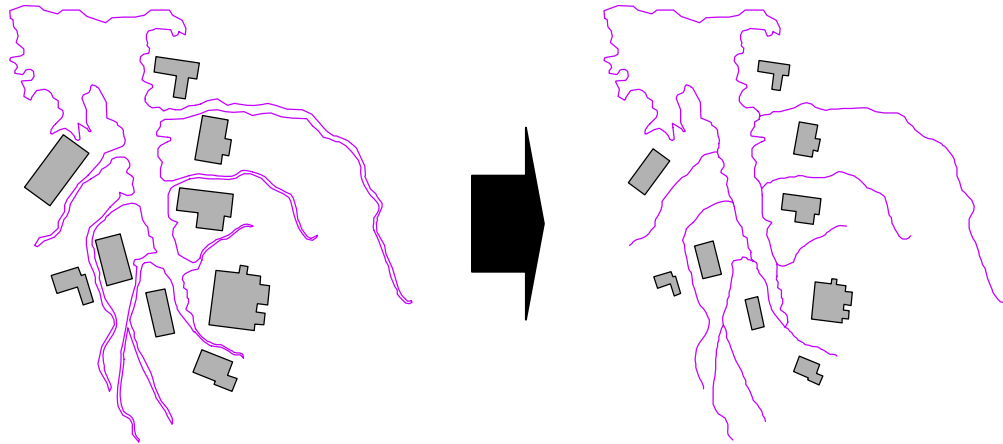


Figure 2.25 -- Collapsing using size or dimensionality reduction for generalization

Displacement – for visual clarity, particular features that have lower semantic priority but are still significant, are moved to resolve conflicts and ensure conformity with minimum separation thresholds, as seen in Figure 2.26.

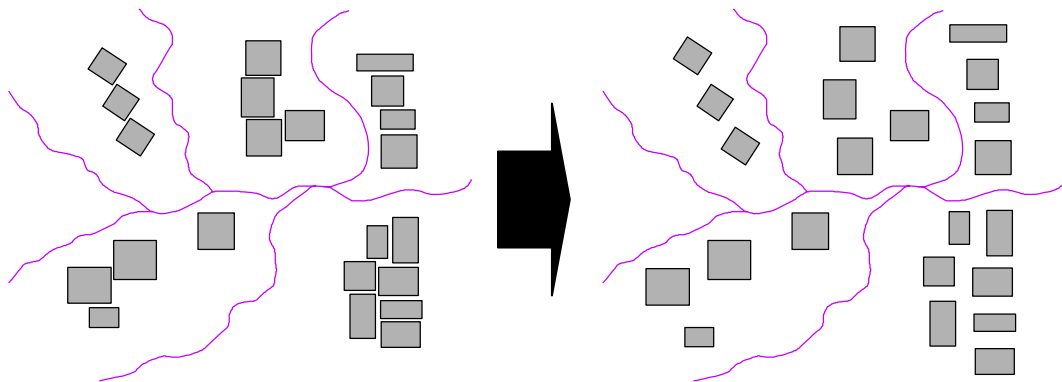


Figure 2.26 -- Translating features in conflict resolution during generalization

Typification – a reduction in the density and detail of small features while maintaining the overall distribution pattern and not compromising on the intuitive structure (such as removing several small building features in order to increase visual clarity, but maintaining a sense of the building distribution) as shown in Figure 2.27.

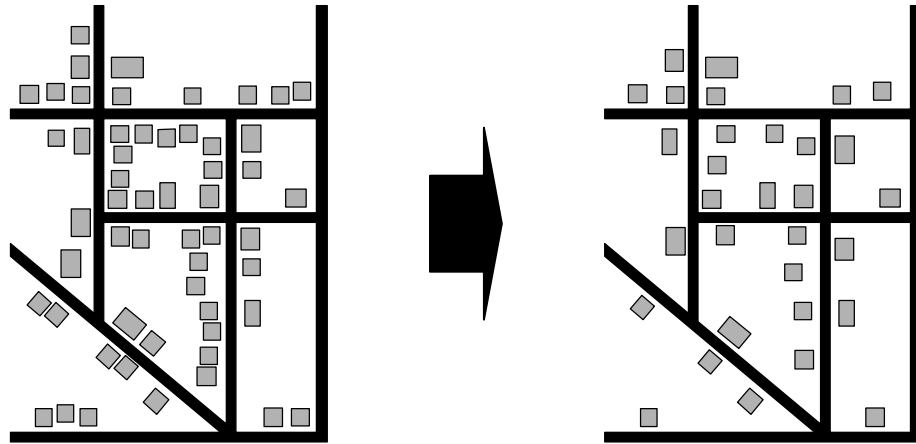


Figure 2.27 -- Removing and moving features in typification during generalization

Refinement – adjusting the geometry of features in order to improve its visual representation and to conform to reality (for instance, smoothing river features, or orthogonalizing building corners) – see Figure 2.28.

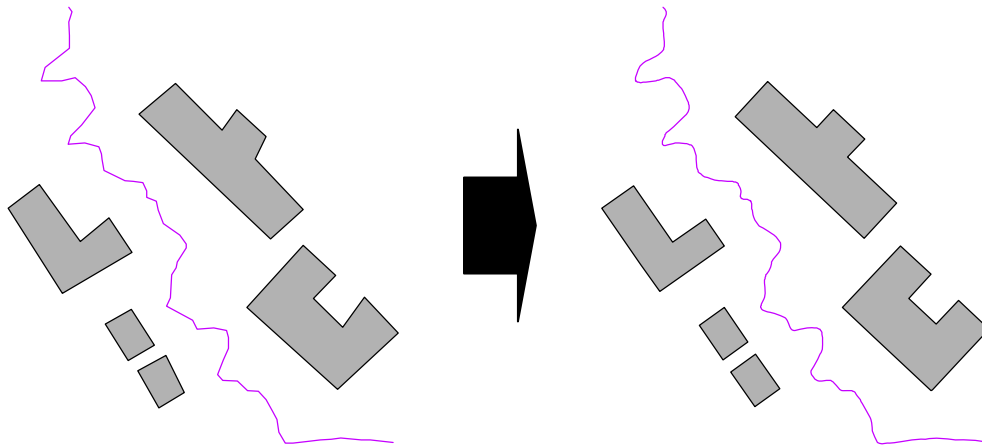


Figure 2.28 -- Refining feature geometry during generalization

Symbolization – creating new features based on lower level discrete features that share some attribute (creating “Industrial Use” polygons from lower level land use polygons that contain various levels of industrial use, such as light industrial, heavy industrial, pharmaceutical, etc.), as described by Figure 2.29.

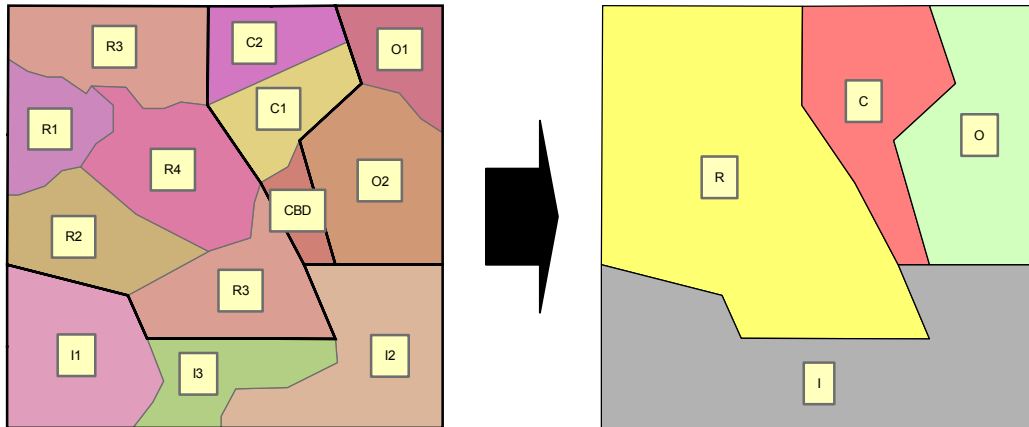


Figure 2.29 -- Symbolizing lower level into higher groups during generalization

Based on these definitions and combined with semantic and geometric rules, generalization operators may be designed for a particular application. Note that all operators may not be necessary for all applications – specific generalization operations may be iterated with increasing threshold parameters and combined into a sequence that solves the problem at hand.

2.5.8. Line Simplification

Line simplification is of special significance in this research. Since input data for shape analysis are derived from GIS, line simplification routines would be extremely useful in preprocessing the shape before features are extracted or the shape is described. Simplifying building edges would extract vertices of special significance that would serve as landmarks – these landmarks would be used to compute, measure, transform, extract and analyze the shapes and perhaps even directly applied during shape recognition based on syntactic methods.

Several algorithms have been developed since the 1960s for line simplification, drawing primarily on Attneave's (1954) identification of curvature-based vertex

significance. McMaster and Shea (1989) classify line generalizing algorithms primarily by the processing context – their first three classes are based on vertex processing, while their next two are based on extending the processing context beyond the vertex-based neighborhoods.

Simplification Using One Vertex

Here, the vertex being processed has no formal relationship with other vertices other than sequence. A crude example would be a routine that decimates every third vertex in the contour sequence

Simplification Using Vertex and Immediate Neighbors

Here, the preceding and succeeding vertices are included in a mathematical relationship with the processing vertex. An example would be a routine that compares the perpendicular distance between the vertex being processed and the chord joining the preceding and succeeding vertices with the thresholding tolerance (Chang et al. 1991).

Simplification by Processing Extended Vertex Neighborhoods

Here, the routine uses other descriptors such as angle or distance or a minimum number of points as larger contexts in mathematical relationships beyond the immediate neighboring vertices. An example is a routine that compares pairs of sequential segments that have one segment in common and decimates vertices based on Euclidean or Hausdorff metrics (Leu and Chen 1988; Boxer et al. 1993)

Simplification Using Extended Local Processing

These routines use the complexity of the line's geometry to search beyond the neighborhoods described above. An example of such a routine would apply shape

recognition techniques to detect bends, analyze their curvature characteristics in a local context, and eliminate insignificant ones. Thus, a bend that is too narrow will be widened slightly to satisfy the tolerance and the resulting line is more faithful to the original and shows better cartographic quality (Lee 2003).

Simplification by Global Routines

Here, all the vertices that constitute the line are taken in the processing framework. Examples include the Douglas-Peucker (1973) line generalization algorithm, illustrated in Figure 2.30, or the Chaikin (1974) line smoothing algorithm.

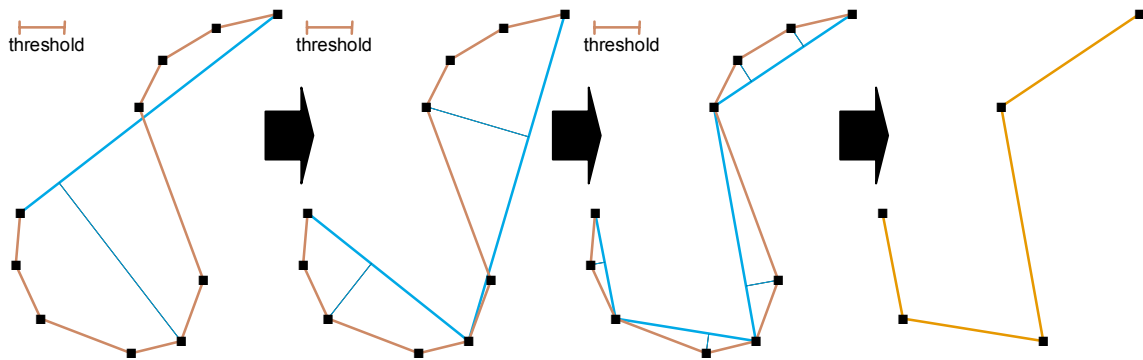


Figure 2.30 -- The Douglas-Peucker algorithm for line simplification

Perhaps the most common algorithm applied in built-in GIS simplification routines is the Douglas-Peucker algorithm. In the Douglas-Peucker algorithm, a temporary line is constructed joining the first and last points of the original polyline. The vertex that is furthest away from this temporary line is added. The distance of each vertex from the modified line is recomputed, and the farthest vertex is added to the temporary line. The process is repeated till the distance of the vertex farthest away is smaller than the thresholding tolerance – at this limit, the original line's geometry is replaced by the temporary line.

2.5.9. Evaluation of Generalization and Simplification Algorithms

Several researchers (McMaster and Shea 1988; Visvalingam and Whyatt 1990) have systematically evaluated the existing line generalization algorithms and have consistently enhanced the performance of these routines. Ruas and Plazanet (1996) have developed a routine based on polygonal approximation by evaluating curvature functions, while Visvalingam and Whyatt (1990) have modified the Douglas-Peucker algorithm and base their simplification on decimation by evaluating effective areas – points with the least areal displacement from the current part-simplified line are iteratively dropped. They chose area because line morphology becomes significant only when the size of the feature becomes larger than a perceptible threshold. Gribov and Bodansky (2004) include noise filtering in their piecewise linear approximation approach, by decomposing the source polyline into optimal segment clusters based on the squared error of approximation, and then replacing each cluster with a straight line segment. Researchers and practitioners are also evaluating generalization tools in software applications based on quantitative measures of computational efficiency, initial assumptions and assessments of results (Weibel and Jones 1998; Skopeliti and Tsoulos 2001).

While generalization routines work reasonably well, there are numerous instances where the performance is less than effective. This occurs particularly in the case of orthogonal segments that deviate more than the thresholding tolerance in one dimension, but not in the other. This shortcoming is difficult in the context of building footprint simplification, where edge segments are usually orthogonal and deviate considerably from the threshold tolerance. Figure 2.31 illustrates the deficiency in the Douglas-Peucker algorithm for building simplification.

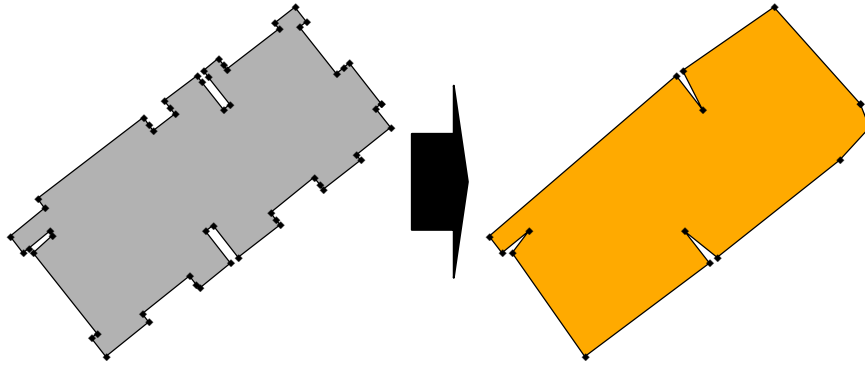


Figure 2.31 -- Limitations of the Douglas-Peucker algorithm for orthogonal edges

Thus, line generalization tools are not complete and perfect solutions (Limeng and Lixin 2001; Kazemi 2003) and may require some level of preprocessing for automated applications, and often, manual intervention. The result is often dependent on the geometry of the input building footprint polygons (that can vary considerably based on the source acquisition methodology) and additionally, after simplification, in many instances, topological errors occur that require manual corrections again (Kazemi et al. 2001).

2.6. Building Valuation

Quantifying economic losses from natural hazards is a vital element in evaluating risk, assessing the appropriateness of mitigation planning alternatives, estimating the efficient level of disaster assistance and informing the relevant stakeholders of their potential liability. Several recent research efforts have emphasized both the necessity and appropriateness of the various methodologies associated with hazard economic loss estimation (Chang 1998; Shinozuka et al. 1998; National Research Council 1999; Chang 2001).

In general, most loss estimation studies first estimate damage to the physical inventory, and then translate these into economic losses. Economic losses have been

typically differentiated into direct and indirect effects, which are not clear-cut in nature (Rose 2004). ATC-21 (1991) clarified that direct losses are attributed to property damage, while business interruption losses tend to be indirect, but Rose (Rose and Lim 2002; Rose and Kunreuther 2004) suggests that this distinction is confusing, because both types of losses have direct and indirect components. However, the National Research Council (1999) makes a worthwhile distinction and defines direct losses as those that arise from the premises housing the business being damaged directly by the hazard, while business interruption losses stemming from utility or infrastructure interruption are termed secondary direct losses (Rose et al. 1997). All other losses based on linkages with other business entities are termed indirect losses.

This dissertation limits the discussion to damage caused by the hazard to the building itself, and more specifically, to the replacement cost of the building. There have been several studies regarding the cost of construction, and specifically the cost of seismic construction upgrading (FEMA 1992a, 1992b, 1994, 1995). ATC-13 (1985) also performed a significant study for buildings in California. These studies used a regional perspective and looked for central tendencies in the building inventories, expressing costs in dollars per square foot for several occupancy classes. In addition, these studies were limited to lateral forces on buildings. This is somewhat surprising, considering that the San Fernando earthquake of 1971 made it apparent that damage to nonstructural components not only resulted in major economic loss, but also posed real threats to life safety. Nonstructural damage accounted for nearly 50% of the total loss of about \$18.5 billion in the Northridge earthquake (Kircher 2003). Since then, there have been several other studies evaluating the direct components of damage in buildings. Nonstructural components and building contents represent a significant part of the overall value of

most buildings, and a large component of direct losses to buildings in earthquake events may be attributed to nonstructural aspects of the building (Whittaker and Soong 2003).

There are several projects being conducted at the various earthquake research centers in the US, but these efforts deal with individual aspects of building components. However, there have been several studies that attempt to define the various elements of the building (Porter et al. 2001; Porter 2005; ATC-69 2008). Before proceeding to the components of a building sensitive to earthquake stresses, it would be worthwhile to investigate how existing models of loss estimation estimate replacement costs of buildings.

2.6.1. Replacement Costs of Buildings

Simply defined, the replacement cost of a building is the amount in dollars to reconstruct the building today at the same site for the same functionality using the same materials, and ensuring that the building follows the current building code. An important aspect of this working definition of replacement cost is building use. In other words, the specific occupancy of the building is an important driver of replacement costs.

Let us begin the discussion by examining replacement cost models in typical loss estimation applications. For regional loss estimation, HAZUS MH MR-3 bases building replacement costs for each specific occupancy class (see Table 1.2 for a list of specific occupancies in HAZUS MH MR-3) using industry-standard cost-estimation models published in the Means Square Foot Costs (R. S. Means 2008). For each specific occupancy class, HAZUS establishes a default model, using averages of the square foot costs for various alternatives of the exterior wall construction (FEMA - DHS 2007). Table 2.4 shows an extract of the default listing of Means building models and

associated 2002 replacement costs in dollars per square foot, for the various residential specific occupancy classes, excluding single-family residential.

Table 2.4 -- HAZUS MH MR-3 2002 Residential Replacement costs (in \$/sq. ft.)

Use	Description	Sub-category	Means Model Description	Cost/sq. ft.
RES2	Manufactured Housing	Manufactured Housing	Manufactured Housing	\$ 30.90
RES3	Multi-family Dwelling (small)	Duplex	SFR Avg 2 St., MF adj, 3000 SF	\$ 30.90
		Triplex/Quads	SFR Avg 2 St., MF adj, 3000 SF	\$ 67.24
	Multi Family Dwelling (medium)	5-9 units	Apt, 1-3 st, 8,000 SF (M.010)	\$ 73.08
		10-19 units	Apt., 1-3 st., 12,000 SF (M.010)	\$ 125.63
	Multi Family Dwelling (large)	20-49 units	Apt., 4-7 st., 40,000 SF (M.020)	\$ 112.73
		50+ units	Apt., 4-7 st., 60,000 SF (M.020)	\$ 108.86
		High-rise Apartment	Apt., 8-24 st., 145,000 SF (M.030)	\$ 106.13
RES4	Temp. Lodging	Hotel (medium)	Hotel, 4-7 st., 135,000 SF (M.350)	\$ 111.69
		Hotel (large)	Hotel, 8-24 st., 450,000 SF (M.360)	\$ 104.63
		Motel (small)	Motel, 1 st., 8,000 SF (M.420)	\$ 93.47
		Motel (medium)	Motel, 2-3 st., 49,000 SF (M.430)	\$ 94.13
RES5	Institutional Dormitory	Dorm (small)	Frat House, 2 st., 10,000 SF (M.240)	\$ 110.03
		Dorm (medium)	College Dorm, 2-3 st, 25,000 SF (M.130)	\$ 118.82
		Dorm (large)	College Dorm, 4-8 st, 85,000 SF (M.140)	\$ 113.31
RES6	Nursing Home	Nursing home	Nursing Home, 2 st., 25,000 SF (M.450)	\$ 99.50

Similarly, the application contains default Means model types and square footage costs for all specific occupancy categories, and several alternative models of each occupancy. None of the non-residential specific occupancy categories have basements included in the default costs. For single-family residential structures, again, based on Means square foot costs, HAZUS breaks up the inventory into four classes of single family residences, including Economy, Average, Custom and Luxury, sub-classified by height (number of stories), presence of a finished or unfinished basement and adjusted for car garages.

The replacement value for the region's buildings is based on the derived counts for each specific occupancy, described in Section 1.4.3. While the intentions are certainly merit-worthy, the various sub-categories or alternative Means models are not used in the application. Further, the replacement costs make no allowance for the

structure type of the building – there is a considerable difference in the per square foot costs for buildings made of wood versus concrete or steel. In addition, the Means model costs describe typical buildings of specified area, and in reality, buildings in any occupancy category exhibit considerable variance in the square footage, and may be substantially less or more than the square foot range specified in Means.

2.6.2.1. Structural and Nonstructural Building Components

In most developed countries, seismic safety codes have influenced the design and construction of buildings to the extent of significantly mitigating catastrophic structural collapse. However, the structural system of a designed building typically represents less than a quarter of the total replacement costs. Admittedly, this fraction may be different for a specific subset of buildings, but in general, nonstructural building components and building contents hold a significant portion of the total cost of construction. In addition, hazard-related damage to nonstructural components can potentially threaten life safety (Whittaker and Soong 2003). The significance of nonstructural building elements has been facing greater scrutiny, particularly in earthquake engineering research, and the variety and complexity of nonstructural and content elements will continue to dominate the challenges to the overall seismic performance of buildings and inform mitigation planning efforts (ATC-69 2008).

Every earthquake event has had some impact on nonstructural building components and building contents. Consider the Modified Mercalli Index (MMI), first proposed in 1931 – earthquake intensity levels are almost completely defined in terms of the behavior of nonstructural or content elements (Richter 1957). However, a systematic analysis of the performance of nonstructural building components has been problematic owing to the lack of data. To date, while descriptions of every earthquake event include some documentation of nonstructural damage, there is little systematic information on

how the failure of or damage to nonstructural elements and contents pose threats to life safety or cause direct damage/business interruption losses. ATC-69 (2008) suggests that the main reason for this situation is that the division of direct damage (as structural and nonstructural) is not consistent with owners or underwriters – building owners bear the responsibility for damage to the structural and nonstructural components, while building users or tenants are responsible for the inventory and contents. In addition, while building ownership data (and related insurance claims for earthquake-related damage) are relatively easy to find, a building may have several tenants who face differing amounts of content damage or interruption losses, the data for which is likely to be dispersed. Even where damage has occurred, research teams collecting damage data tend to focus on the dramatic aspects of structural damage first, then on obvious nonstructural damage like broken sprinkler systems or collapsed ceilings/interior partitions, etc. (all of which photograph well!) and little or no attention is directed to situations of minor or even functional nonstructural elements (even successes can teach us something!). Content-related damage is often cleaned up before studies document them, and repair of nonstructural elements have an extremely long time frame, depending on the criticality of the nonstructural element to occupancy. Finally, collecting information on the performance of nonstructural elements is time-consuming and resource-intensive (Reitherman 1998). Coupled with the added problems of lack of standardization in the collection and presentation of data, nonstructural damage research has proven to be almost intractable.

2.6.2.2. Earthquake-related Damage to Nonstructural Components and Contents

Damage to architectural, mechanical, electrical and plumbing and water supply systems has occurred in the past. Direct damage to nonstructural elements has been exacerbated by exposure to water, forceful running water, chemicals or other hazardous

substances. Nonstructural elements or construction assemblies that have consistently been damaged in earthquake events include:

- *Architectural elements* – cladding, glazing, external non-load-bearing walls, parapets, chimneys, partitions, false ceilings, etc.
- *HVAC, Electrical, Water supply, Fire protection, Plumbing and Conveyance* – sprinkler systems, pipes, piping connections, ductwork, lighting, escalators and elevators, tanks, conduits and trays, equipment, etc.
- *Contents* – shelves, cabinets, book cases, furniture, appliances, storage racks, equipment, computers and servers, etc.

A number of data collection efforts have been executed primarily by insurance companies and underwriters, but these datasets tend to be confidential and proprietary, and not available for research (Porter 2002).

The separation of damage into structural and nonstructural components is important because the systems behave differently under earthquake stresses. The general technique for loss estimation is to develop mean repair cost ratios for discrete damage state probabilities (derived from earthshaking levels and fragility curves) and estimate mean total repair cost by component category. The approach is fairly deterministic and several different building components are grouped under common fragility functions – all electrical, mechanical and plumbing elements are represented by just one or two fragility functions (Porter et al. 2001). In addition, the cumulative effect of different components is never considered – for instance, if a false floor fails under a moderate level of shaking, and several pieces of heavy equipment (that will tip over only under extreme earth shaking) rest on it, the cumulative effect of the failure of the false floor would be that the pieces of heavy equipment would also tip over.



Porter (Porter et al. 2001; Porter 2005) suggests a taxonomy of components or assemblies of components that include structural and nonstructural elements, installation conditions, detailed inventories of equipments and contents and architectural and service assemblies. He broadly classifies structural elements as those that are part of the building's vertical- or lateral-force-resisting system and nonstructural components as those that are attached or rest on the structural system, but are not part of any force-resisting system. Nonstructural elements may then be grouped based on a set of rules that allow the development of representative fragility functions for that class, with emphasis on potential repair costs, life safety or interruption of use. A consistent taxonomy would then enable an effective evaluation of the building's seismic performance and enable quantification of the potential benefits of retrofits or design proposals.

Thus, a detailed and careful classification of structural and nonstructural elements of a building would enable loss estimation routines to apply specific damage functions to component category groups that behave similarly, just as the GBS is classified by occupancy, height, structure type and design level into groups that behave similarly under hazard stresses. Grouping several disparate elements that have different damageability functions will introduce a large amount of uncertainty in the loss estimation process. The International Building Code (ICC 2000) and the ASCE 7-05 (ACSE 2005) Minimum Design Loads for Buildings and other Structures both specify a series of seismic design requirements for architectural and mechanical and electrical components respectively. The International Building Code specifies components such as interior walls and partitions, braced and unbraced cantilevers, veneers, ceilings, cabinets, etc. The ASCE 7-05 includes mechanical and electrical components with conveyance equipment included under electrical and distribution systems.

Similarly, HAZUS MH MR-3 arranges common nonstructural elements of buildings as seen in Table 2.5 into simple, tractable groups and works well for regional loss estimation. However, suspended ceilings and glazing categories are conspicuously missing. Nonstructural elements are grouped as either “drift-sensitive” or “acceleration-sensitive” components. Damage to drift-sensitive components is largely a function of interstory displacement, while damage to contents and acceleration-sensitive components is influenced by floor acceleration.

Table 2.5 – HAZUS MH MR-3 division of nonstructural elements and contents

Type	Description	Drift-sensitive	Acceleration-sensitive
Architectural	Nonbearing Walls/Partitions	Primary	Secondary
	Cantilever Elements and Parapets		Secondary
	Exterior Wall Panels	Primary	Secondary
	Veneer and Finishes	Primary	
	Penthouses	Primary	
	Racks and Cabinets		Primary
	Access Floors		Primary
	Appendages and Ornaments		Primary
Mechanical and Electrical	General Mechanical (boilers, etc.)		Primary
	Manufacturing and Process Machinery		Primary
	Piping Systems	Secondary	Primary
	Storage Tanks and Spheres		Primary
	HVAC Systems (chillers, ductwork, etc.)	Secondary	Primary
	Elevators	Secondary	Primary
	Trussed Towers		Primary
	General Electrical (switchgear, ducts, etc.)	Secondary	Primary
	Lighting Fixtures		Primary
Contents	File Cabinets, Bookcases, etc.		Primary
	Office Equipment and Furnishings		Primary
	Computer/Communication Equipment		Primary
	Nonpermanent Manufacturing Equipment		Primary
	Manufacturing/Storage Inventory		Primary
	Art and other Valuable Objects		Primary

 Primary cause of Damage
 Secondary Cause of Damage

Taghavi and Miranda (2003) designed and implemented a Microsoft Access database of nonstructural elements for commercial buildings based entirely on R.S. Means' assemblies – the assemblies found in R.S. Means are industry standard and a wide variety of building industry professionals are familiar with the terms. They included a taxonomy of components, photographs, fragility functions, repair costs, repair methods and functionality to each of the component groups.

Structural and nonstructural divisions for residential buildings are harder to find. Saeki et al (2000) surveyed nearly a 1000 insurance company employees regarding property ownership and damage to ten categories of contents, as seen in Table 2.6. They found that the most commonly damaged items were tableware, while heaters and coolers remained relatively undamaged.

Table 2.6 -- Taxonomy of household contents (Saeki et al 2000)

Goods		Household Property		
Type	Code	Damage type	Description	Example
Durable Possessions	A	overturning	large, self-standing furniture for storage	chests, bookshelves, cupboards
	B	overturning	household electrical appliances	refrigerators, washing machines
	C	falling, toppling over	household electrical appliances	microwave ovens
	D	falling, toppling over	entertainment equipment	audiovisual, computers, telecommunications equipment
	E	crushing	floor-standing furniture, tables and chairs	dining tables, chairs, living room furniture, stoves
	F	crushing, overturning	heaters and coolers	air conditioners and heaters
Non-durable Possessions	G	crushing	indoor and miscellaneous items	curtains, sliding doors/screens, medical equipment, shoes, carpets
	H	falling, toppling over	tableware	tableware -- knives, spoons and forks
	I	falling, toppling over	home entertainment items	clocks, cameras, lighting fixtures, records, CDs, toys
	J	damaged, spills	clothes, bed linen, bed clothes	clothes and bed linen

Several other taxonomies have been proposed – the interested user is asked to see Porter (2005) for a comprehensive literature survey of these taxonomies. Porter’s taxonomies are very detailed and suited to analyses of single buildings. The R.S. Means or the Taghavi and Miranda typologies offer the best potential for application in regional loss estimation modeling.

2.6.2.3. Content Value of Buildings

The literature for determining the content value of buildings by occupancy or indeed by any other classification is non-existent. Interviews with personnel in valuation companies revealed that contents of buildings were individually surveyed and inventoried, and aggregated on a building-by-building basis for the purposes of insurance and/or portfolio management. The accepted methodology was straightforward in application, and included depreciation using standard methods. However, none of the valuation companies or insurance agencies was willing to share their data, citing confidentiality issues or that the data was proprietary.

The only other source was HAZUS MR-3, in which content value was expressed as a percentage of replacement cost by specific occupancy. The technical documents made no mention of the source for the default specifications. Table 2.7 shows the content value expressed as a percentage of replacement costs by specific occupancy.

Table 2.7 -- Content value as percentage Replacement cost (HAZUS MH MR-3)

Occupancy Type	Content Value	Occupancy Type	Content Value
All Residential Units	50%	Food and Entertainment	100%
Retail Trade	100%	Parking Garages	50%
Wholesale Trade	100%	All Industrial	150%
Personal/Repair Services	100%	Religious	100%
Commercial Offices	100%	Emergency Response	150%
Banks	100%	Schools	100%
Health care related	150%	Colleges and Universities	150%

2.7. Conclusion

This literature review is fairly comprehensive and serves two purposes. First, I have attempted to describe relevant information about the background and applications of advanced techniques that the average audience interested in loss estimation and risk assessment modeling would find useful or illuminating. Second, the literature review is also aimed at identifying methods and innovative approaches that would inform the methodology section of the dissertation. In fact, several approaches advocated by the literature were implemented during the course of this research, including a decision-rule or knowledge-based classification model of structure type, and a shape recognition application based on invariant shape representations and statistical moment functions. In both cases, performances belied expectations, and the shape recognition application performance was particularly unsatisfactory, but the process of actually implementing previous research approaches clearly demonstrated the pros and cons of particular strategies.

Chapter 3 . METHODOLOGY

Based in part on the literature review, this chapter describes the design of methods for determining the structure type of buildings, classification of building footprints by shape and estimation of building costs (replacement, acceleration- and drift-sensitive costs, content value). It is quite possible to design the methods to estimate the various parameters required for assessing risk to buildings, but validity concerns prompted the necessity for demonstrating the methods in the context of the real world. Consequently, the MAEC decided to showcase all its projects and demonstrate proof-of-concept for Shelby County, Tennessee.

Shelby County is located in the western-most part of Tennessee in Mid-America, comprising mainly of the City of Memphis. The Mid-America region or the New Madrid Zone stretches to the southwest from New Madrid, Missouri, and is characterized by the New Madrid fault line. The zone covers parts of Missouri, Arkansas, Illinois, Kentucky and Tennessee, and is seismically active with the potential to produce significant earthquakes. Between 1811 and 1812, several earthquakes with estimated magnitudes greater than 7 rocked the area of the Mississippi valley. Since then however, there have been several insignificant earthquakes, and building practices in the region tend to reflect this lack of memory. The scientific community believes that we are long overdue for a seismically significant event, with a 90% chance of a 6.0 or greater event occurring by the year 2040. An earthquake of 7 or greater therefore has great potential to cause significant damage to life and property, which may be reduced to a large extent with research, public awareness and mitigation planning. The study area, also called the MTB, is therefore an apt choice, with a large urban population.

This chapter describes the existing primary data available for the MTB and its extraction for the inventory modeling exercise. The following section describes the survey of buildings in Memphis for generating a calibration and validation sample. The design of the specific methodologies for classifying buildings by structure and shape and estimating the value of the building inventory forms the remainder of the chapter. Figure 3.1 describes the overall methodology through a schematic process, using both primary and derived data.

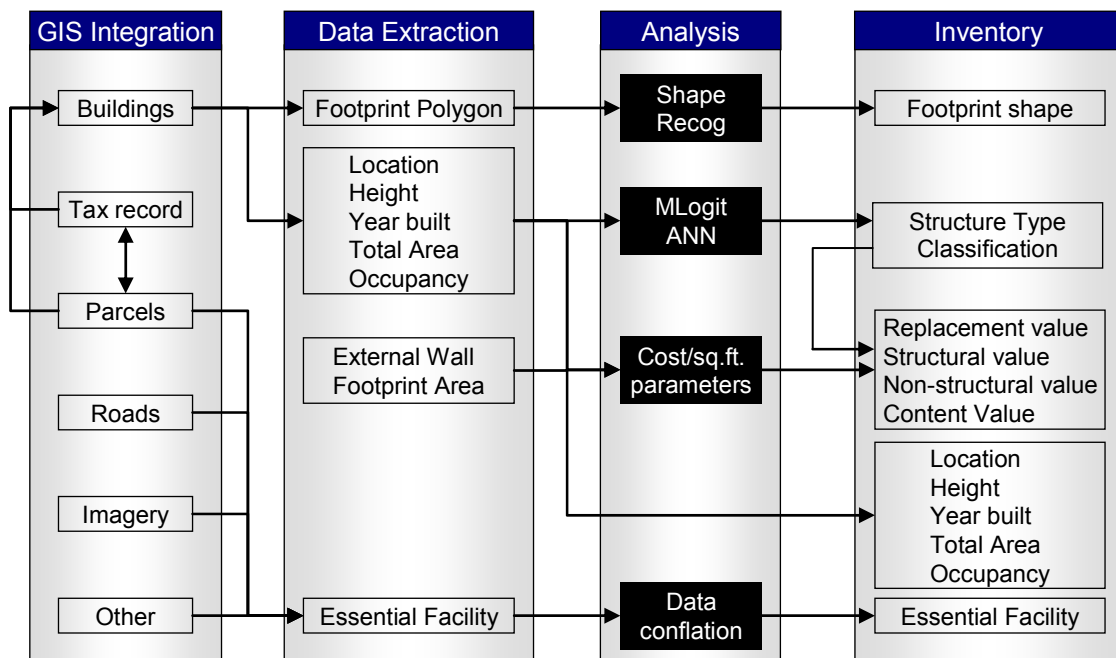


Figure 3.1 -- Research methodology described by a schematic process

Primary data, in the form of tax records from the Shelby County Tax Assessor’s database, roads, parcels, aerial imagery, and corollary datasets for schools and other essential facilities are first analyzed and integrated within a GIS. Specific variables are extracted and analyzed to produce the necessary attributes of the building inventory outlined in the scope of the dissertation. The analytical module consists of four separate components that will be used in order to :

- classify buildings by their footprint configuration using shape recognition algorithms designed and written in the GIS environment
- derive the structure type of buildings using multinomial logistic regression or ANNs
- estimate the replacement value, the structural, acceleration- and drift-sensitive nonstructural components of the replacement value and the content value using parameterized curves generated by curve-fitting routines
- add missing essential facilities (schools, fire stations and police stations) and churches manually to the spatial building inventory database by geocoding and inspection of aerial images

3.1. Tax Assessor’s Data for Shelby County

The Shelby County Tax Assessor’s database (henceforth Tax records) comprised of 20 separate tables, as seen in Table 3.1. Most of the tables contained information pertaining to the parcel, identified by a unique identifier, or to the improvement(s) in the parcel. Improvements made to the parcel, as captured in the Tax Records, represent single buildings or multiple identical buildings. While some of the documentation for the relational database was adequate, there were significant gaps in the descriptions and specific relationships between particular fields across the tables. For instance, the field “UNITS” appeared in two different tables and was documented as “Number of units” – the numbers however, did not tally across tables for the same improvement. *Users are cautioned that reconstructing or re-engineering tables in order to create new synthetic tables is not a trivial task and requires expertise on handling relational database management system concepts.*

Table 3.1 -- Shelby County Tax Assessor's database tables

S. No.	Table	Basis	Description
1	ADDN	parcel	Contains improvements and additions made to the property
2	AEDIT	none	Contains all the edit codes in the database and their description
3	AGAPPL	parcel	Agricultural application number table or farmland table
4	AGLAND	parcel	Contains all the agricultural land description
5	ASMT	parcel	Contains Appraisal and Assessment Value information
6	COMAPT	improvement	Contains commercial apartments data
7	COMDAT	improvement	Contains all commercial building data
8	COMFEAT	improvement	Commercial features information
9	COMINTEXT	improvement	Contains commercial interior exterior information
10	COMNT	parcel	Contains comment number, comment code and description
11	DWELDAT	improvement	Contains dwelling information
12	ENTER	parcel	Contains survey information
13	IEPRCL	parcel	Parcel information for income valuation/modeling
14	LAND	parcel	Contains land information
15	LEGDAT	parcel	Contains legal data information
16	OBY	parcel	Contains other building and yard information
17	OWNDAT	parcel	Contains owner information
18	PARDAT	parcel	Contains Parcel Level Information
19	PERMIT	improvement	Contains permits information
20	SALES	parcel	Contains sales information

3.1.1. Generating Unique Identifiers for Tax Records

There were no unique identifiers for improvements or sections of improvements and unique identifiers for improvements were generated by concatenating the parcel identifier with the numerical sequence number of the improvement. Since the improvement records could represent one or more buildings, they could not be used directly to identify buildings. All improvements that represented multiple buildings were cloned by the number of buildings that each improvement represented, with a sequence number for each clone of the original improvement record. If an improvement record represented one building, the sequence number would be “1”. If an improvement record represented 3 buildings, the table would have the original improvement and 2 clones, with sequence numbers of “1”, “2” and “3”. This process resulted in one improvement record for every building. Defining a unique building identifier was trivial after the cloning

operation and generated by concatenating the improvement identifier and the sequence number. Thus, the parcel identifier would serve as a primary key for the parcel database, and the building identifier would serve as a primary key for the building inventory database. Thus, each building in the building database may be identified uniquely and further, the building record also identifies the land parcel where the building is sited. Multiple buildings that are located in the same parcel have the same parcel identifier value. Table 3.2, extracted from the building inventory database describes the identification pattern using three land parcels.

Table 3.2 -- Parcel, Improvement and Building identifiers

S. No.	Parcel	Improvement	Building
1	001001 00025	001001 00025_1	001001 00025_1_1
2	001001 00026	001001 00026_1	001001 00026_1_1
	001001 00026	001001 00026_1	001001 00026_1_2
	001001 00026	001001 00026_2	001001 00026_2_1
	001001 00026	001001 00026_3	001001 00026_3_1
	001001 00026	001001 00026_3	001001 00026_3_2
	001001 00026	001001 00026_3	001001 00026_3_3
	001001 00026	001001 00026_3	001001 00026_3_3
3	001057 00002	001057 00002_1	001057 00002_1_1
	001057 00002	001057 00002_2	001057 00002_2_1

The first parcel, identified by “001001 00025” has only one improvement and one building identified by “001001 00025_1” and “001001 00025_1_1” respectively. The second parcel identified by “001001 00026” has three improvements specified by “001001 00026_1”, “001001 00026_2” and “001001 00026_3”. The first improvement consists of two identical buildings, each uniquely identified by “001001 00026_1_1” and “001001 00026_1_2” respectively. The second improvement consists of one building identified as “001001 00026_2_1”. The third improvement consists of three identical buildings, each uniquely identified by “001001 00026_3_1”, “001001 00026_3_2” and “001001 00026_3_3” respectively. This parcel therefore has a total of six buildings.

The third parcel, “001057 00002”, has two improvements consisting of one building each, specified by “001057 00002_1_1” and “001057 00002_2_1” respectively.

The raw entities representing the information from the Tax Assessor are compiled into a complex relational database structure. Understanding the entities and their relationships is a vital component to the creation of synthetic tables. Relational databases are designed and implemented for a specific audience and a specific purpose – in the Shelby County Tax Assessor’s case, the system was designed to keep track of taxable improvements, and not buildings, requiring the extraction and conversion of specific items into synthetic tables related to seismic risk assessment. While the Shelby County Tax Assessor’s data records had adequate documentation, relationships between and among items were not clearly specified and had to be reconstructed through trial and error. Users are cautioned that extracting information pertinent to earthquake risk assessment and damage modeling from the relational database is relatively complicated, with potential for large errors that could propagate throughout the models -- this process requires care and expertise, first to understand the relationships and then to manipulate the data through join, summarizing and extraction operations.

3.1.2. Single-family Residential Building Extraction

The extraction of single-family residential buildings process was relatively straightforward, since all the relevant information was contained in the DWELDAT table. The Tax Records contained multiple instances of the parcel identifiers in situations where several single-family residential units were sited in one parcel. These specific records, where there were multiple single-family residential units, were identified for parcel counts greater than 1 and cloned, using the cloning process described in Section 3.1.1. The relevant variables extracted or generated included the parcel, improvement and building identifiers, the number of stories, the year of construction, the type of

basement, the total and ground floor areas in square feet, the major use of the parcel, the exterior wall type, the overall condition of the building and the appraised value of the building. Based on the square footage and the overall condition of the building, single-family residential units were classified as “Economy”, “Average”, “Custom” and “Luxury”. As expected, “Economy” and “Average” construction types dominate the single-family stock.

3.1.3. Multi-family and Commercial/Industrial Building Extraction

Extracting commercial and industrial buildings was much more complicated. The documentation provided with the Tax Records was inadequate and the tables had to be thoroughly analyzed and re-engineered to understand the linkages between the various files. Commercial and Industrial building data was distributed between the COMDAT and COMINTEXT tables. Multi-family residential (apartments and condominiums) data was distributed among the COMAPT, COMDAT and COMINTEXT tables. In addition, as mentioned earlier, the Tax Records in these tables consisted of improvements or sections of improvements and each improvement could represent one or more buildings. In particular, the specific use of the building, the number of dwelling units, the number of stories and the square footage information was contained in these tables. The process of extraction is best explained using the example of 1 land parcel identified by containing 5 separate improvements and 10 buildings.

Table 3.3 shows the COMAPT records for parcel “001001 00026”. The “Count” field specifies the number of dwelling units in identical buildings, while the “Units” field specifies the total number of dwelling units in that record. The Improvement field contains the Improvement Identifier, so this table specifies that there are 4 improvements in that parcel (the number of unique improvement identifiers). The

column named "Impr Units" highlighted in yellow is a computed field containing the total number of dwelling units in that parcel and is used for error-checking and quality control.

Table 3.3 -- COMAPT extract for Parcel "001001 00026"

Parcel	ImpID	Improvement	Count	Units	Year	Impr Units
001001 00026	1	001001 00026_1	6	12	1991	24
001001 00026	1	001001 00026_1	6	12	1991	
001001 00026	2	001001 00026_2	12	12	1991	18
001001 00026	2	001001 00026_2	6	6	1991	
001001 00026	3	001001 00026_3	12	48	1991	48
001001 00026	4	001001 00026_4	12	24	1991	24

Table 3.4 shows the original COMDAT records for the same parcel, organized by the improvement identifier, with a total of 5 improvements, while Table 3.5 shows the COMDAT records after the cloning operation. The "Units" field in Table 3.4 specifies the total number of dwelling units in each example building specified by that improvement record. The "NumIdent" field specifies the number of identical buildings specified in that improvement. Thus, for Improvement = "001001 00026_3", there are 4 identical buildings. The "Sum Area" field specifies the total area of all the buildings represented by that improvement record. The "Impr Units" column (with the yellow highlight) is a computed column derived by multiplying the "Units" value and the "NumIdent" value and contains the total number of dwelling units in that improvement.

Table 3.4 -- COMDAT extract for Parcel "001001 00026"

Parcel	ImpID	Improvement	Units	NumIdent	Sum Area	Impr Units
001001 00026	1	001001 00026_1	12	2	21140	24
001001 00026	2	001001 00026_2	18	1	16082	18
001001 00026	3	001001 00026_3	12	4	29648	48
001001 00026	4	001001 00026_4	12	2	25272	24
001001 00026	5	001001 00026_5	0	1	1803	0

Table 3.5 -- Cloned COMDAT extract for Parcel "001001 00026"

Parcel	ImpID	Improvement	Units	NumIdent	Sum Area	Bldg Units	S. No.
001001 00026	1	001001 00026_1	12	2	10570	12	1
001001 00026	1	001001 00026_1	12	2	10570	12	2
001001 00026	2	001001 00026_2	18	1	16082	18	1
001001 00026	3	001001 00026_3	12	4	7412	12	1
001001 00026	3	001001 00026_3	12	4	7412	12	2
001001 00026	3	001001 00026_3	12	4	7412	12	3
001001 00026	3	001001 00026_3	12	4	7412	12	4
001001 00026	4	001001 00026_4	12	2	12636	12	1
001001 00026	4	001001 00026_4	12	2	12636	12	2
001001 00026	5	001001 00026_5	0	1	1803	0	1

Note that in Table 3.5, the number of records has increased to 10, corresponding to the total of "NumIdent" for that parcel. Note also Improvement = "001001 00026_3" has been cloned and there are now 4 instances of that same improvement identifier. "Sum Area" has been reduced by dividing the original "Sum Area" by the number of identical buildings. The fields highlighted in yellow are computed fields. "Bldg Units" is computed by dividing "Tot Units" by the number of identical buildings. The "S. No" field, part of the cloning process, generates a sequence number for each cloned record and is concatenated with the improvement identifier in order to generate unique building identifiers. The cloned table thus reflects each record as a specific building with the correct square footage and number of dwelling units.

Table 3.6 shows the COMINTEXT records for the same parcel, and consists of improvement sections – this table contained a lot of valuable information related to the improvement, such as Area, Occupancy, Fire rating, Number of stories, External wall and other such details. The "Area" field specifies the square footage of the improvement section, while the "Total Area" field specifies the total square footage of all buildings in that improvement section. The field "Use" details the specific use of the building. Note that the last improvement is a multi-use office and clubhouse.

Table 3.6 -- COMINTEXT extract for Parcel "001001 00026"

Parcel	ImplD	Improvement	Area	Use	Total Area	Impr Area
001001 00026	1	001001 00026_1	1525	Apartment	3050	10570
001001 00026	1	001001 00026_1	984	Apartment	984	
001001 00026	1	001001 00026_1	997	Apartment	1994	
001001 00026	1	001001 00026_1	1590	Apartment	1590	
001001 00026	1	001001 00026_1	984	Apartment	1968	
001001 00026	1	001001 00026_1	984	Apartment	984	
001001 00026	2	001001 00026_2	1027	Apartment	2054	16082
001001 00026	2	001001 00026_2	1804	Apartment	1804	
001001 00026	2	001001 00026_2	1105	Apartment	1105	
001001 00026	2	001001 00026_2	1733	Apartment	1733	
001001 00026	2	001001 00026_2	1027	Apartment	2054	
001001 00026	2	001001 00026_2	1027	Apartment	2054	
001001 00026	2	001001 00026_2	1014	Apartment	1014	
001001 00026	2	001001 00026_2	1105	Apartment	1105	
001001 00026	2	001001 00026_2	1105	Apartment	1105	
001001 00026	2	001001 00026_2	1027	Apartment	2054	
001001 00026	3	001001 00026_3	1213	Apartment	2426	7412
001001 00026	3	001001 00026_3	1280	Apartment	1280	
001001 00026	3	001001 00026_3	1280	Apartment	1280	
001001 00026	3	001001 00026_3	1213	Apartment	2426	
001001 00026	4	001001 00026_4	1105	Apartment	1105	12636
001001 00026	4	001001 00026_4	1027	Apartment	2054	
001001 00026	4	001001 00026_4	1105	Apartment	1105	
001001 00026	4	001001 00026_4	1105	Apartment	1105	
001001 00026	4	001001 00026_4	1027	Apartment	2054	
001001 00026	4	001001 00026_4	1105	Apartment	1105	
001001 00026	4	001001 00026_4	1027	Apartment	2054	
001001 00026	4	001001 00026_4	1027	Apartment	2054	
001001 00026	5	001001 00026_5	1803	Clubhouse	1803	1803

The column "Impr Area" (highlighted in yellow) is a computed field generated by summing the total area for all improvement sections for a single building represented by that improvement. This tallies with the "Sum Area" field of the cloned COMAPT records depicted in Table 3.5. The COMINTEXT table also had the number of stories specified for each improvement section, the exterior wall type and the fire rating of the structure (Fireproof, Fire Resistant, Pre-engineered Steel and Wood joists). The number of stories was alphanumeric, and included basement and penthouse codes in addition to the numeric stories. Based on area thresholds of 2,500 square feet, basement and

penthouse floors were included in the total number of stories (the area threshold would eliminate service areas like elevator machine rooms in the penthouse floors or utility rooms in the basement). Accordingly, the extracted variables included the parcel, improvement and building identifiers, the major use of the parcel, the specific use of the building, the number of stories (above and below ground), year built, total square footage, basement type and basement square footage, number of dwelling units and appraised value. Appraised value for the building was computed by dividing the total building appraised value in the parcel by the square footage of each building.

3.1.4. Imputation of Missing Data and Data Refinement

Several buildings in the Tax records lacked complete information – missing information included combinations of square footage, number of dwelling units, stories, year built, etc. In addition, we acquired information on several churches, fire and police stations and schools from a variety of other sources that lacked similar information. In the interests of completing the database and not discarding otherwise useful information, the missing information was imputed, and the record marked as “imputed”.

Information was acquired from GDT, Inc. for churches and schools that were missing from the Tax Records and physically moved to the correct parcel using aerial images (churches were often identifiable through shadows of steeples and domes, while school buildings had distinctive footprints and often had baseball or athletic tracks in the vicinity). Subsamples of each were digitized and computed average area measures for two size categories, small and large churches (6,000 sq. ft and 9,000 sq. ft.), and elementary and high schools (25,000 sq. ft. and 75,000 sq. ft.). A similar approach was taken for police- and fire-stations, except that the approximate area for each of the buildings was calculated and recorded. For year of construction, improvements in the vicinity were analyzed, particularly if the imputed buildings were part of a multiple use

parcel. Structure type was imputed as the most probable structure type for that occupancy in that decade. Imputations were made for about 2.52% of the total number of buildings, most of which occurred in schools, colleges, emergency response, mobile homes, apartment homes and hotels, as shown in Table 3.7.

Table 3.7 -- Imputations by occupancy categories

Occ Type	Occupancy Description	Imputed Count	Total Count	Percent (Occupancy)	Percent (Imputed Total)	Percent (Total)
COM1	Retail Trade	3	4,020	0.07%	0.04%	0.00%
COM2	Wholesale Trade	7	4,891	0.14%	0.09%	0.00%
COM3	Repair Services	7	1,576	0.44%	0.09%	0.00%
COM4	Commercial Offices	12	2,930	0.41%	0.16%	0.00%
COM5	Banks	3	220	1.36%	0.04%	0.00%
COM6	Hospitals	-	22	0.00%	0.00%	0.00%
COM7	Medical Offices	-	408	0.00%	0.00%	0.00%
COM8	Restaurants	8	1,322	0.61%	0.10%	0.00%
COM9	Theaters	-	28	0.00%	0.00%	0.00%
COM10	Parking	-	50	0.00%	0.00%	0.00%
EDU1	Schools	250	280	89.29%	3.24%	0.08%
EDU2	Colleges	16	16	100.00%	0.21%	0.01%
GOV2	Police/Fire	39	48	81.25%	0.51%	0.01%
IND1	Heavy Industrial	2	709	0.28%	0.03%	0.00%
IND2	Light Industrial	8	324	2.47%	0.10%	0.00%
IND4	Extraction	-	27	0.00%	0.00%	0.00%
IND5	High Tech	-	14	0.00%	0.00%	0.00%
REL1	Religious	805	1,021	78.84%	10.43%	0.26%
RES1	Single-family	172	269,442	0.06%	2.23%	0.06%
RES2	Mobile Homes	15	43	34.88%	0.19%	0.00%
RES3	Apartments	6,225	18,135	34.33%	80.67%	2.03%
RES4	Hotel/Motel	127	331	38.37%	1.65%	0.04%
RES5	Dormitories	16	59	27.12%	0.21%	0.01%
RES6	Nursing Homes	2	87	2.30%	0.03%	0.00%
Totals		7,717	306,003	2.52%	100.00%	2.52%

Over 80% of the imputed buildings were multi-family residential, 10% were churches, and the rest were distributed over schools, single-family and hotels as highlighted in the table. The Emergency Response, Schools and Colleges category imputations, though small in number, are significant because of their individual sizes and consequently the capital investment in these buildings.

The multi-family apartment imputations are significant because of the sheer number of imputations – over a third of the apartment buildings were imputed for number of dwelling units, square footage or both. The consequences of the imputations in terms of the number of dwelling units in multi-family structures is explained in the validation section of the concluding chapter in Section 5.1. Table 3.8 shows the dwelling unit imputations for multi-family apartment buildings by decade of construction. Square footage imputations for apartments by decade echoed the trends seen in Table 3.8.

Table 3.8 -- Multi-family residential imputations of dwelling units by decade

Decade	Imputed DU	Unimputed DU	Total DU	Percent by Decade	Percent by Imputed	Percent by Total
Pre-1939	631	7,947	8,578	7.36%	1.16%	0.49%
40-49	863	3,792	4,655	18.54%	1.58%	0.67%
50-59	1,497	5,542	7,039	21.27%	2.75%	1.16%
60-69	9,127	17,388	26,515	34.42%	16.75%	7.06%
70-79	16,361	20,418	36,779	44.48%	30.02%	12.65%
80-89	9,594	8,948	18,542	51.74%	17.60%	7.42%
90-99	8,391	4,978	13,369	62.76%	15.40%	6.49%
Post-2000	8,032	5,803	13,835	58.06%	14.74%	6.21%
Totals	54,496	74,816	129,312	42.14%	100.00%	42.14%

Imputations were based on similarity between the record with incomplete information and other similar building records, with similarities based on decade of construction, frequency measures in exterior walls and structure types, building sizes, etc. For instance, if a multi-family residential building did not have information on the number of dwelling units, we imputed the number of dwelling units by computing the square footage per dwelling unit for structures similar in area and built in the same decade. A similar approach was used if the number of dwelling units was available, but not the square footage. If neither dwelling unit nor square footage was available, apartment buildings were inspected and crudely digitized from aerial photographs to compute approximate square footages. Several apartment complexes seemed to be

vacant (particularly in the 1960 and 1970 decades) from the aerial photographs, marked by the complete absence of parked cars in the parking lots. These were not recorded as vacant, since the inspections were incidental to the imputation process and not comprehensive in nature (the visual inspection process was exhaustive, time-consuming and certainly felt comprehensive!). Most of the imputations occurred in multi-family residential apartments and condominiums.

Table 3.9 shows the imputed, unimputed and total average areas per square foot for apartments by decade of construction. Note that the imputation averages of square foot per dwelling unit resulted in remarkably consistent numbers for each decade. This table therefore also validates the multi-family residential building imputation process.

The imputations are of primary concern because they provide input to other calculated fields. In addition, several imputations occurred in buildings that have a higher concentration of capital invested per unit.

Table 3.9 -- Multi-family residential imputations for dwelling size by decade

Decade	Imputed Square Feet per Dwelling Unit	Unimputed Square Feet per Dwelling Unit	Total Square Feet per Dwelling Unit
Pre-1939	911	972	968
40-49	725	790	778
50-59	684	691	690
60-69	797	814	808
70-79	946	943	944
80-89	897	875	886
90-99	751	773	759
Post-2000	899	953	922
Totals	864	871	868

3.1.5. Spatial Representation of Extracted Buildings

The Tax records also contained a spatial dataset of parcel boundaries for Shelby County with the correct parcel identifier. Spatial X-Y coordinate information was

generated for each building by extracting the parcel centroid spatial X-Y coordinates and linking them to the parcel identifier in the building inventory database. In other words, each building was geocoded to the centroid of the parcel that it was sited in. While this is not spatially precise, it at least ensures that each building is necessarily located within the boundary of the parcel polygon that contains it. The location of the building is specified by the LAT and LON coordinates, as well as projected X and Y coordinates in the building dataset – note that all the buildings within a parcel share the same location specified by the centroid coordinates of the parcel and are therefore coincident. Figure 3.2 shows an extract for a residential area in the central portion of Memphis, showing parcels and building points.

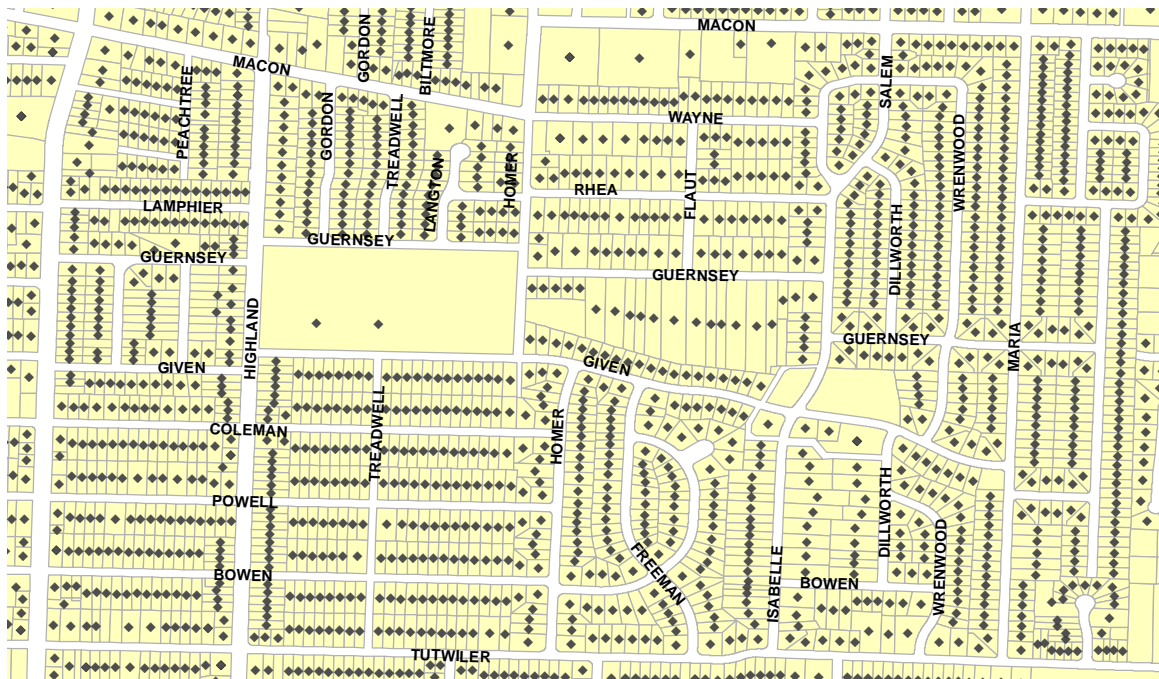


Figure 3.2 -- Extract of parcels and buildings in Central Memphis

3.2. Sample Data Collection

Determining the structure type of buildings required a training or calibration dataset, which contained all combinations of the input variables and the known or desired structure type. This would enable the estimation of the structure type model parameters and an understanding of the independent variables that were significant in terms of associations with the structure type class. In other words, the calibration dataset would enable the creation of a model that could then be used to predict the structure type for the remaining population of buildings.

From the initial analysis of the Tax Records, there were over 280,000 structures in Shelby County, with almost 90% being single-family residential structures. Since capital investment is concentrated in larger non-single-family residential buildings, field surveys were conducted in order to generate sample datasets for calibration and validation of non-single-family structures only. The choice to sample only commercial and industrial buildings is explained further in Section 3.3. Graphic details of the location of predominant sampling areas within Shelby County, Tennessee, are shown in Figure 3.3, showing the samples collected along with labels listing the corridors chosen for the survey.

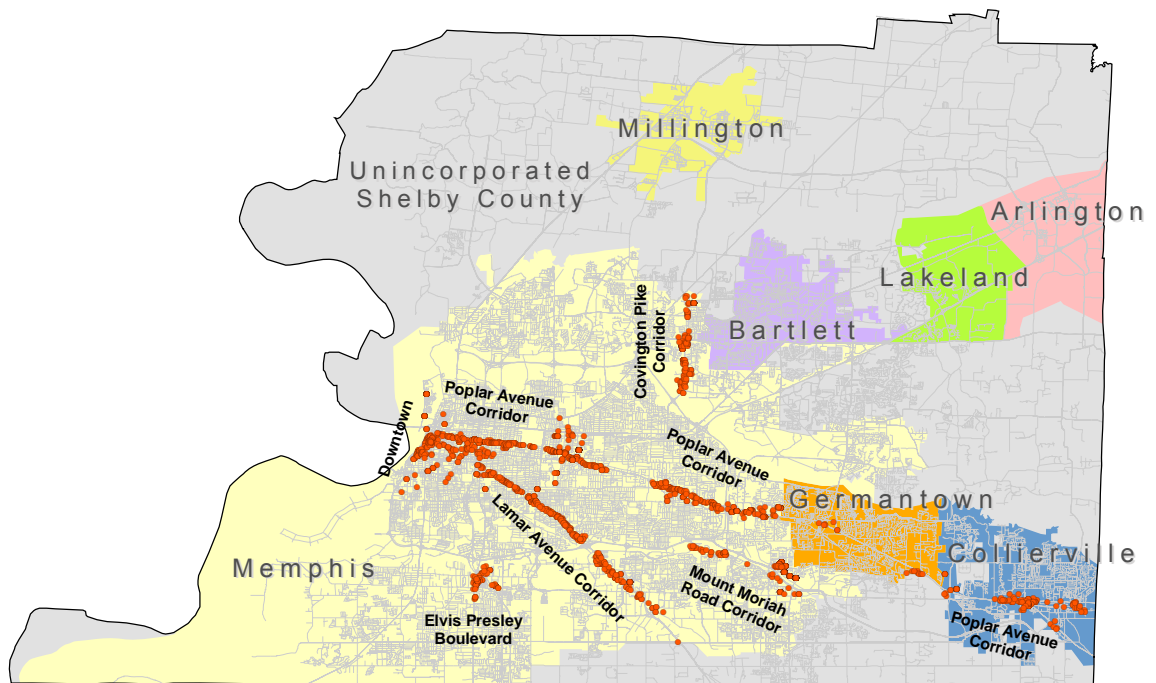


Figure 3.3 -- Survey sample collection areas in Shelby County

Based on FEMA guidelines for the rapid visual screening of buildings for potential seismic hazards (ATC-21 1988), two field surveys of non-single-family residential structures were conducted in May and October of 2003. The survey team consisted of one graduate research assistant and myself. We anticipated vehicle-based windshield surveys in areas of low traffic and walking surveys where slow driving would pose hazards (non-seismic, of course!). Where it was not possible to judge the structural system of the building (particularly for larger steel or reinforced concrete buildings), we would enter the building with permission for closer inspection (which caused the US Secret Service, Memphis Field Office to take an undue interest in our research activities). Since non-single-family structures tend to be located along major boulevards and arterials, the survey areas were designed to sample buildings along commercial-intensive corridors, rather than a cluster-oriented design. Major streets from the Shelby

County street database were overlaid on aerial images and corridors inspected for concentrations of large buildings.

The downtown area was extensively surveyed (based on walking) since a large amount of capital is invested in the central business district. Poplar Avenue, a long corridor running across Shelby County, was extensively surveyed, since it contained a significant number of commercial and industrial buildings of various sizes. The Lamar Avenue and Elvis Presley Boulevard corridors were chosen because of the fairly high concentration of industrial structures. The Covington Pike corridor also contained several industrial and warehouse type buildings. Towards the end of the day, incidental samples of smaller areas were collected in clusters dispersed through the study area. To be entirely honest, despite this initial preparation, we had no idea of how long it would take to collect the structural information on a building-by-building basis, and second, on how many samples we could gather. Further, since we had only the addresses and aerial photographs to initially design survey routes, we could not control the generation of sufficient samples for intersections of variables – in other words, we could not a priori determine the minimum number of samples for each cell in a cross-tabulation of structure type and occupancy.

The two surveys yielded 1831 buildings over 1043 addresses. The sample database is described in the next section.

3.2.1. Description of Sample Data

Only address and corridor information items were available at the beginning of the survey. After the surveys were complete, the samples were reconciled with the correct address and the parcel number. In cases where only one or two buildings were observed in the sample (typically in the case of apartment complexes), all other buildings

in the parcel were coded to the same structure type. The assumption is not troublesome because most buildings in a parcel, particularly in the case of multi-family apartments, tend to be built in the same time period, using the same methods of construction. While some additional samples were collected in the field for other addresses in the vicinity of the survey, these could not be reconciled with the parcel database, owing to incorrect or missing address information in the Tax Records, and about 120 samples were not used. After the samples were reconciled with the parcel identifiers, the remainder of the Tax record attributes were attached and analyzed. Table 3.10 shows the distribution of structure types for the sample.

Table 3.10 -- Sample structure type frequency

General Structure Type	Code	Count	Percent
Concrete Moment Resisting Frame	C1	99	5.41%
Concrete Frame with Concrete Shear Wall	C2	30	1.64%
Concrete Tilt-up	PC1	67	3.66%
Precast Concrete Frame	PC2	16	0.87%
Reinforced Masonry	RM	184	10.05%
Steel Frame	S1	245	13.38%
Light Metal Frame	S3	185	10.10%
Unreinforced Masonry	URM	301	16.44%
Light Wood Frame	W1	321	17.53%
Commercial Wood Frame	W2	383	20.92%
Totals		1,831	100.00%

Note that concrete structures seem undersampled, while wood structures occur most frequently. Table 3.11 shows the distribution of occupancy types for the sample. Again, note that several occupancy classes did not occur in the sample, while some categories were undersampled.

Table 3.11 -- Sample occupancy type frequency

Occupancy Description	Occupancy Type Code	Count	Percent
Retail Trade	COM1	291	15.89%
Wholesale Trade	COM2	327	17.86%
Personal and Repair Services	COM3	139	7.59%
Professional/Technical Services	COM4	181	9.89%
Banks	COM5	42	2.29%
Hospital	COM6	2	0.11%
Medical Office/Clinic	COM7	29	1.58%
Restaurants and Bars	COM8	96	5.24%
Theaters	COM9	1	0.05%
Parking Garages	COM10	27	1.47%
Education (Grade Schools)	EDU1	4	0.22%
Education (Colleges)	EDU2	0	0.00%
Emergency Services (Police/Fire/EOC)	GOV2	0	0.00%
Heavy Industrial	IND1	46	2.51%
Light Industrial	IND2	18	0.98%
Food/Drugs/Chemicals	IND3	0	0.00%
High Technology	IND4	0	0.00%
Place of Worship	REL1	15	0.82%
Single-family Residential	RES1	0	0.00%
Mobile Home	RES2	0	0.00%
Multi-family Residential	RES3	581	31.73%
Temporary Lodging (Hotel/Motel)	RES4	30	1.64%
Institutional Dormitory	RES5	0	0.00%
Nursing Home	RES6	2	0.11%
Totals		1,831	100.00%

Table 3.12 shows the cross tabulation of structure by broad occupancy types. This table clearly identifies gaps in the sample, particularly for concrete structure types and for some uses. To some extent, this is expected – for instance, construction practices preclude the use of wood as a structural system for hospital occupancies. In some cases, the frequency of the occurrence of that specific occupancy might be low in the general population of buildings, or may have been located away from major arterials, or even simply not have been located in our sampling areas. These gaps may have implications for the modeling exercise.

Table 3.13 shows the cross tabulation of structure type by decade of construction. The sampling frequency of decade is consistent with Memphis' growth

periods. Further, the structure types show good consistency between construction practices and decade. For instance, concrete tilt-ups are not seen till the late 1950s. In masonry buildings, reinforced masonry structures are seen only after 1974 when unreinforced masonry structures disappear. To be fair however, the survey coded a structure as “masonry” and the assignment to either the “RM” (Reinforced Masonry) or “URM” (Unreinforced Masonry) class for masonry samples was based on the year of construction.

Table 3.12 -- Sample cross tabulation of Structure type by broad occupancy

Structure Type	Retail	Wholesale	Office	Restaurants	Hospitals
C1	7	12	59	1	2
C2	-	-	18	-	-
PC1	7	38	3	-	-
PC2	-	-	-	-	-
RM	42	59	14	25	-
S1	84	32	60	2	-
S3	15	111	3	2	-
URM	110	49	47	12	-
W1	23	26	36	46	-
W2	3	-	12	8	-
Totals	291	327	252	96	2
Percent	15.89%	17.86%	13.76%	5.24%	0.11%
Structure Type	Parking	Industrial	Schools	Churches	Multi-family
C1	1	10	3	-	4
C2	-	-	-	-	12
PC1	-	19	-	-	-
PC2	16	-	-	-	-
RM	-	42	-	1	1
S1	10	24	-	4	29
S3	-	54	-	-	-
URM	-	50	1	1	31
W1	-	5	-	5	180
W2	-	-	-	4	356
Totals	27	204	4	15	613
Percent	1.47%	11.14%	0.22%	0.82%	33.48%

Table 3.13 -- Sample cross tabulation of Structure type by decade

Structure type	Pre-1939	40-49	50-59	60-69
C1	29	2	17	28
C2	9	-	4	7
PC1	1	-	3	12
PC2	-	-	2	1
RM	-	-	-	1
S1	8	14	28	29
S3	2	1	12	29
URM	114	47	69	60
W1	9	50	25	31
W2	13	2	31	50
Totals	185	116	191	248
Percent	10.10%	6.34%	10.43%	13.54%
Structure type	70-79	80-89	90-99	Post-2000
C1	16	7	-	-
C2	3	6	1	-
PC1	17	23	9	2
PC2	1	7	2	3
RM	43	76	57	7
S1	27	68	49	22
S3	38	46	47	10
URM	11	-	-	-
W1	31	132	38	5
W2	65	210	10	2
Totals	252	575	213	51
Percent	13.76%	31.40%	11.63%	2.79%

3.3. Structure Type Classification

Preliminary analyses of appraised values in Shelby County and HAZUS MH MR-3 GBS showed that while the total building value was concentrated in the single-family occupancy category, the average value of the single-family occupancy was among the lowest. Single-family homes in Shelby County had an overall value of almost \$25 billion – however, this amount was distributed over more than 266,000 structures, yielding low average values. The distribution of average square footage also followed this pattern, where single-family homes had the lowest average square footage, while larger building areas were observed in commercial and industrial occupancies. The preliminary analyses also revealed that the bulk of capital invested in buildings was concentrated in

commercial and industrial occupancies. Previous studies had established the resiliency of single-family wood structures in the context of earthquakes – in past earthquakes, single-family wood structures showed little damage compared with other structures (ATC-13 1985; FEMA 2004; FEMA - DHS 2007). Consequently, we separated the dataset into single-family and non-single-family structures and decided to perform the structure classification modeling for only the 36,561 non-single-family structures.

3.3.1. Multinomial Logistic Regression

Since structure type classifications show remarkable consistency in urban areas in relationships with occupancy and year of construction and size, the models for classifying structure types for buildings could be specified in terms of these relationships. Hence, an input parameter space consisting of both factors (categorical) and covariates (continuous) could be analyzed in order to identify patterns that correspond to the structure type categories. Accordingly, the structure type classification was assumed to be a function of the building size (area), height (number of stories), year of construction, and building occupancy.

The separation of Concrete structures into Concrete Moment Frame, Concrete Frame with Shear Wall, Precast Concrete and Concrete tilt-up was not always possible in the field. Parking structures were predominantly supported by precast concrete columns and beams, but it was difficult to ascertain the precast nature of concrete when the concrete was covered by some other finish. Concrete tilt-ups were easier to spot, but that depended on whether the particular tilt-up walls were in the survey view. For instance, there were several cases where the survey had identified buildings as supported by a Steel Moment Frame, when the inspection of the Tax Records external wall code revealed that they were a concrete tilt-up structures. It is quite possible for the side walls of the building to have tilt-up panels, while the front and back walls are

fenestrated or have more conventional construction. For the modeling exercise, all concrete structure types were grouped into a single category, and the individual concrete differentiations could be extracted by inspections of supporting attributes from the Tax records.

Building square footage, number of stories and year of construction were directly acquired from the building inventory database. Building occupancy in the database consisted of 82 categories of detailed occupancies, or 29 HAZUS MH MR-3 categories of specific occupancies, or 12 general occupancies. Using so many levels for occupancy could pose two potential problems. First, the number of samples available for training may not be adequate, in the sense that there may be too few or even no exemplars of a certain structure type-occupancy combination -- the model would not be able to estimate the parameters adequately for this input set. Second, even if there were enough samples, the number of levels in the results would make interpretation very complicated. Consequently, the occupancy categories were collapsed into a set of 8 categories, including "Retail Trade", "Wholesale Trade", "Commercial Offices", "Banks", "Restaurants", "Heavy Industrial", "Light Industrial" and "Multi-family Apartments".

The structural fire rating variable in the Tax Records had a very consistent relationship with structure type. The ratings were categorized as "Fire Resistant", "Fire Proof", "Engineered Steel" and "Wood Joists". In the sample, wood joist levels were overwhelmingly Wood structure types. Concrete structure types were split almost equally among fire proof, fire resistant and engineered steel ratings. Steel frames were split predominantly between fire resistant and engineered steel ratings. Table 3.14 shows the cross tabulation between model structure type and the structural fire rating code.

Table 3.14 -- Sample Structure type and Structural fire rating category

Structure Type	Structural Fire Rating Code				Row Total	Row Percent
	Fire Proof	Fire Resistant	Engineered Steel	Wood Joists		
C	58	69	67	4	198	10.81%
RM	-	109	4	69	182	9.94%
S1	8	157	67	13	245	13.38%
S3	-	1	181	3	185	10.10%
URM	14	154	-	149	317	17.31%
W	1	5	-	698	704	38.45%
Totals	81	495	319	936	1,831	100%
Percent	4.42%	27.03%	17.42%	51.12%	100%	

Based on the concentration of non-single-family residential built square footage per acre generated for structures built before 1940, a density grid was generated in order to approximately demarcate a polygon identifying a historic zone. Structures within this zone were given a true value for a historic zone dummy variable.

The multinomial logistic regression therefore attempts to classify structure type on the basis of building area, number of stories, year of construction, presence in a historic zone, occupancy and structural fire rating characteristics.

3.3.2. Design of Neural Network Topology for Classification

All the topologies suggested in this section may be implemented in the NeuroSolutions application software released by NeuroDimension, Inc. Five specifications of ANN topologies are suggested for the classification problem.

The input data and desired samples for the ANN model were exactly the same as described in the previous section, where the ANN attempts to classify structure type on the basis of building area, number of stories, year of construction, presence in a historic zone, occupancy and structural fire rating characteristics.

3.3.2.1. Multilayer Perceptron

Multilayer Perceptrons (MLPs) extend Rosenblatt's (Rosenblatt 1958) perceptron (a single layer perceptron) that could solve only linearly separable classification problems into a classification device capable of nonlinear classification. In the MLP, each PE is characterized by a smooth non-linear function (either the logistic or the hyperbolic tangent function) and the PEs are massively and fully interconnected in a manner that any PE in a layer connects to every other PE in the succeeding layer. The MLP is trained with error correction learning. Using the gradient descent construct, each weight in the network is changed using a function of the inputs and the instantaneous error at that iteration. The total local error computed at the output PE is distributed backwards through the network based on the output sensitivity to that weight, using only local information (Rumelhart et al. 1986). Momentum learning allows a memory term (previous increments or decrements to the weight) to speed up convergence and avoid getting trapped in local minima or flat areas of the input space (Principe et al. 2000). MLPs are extremely powerful classifiers capable of reproducing almost any input-output combination set, but require lots of exemplars and may train slowly. Figure 3.4 shows the schematic topology for an MLP network with 1 hidden layer. Note that based on the input variables, there are 21 PEs in the input layer massively connected to the 8 PEs in the hidden layer. The 8 PEs in the hidden layer are again, massively connected to the 8 PEs in the output layer. Each PE in the output layer is used to test the probability of one structure type against all the others. The step size parameter was set at .1 and the momentum parameter at 0.7.

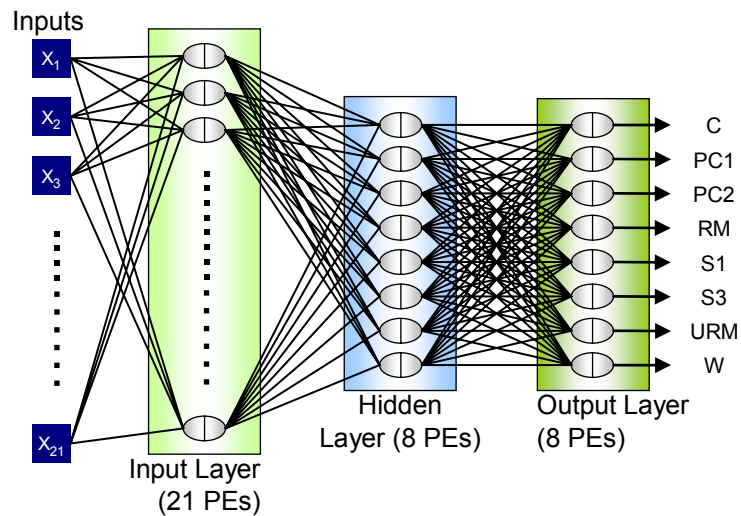


Figure 3.4 -- Schematic of MLP network for structure type classification

3.3.2.2. Generalized Feed Forward Network

A Generalized Feed Forward (GFF) network is an extension of the MLP, except that connections can jump over one or more layers. While in theory, an MLP can solve any classification problem, the GFF solves the problem much more efficiently, because weight modification can potentially proceed forward by skipping layers that have little effect on the output. The caveat is that too many hidden layers will result in overtraining and performance in testing or unseen exemplars is heavily degraded. Figure 3.5 shows the schematic topology for a GFF network that is essentially the same as the previous MLP network, except that here, the input layer is also massively connected to the output layer directly (as marked by the curved arrow in the figure).

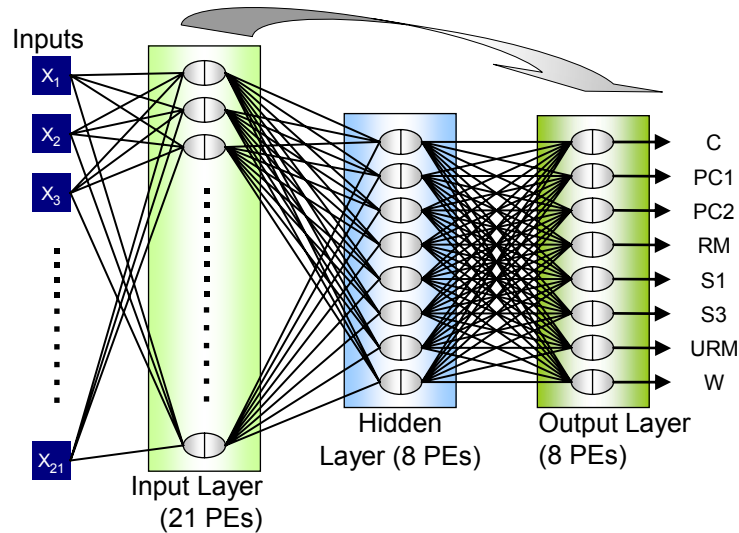


Figure 3.5 -- Schematic of GFF network for structure type classification

3.3.2.3. Modular Neural Network

Modular Neural Networks (MNNs) are again special cases of MLPs, where the layers are divided into modules. Unlike the MLP, MNNs do not have massive interconnectivity between layers, and therefore fewer network weights are required. This topology often speeds up the training and achieves the same relative level of accuracy with fewer exemplars than an MLP. Creating the network topological structure in this case is essentially an exercise in segmenting each hidden layer into modules, and specializations of functions in each sub-module have been observed. In practice however, there is no guarantee that the specialization occurs with the same combination of input data consistently, nor are there guidelines for the best modular design among the various alternatives. Figure 3.6 shows the schematic topology for an MNN with 2 hidden layers that are identical, each with 8 PEs. The hidden layers are segmented into 2 modules each, consisting of 4 PEs. Unlike the previous topologies, the input layer is only massively connected with each module of the hidden layer. The modules of the first

hidden layer are connected to the corresponding modules in the second hidden layer as well as to the output layer (as marked by the curved arrows in the figure).

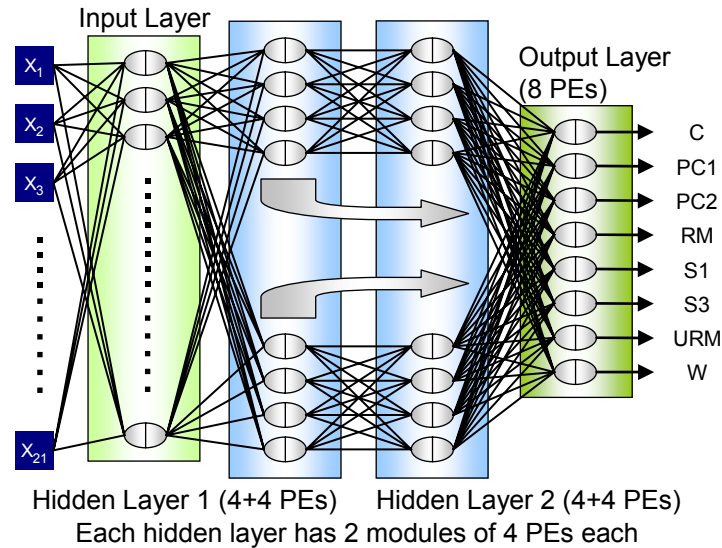


Figure 3.6 -- Schematic of MNN for structure type classification

3.3.2.4. Radial Basis Functions

The accuracy of a classifier depends on the location and shape of the decision boundary in the input space. Since the decision boundary is determined by solving discriminant functions, the location and shape of the discriminant is critical in designing classifiers. In reality, our underlying functional data distributions may be incorrect or we may have too few samples for an optimal parametric classifier. In many cases, classification becomes a trade-off between optimality and robustness (Rojas 1995). Classifiers based on Radial Basis Functions (RBF) offer some potential solutions.

Radial functions are characterized by monotonic increase or decrease in the response based on distance from a central point, and its parameters include the center, the distance scale and the shape of the function. One might conceptualize basis functions as a linear sequence of local functions in the input space with parameters that

alter the location, center and shape of each local function, thereby allowing the sequence to approximate the input space. Typically, a Gaussian function serves as a local function sequence (ibid). Figure 3.7 shows the linear combination of Gaussian functions that approximate the input space. The Gaussians are centered (location of the means) and stretched (variance spread) and use other properties (skewness) to alter the heights of the Gaussians.

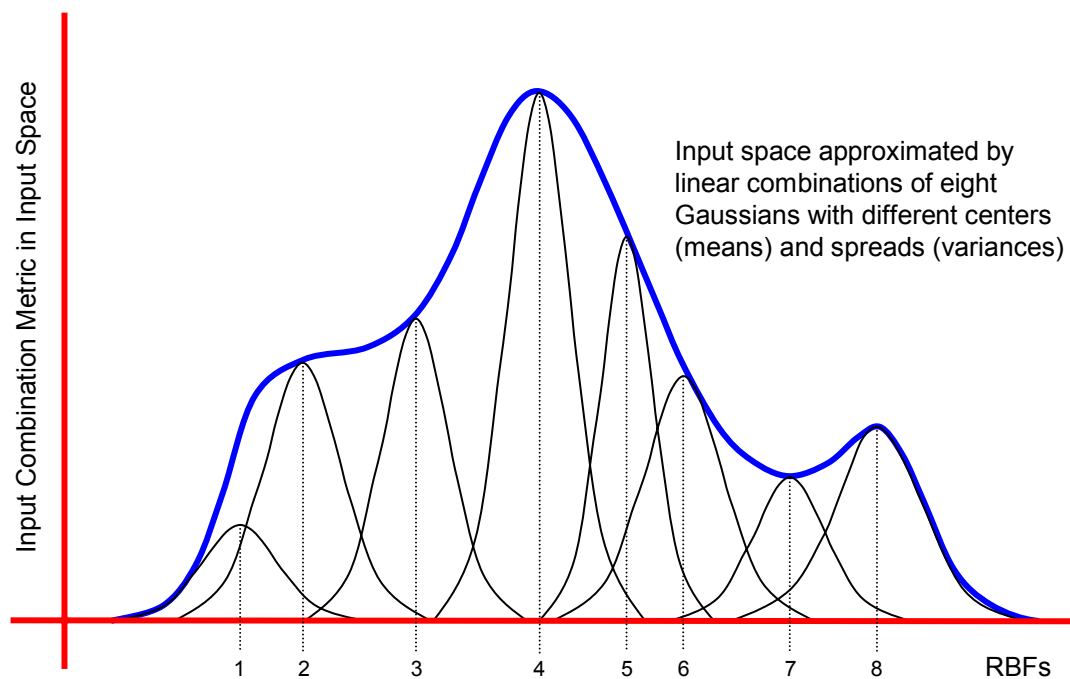


Figure 3.7 -- Linear combination of RBFs used for approximation

Obviously, the centers of the Gaussians should locate at the clusters of the data samples in the input space. Given a fixed number of Gaussians, variances can be estimated and altered to cover the input space (Haykin 1994). Once the centers and variances have been computed, a simple soft maximizing classifier can adapt the weights in order to interpret the outputs as probabilities. See Figure 3.8 for a schematic representation of the RBF network.

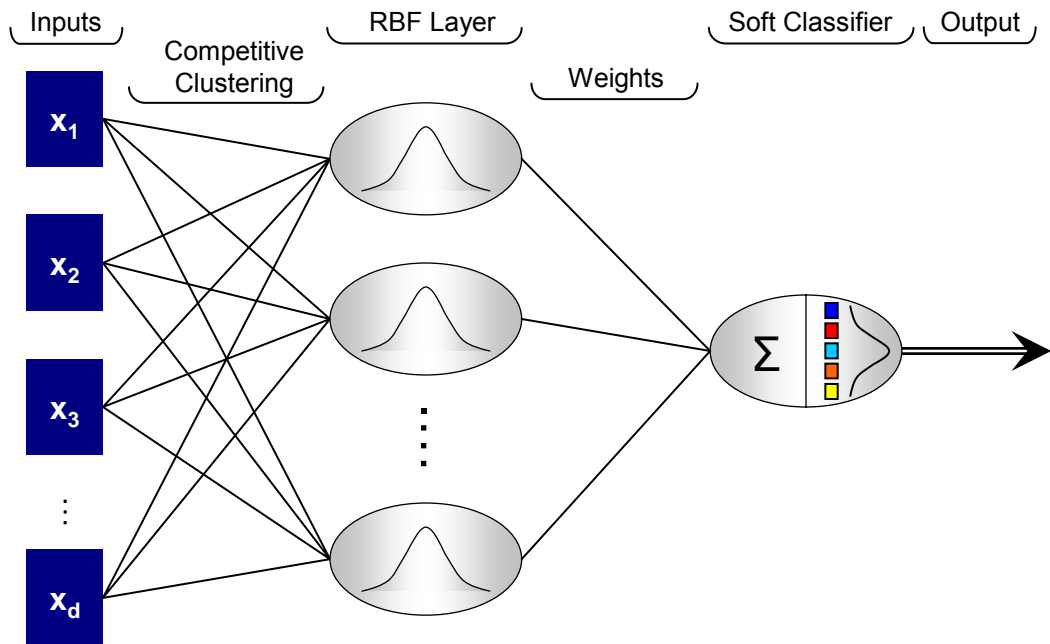


Figure 3.8 -- Schematic representation of RBF network

While there are several methods to center the Gaussians on the data clusters, the most common method is the K-means algorithm with competitive learning. Here, the samples are divided into K clusters, each with an initially randomly assigned center. Centers are then moved to minimize the Euclidean distance between the input cluster and the Gaussian center (Prager and Fallside 1989; Michie et al. 1994). The challenge here is to determine the number of bases – too few and the classification performance is poor; too many and spurious classifications may result in new samples because of overfitting (Geman 1992; Principe et al. 2000). Most neural computing software applications recommend internal validation mechanisms by setting aside some samples for testing and validation or early stopping of training (Hanson 1990; Wynne-Jones 1993).

3.3.2.5. Support Vector Machines

More sophisticated classifiers may be created by mapping the inputs into a higher dimension space and then classifying using linear discriminants. Using Cover's theorem (Cover 1965) that any pattern recognition problem is separable in a sufficiently high dimensionality space, the input space may be nonlinearly transformed into a higher dimension feature space. Consider a three-dimensional input space defined by $[x_1, x_2$ and $x_3]$. Using a kernel function we can convert this three-dimensional space into a nine-dimensional space as $[x_1^2, x_2^2, x_3^2, x_1*x_2, x_2*x_3, x_3*x_1, x_1, x_2, x_3]$ The first three dimensions of the higher dimensional space are computed by multiplying the input with itself, the next three by multiplying each input with the succeeding input, and the final three by using the inputs directly. Subsequently, a linear discriminant function may be constructed for this higher dimension space (Freiss and Harrison 1998). Thus, the ANN architecture consists of a kernel processor followed by a linear classifier (Principe et al. 2000). Further, Vapnik (1999) recently showed that for symmetric kernel functions, the weights can be computed without the requirement of solving the problem in the higher dimension space, giving rise to a new class of classifiers called support vector machines (SVM). Higher dimension spaces produce sparse data clusters with lots of room between clusters, and therefore classification may be effected using very simple classifiers. The SVM is implemented directly in the software application without any need for user-specified topology or other parameters.

3.4. Building Footprint Classification

There are many variations in building footprint polygons that arise from their method of extraction or creation. By merely viewing a particular building footprint polygon, a human could judge the polygon as belonging to a particular class, despite the noise in the building edge. Automating this process and accurately recording the shape

class in a manner that mimics the human classification process would result in considerable time saving and increased shape classification reliability.

Based on the concepts and surveys of the literature, this section outlines guidelines for the shape classification process and then designs a structural approach based on landmark correspondence for classifying the building footprint polygons.

3.4.1. Guidelines for Shape Classification Design Process

Shape analysis methods should be designed for the particular application at hand, and often, the methods of implementation are dictated by a compromise among accuracy, recognition efficiency and computational complexity (Kauppinen et al. 1995). For instance, the need to recognize the license plates of speeding vehicles passing through a checkpoint in real time is very different from the subject of this research – a non-real-time identification of building footprint configuration in a GIS database – consequently, the design of the two applications will be quite different.

Choosing the features that describe the shape is crucial, and optimal features and feature combinations should have high discriminative power (Duda et al. 2001). In general, any shape may be described by a set of feature descriptors. General characteristics of feature descriptors or attributes for consistent shape description and accurate recognition include discriminatory power, robustness to noise, disturbance or occlusion, invariance to geometric transformations, scalability and performance. Optimally chosen features partition the feature space into clearly defined and well separated class groups. The methods for the choice of features are not generally consistent and often likened more to an art than a science (Costa and Cesar 2001a).

Scalar feature descriptors may then be combined into a vector before application in shape classification processes. Thus, a feature vector F_k that consists of 'k' measures is in the real feature space of 'k' dimensions, or $F_k \rightarrow R_k$.

The choice of feature measures should be such that similar shapes will be mapped into points in the feature space that are proximal to one another and quite distant from dissimilar shapes. Highly correlated features should be discarded or indexed into single features. Parsimony in the number of features may yield processes with higher discriminatory power and be less computationally burdensome (Dryden and Mardia 1998; Duda et al. 2001). If shape variations are caused by specific transformations, like rotation or scale, it might be worthwhile to normalize features and make them invariant to those specific transformations (Bishop 1995; Costa and Cesar 2001a). Prior to beginning the shape analysis process, the designer should review all relevant material in the literature and apply existing techniques.

The designer should also become thoroughly familiar with the typology of shapes to be classified and to some extent, understand the process causing the typologies of distinct shapes. A library of typical building footprints was created that could be used for calibration and validation exercises. Further, the very process of creating the library enabled familiarity with the typology of building shapes, particularly in cases of typological transitions, such as a human process that judges an L-shaped building as a rectangle, because the stem of the "L" has a very small dimension. It would also be worthwhile to implement alternative methods and analyze their performance both in terms of computational performance and recognition accuracy.

3.4.2. Preprocessing and Collinear Vertex Decimation

The existing contour edge geometry of the building footprint polygon is processed in order to remove collinear vertices – in other words, based on a user-specified threshold angle parameter that defines linearity between two segments, all vertices that do not substantially alter the generalized linear gradient of the polygon's edge segments are decimated. Figure 3.9 depicts the process by which collinear vertices are decimated.

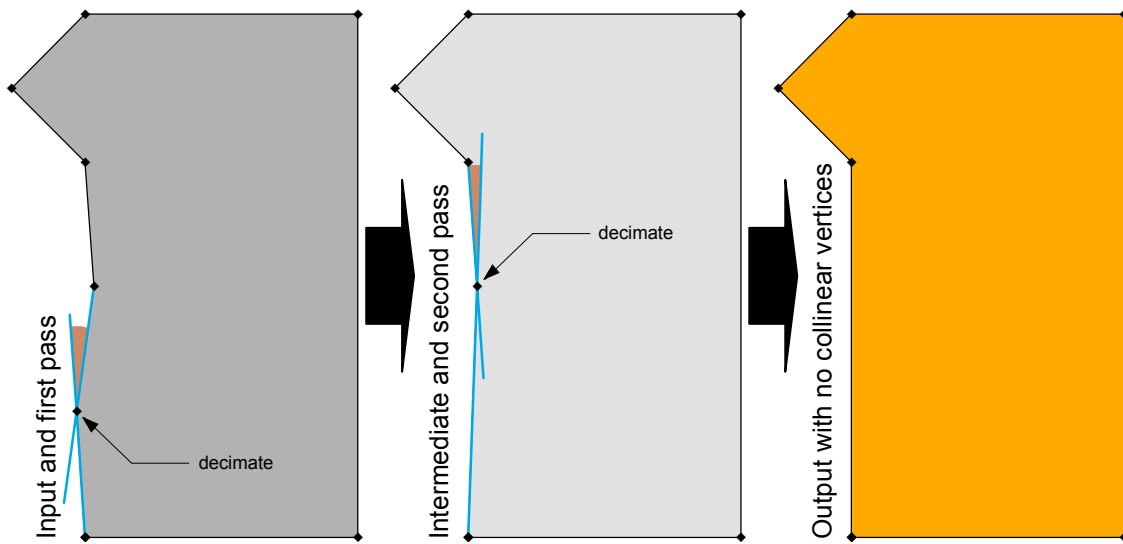


Figure 3.9 -- Removal of collinear vertices from polygon edges

If two successive line segments are collinear or near collinear, then the angle between them is approximately " π " radians or 180° . In such cases, the two line segments may be combined into one line segment, by connecting the first point of the first segment with the end point of the second segment, and decimating the intermediate point. Angles of 170° to 190° between successive line segments may be regarded as artifacts of the extraction process, so a threshold angular tolerance of 10° (the threshold

may be changed by the user) implements segment linearity redefinitions. Angles greater than the specified threshold are assumed to be legitimate edges of the building footprint.

The algorithm begins by extracting the vertex geometry of the polygon in a point collection. The vertices are analyzed in groups of three, where the central angle is computed. If the central angle is within the user-specified threshold of the line connecting the first and third vertices, the second point is removed from the collection. This process is repeated until no more vertices are decimated. The polygon geometry is adjusted to reflect the new vertex collection, which will be less than or equal to the original number of vertices.

3.4.3. Orthogonalization of Polygon Edges by Corner Vertex Adjustment

In general, most buildings have smooth perimeters and corners defined by the rectilinear intersection of sequential edge segments. Automated generalization processes typically include a refinement step in which the polygon geometry is adjusted both for visual clarity and correspondence with reality.

The orthogonalizing algorithm, which is depicted in Figure 3.10, begins by extracting the vertices from the polygon geometry as an array of points. The vertices are analyzed sequentially in groups of 3. The first three vertices are selected, as seen in the second panel of Figure 3.10. The second vertex or the interior or central vertex in the analysis set is moved along the line connecting it to the first vertex subject to the condition that the angle subtended by the three vertices equals 90° . The analysis set is then changed by incrementing the vertices by one index position. Thus, the second, third and fourth vertices form the new analysis set. Note that the coordinate location of the second vertex had been adjusted in the previous step. As seen in the third panel, the third vertex (which is the interior or central vertex in this analysis set), is moved along

the line connecting it with the first vertex in the analysis set again subject to the condition that the angle subtended by the three vertices equals 90° . The analysis set is changed by successive increments of one index position and the process repeated till the coordinate locations of all vertices except the first have been adjusted. The new set of vertices is then used to update or create the orthogonalized polygon geometry, where every corner is defined by a 90° angle.

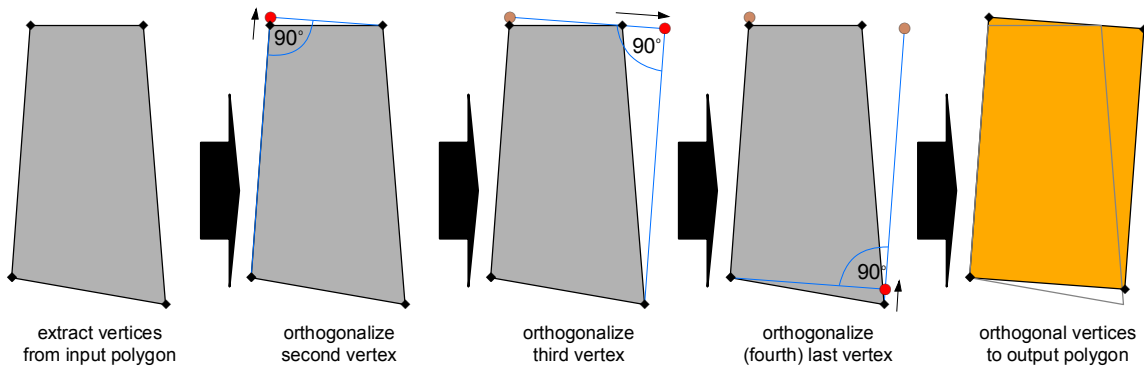


Figure 3.10 -- Orthogonalizing building edges by adjusting corners

An artifact of the orthogonalization process is that the area of the orthogonalized polygon might differ significantly from the input polygon. Consequently, the orthogonalized polygon may be scaled from the centroid of the output polygon by the square root of the ratio of the output and input areas. Alternately, if the starting vertex is designated as the beginning of the longest line segment in the polygon boundary, orthogonalized area does not differ significantly from the input polygon.

3.4.4. Building Footprint Analysis by Landmark Correspondence

There are several potential methods to choose from the existing literature. For instance, each input polygon shape could be transformed into an invariant form. Thus, following Bookstein (1991), each input shape could be rotated till its longest dimension is parallel with the X-axis, translated till the centroid of the shape coincides with the origin

and scaled separately along the coordinate axes such that the overall extents of the shape fits into a unit square ranging from (-0.5, -0.5) to (0.5, 0.5). Then the view invariant transformed shape could be analyzed for feature extraction. The polygon could then be represented using Hu's moments (1962) or Zernike polynomials (Rothe et al. 1996; Zhang et al. 2003), or as a sequence of landmarks (Belongie et al. 2002; Adamek and O'Connor 2003), or as a sequence of vertex distances from the origin (Gupta and Srinath 1988), or as a sequence of line sequences defining the polygon edge (Liu and Srinath 1990). The extracted features of the sample polygon could then be compared with extracted features for reference polygons representative of each class in the analysis and the polygon assigned to the most similar class, based on feature comparisons or Fourier transforms or statistical or structural and syntactic approaches.

The shape recognition design follows a syntactic approach to classifying building footprints based on landmark correspondence. Following typical methods in structural and syntactic shape analysis, first, landmark vertices on the contour of the building footprint polygon are extracted. This is a crucial step in the analysis, because the approach is predicated on building footprint classes to be represented by a specific and atomic set of vertex locations that uniquely define each class. In the application, non-collinear landmarks are defined based on convexity or concavity thresholds. Then, the landmarks of the sample shape are rotationally aligned with those of the reference shapes and classification occurs, based on successful correspondences between the landmarks. Figure 3.11 outlines the flow of tasks in this proposed methodology, beginning with preprocessing the input shape polygon and proceeding to the feature representation by extracting the landmark sequence and convexity properties, and culminating in the rotation-based correspondence algorithm for shape classification.

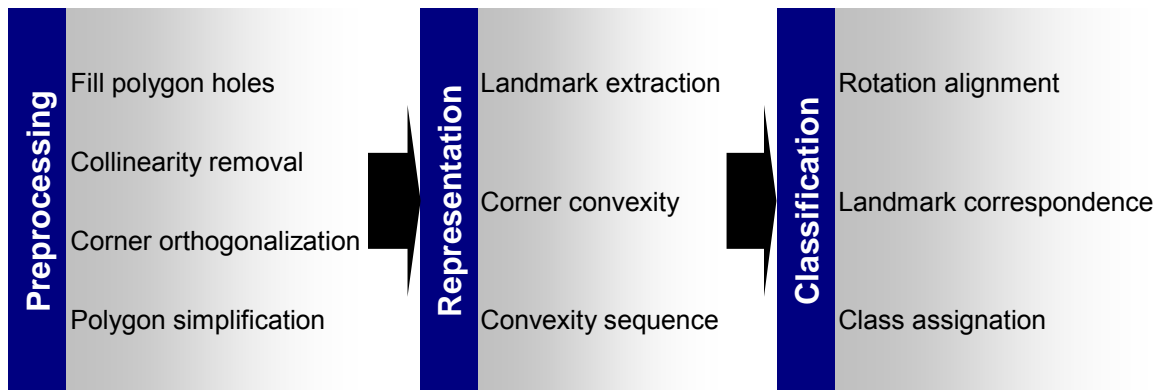


Figure 3.11 -- Footprint classification by landmark correspondence

3.4.4.1. Computing Circularity Indexes to Eliminate Circular Buildings

Circles are represented in GIS as a sequence of straight line segments that approximate the circle boundary. Consequently, a circular building footprint in the GIS is characterized by several (well over 500) vertices. Landmark correspondence measures require the extraction of high curvature or salient points, and in the case of circles, all the vertices have the same curvature. If circular building footprints are identified and eliminated, the landmark correspondence methods can proceed to identify the other shapes. Circular polygons may be identified by their Circularity Ratio, F_{cir} , described in Section 2.4.2.4. The circularity index is a function of the Area and Perimeter of the polygon and all buildings that have values over 0.9 may be eliminated from further analysis as circular buildings.

3.4.5. Building Footprint Polygon Simplification

Since the building footprint polygon is discretely represented through linear polylines between vertex locations (or approximated by linear segments for parametric curves), and the automated extraction or manual digitization process greatly varies with the method or human responsible for the capture of the feature, several irrelevant convexities and concavities may be introduced. Many of the convexities (or protrusions)

and concavities (or intrusions) may be parts of contours following roof lines and therefore may not correspond with the structural system of the building. The roof line and therefore the extracted polygon acts as a proxy for the exterior-most component of the structural system of the building. Structural systems are generally linear arrangements of columns and beams, so a number of the protrusions and intrusions, especially the smaller ones, are mere artifacts of the polygon creation mechanism, and should be removed.

In the context of building polygon simplification, built-in tools in the GIS environment (specifically ArcGIS) for generalization and simplification work well for topologically disconnected buildings. Using these tools, extraneous details in the building edges may be removed without compromising the essential size and shape of the building. Since buildings are orthogonal areas, the tool will convert near 90-degree corners to exactly 90 degrees. Edges of buildings are assumed to comprise of linear segments and usually run parallel. Isolated small offsets in the boundary resulting in small intrusions or extrusions are filled or widened. The final output will contain fewer vertices than the original, but the area will be consistent with the input polygon. Polygons smaller in size than a user-specified threshold will be decimated. See Figure 3.12 for an example of using built-in simplification routines in ESRI ArcGIS 9.2. However, when the connections between buildings are complicated, the tool completely ignores the features, as seen in the top-right polygon of Figure 3.12 – the complicated geometry is labeled for easy recognition and copied to the output, and additional manual processing may be necessary.

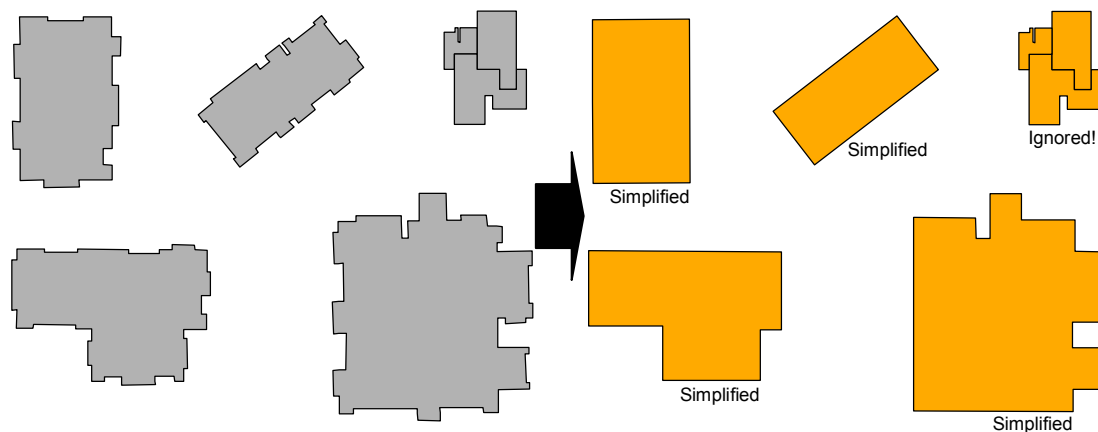


Figure 3.12 -- Performance of built-in building simplification tools in GIS

In the pre-processing stage of this research, it may be necessary to modify building polygon geometry to adjust for angular manual digitizing inaccuracies and artifacts that may be beyond the simplification tool thresholds. This will include removal of collinear vertices and running orthogonalization routines on the input footprint polygons. In addition, polygon footprints will be corrected for specific patterns in the geometry for which standard simplification routines fail. For instance, since we are interested in the overall shape of the structure, any “holes” in the building footprint polygons must be appropriately filled. Within the GIS environment, we will apply routines that identify polygon geometry in terms of rings and remove all interior rings, thus retaining only the outermost ring and filling all the holes.

Additionally, protrusions and intrusions from and into the building footprint polygon greater than the threshold parameter in the direction perpendicular to the contour edge and less in the direction parallel to the edge result in simplification artifacts that are generally larger than they should be. Figure 3.13 depicts two cases of simplification failures, one for protrusions and the other for intrusions. Protrusions are characterized by vertex convexity sequences of concave, convex, convex and concave

from the beginning vertex of the protrusion, while intrusions are characterized by convex, concave, concave and convex vertices, as shown in Figure 3.13 – in both cases, the beginning vertex of the deviation from the polygon edge is reached through a clockwise traversal of vertices. These specific cases for protrusions and intrusions will be appropriately modified before simplification through automated pattern-spotting and cleaning routines within the GIS environment. After appropriately adjusting the footprint geometry to lie within the simplification tool thresholds, the polygons may be simplified and submitted as inputs to the next stage of feature extraction and/or analysis.

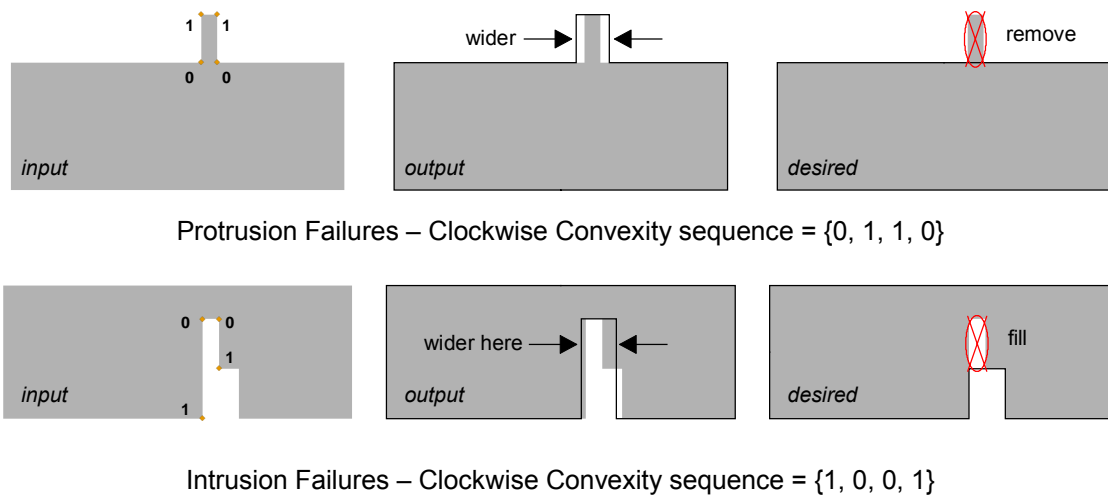


Figure 3.13 -- Simplification failure artifacts and desired results

3.4.6. Identification of Salient Points

The identification of salient points or landmarks for each building footprint polygon is indeed the crux of this application. As specified earlier, 2D building shape types in this research include square, rectangle, L-, C-, T-, H-, Z-, octagonal, circular, cruciform and irregular. For each of these classes, a sequence of landmarks that uniquely identifies each class needs to be extracted.

Based on the Dryden and Mardia (1998) definition of landmarks, significant and unique vertex sets that correspond strongly within classes are extracted. In the context of building footprint polygons, a vertex is deemed significant if it is an inflection point that defines a strong change in the curvature of the edge between the preceding and succeeding arcs (Fischler and Wolf 1994; Cesar and Costa 1995, 1996). Based on significance alone however, a large number of vertices could potentially be extracted, depending on the detail of the input building footprint polygon edge. Several vertices from the potential set of landmarks should therefore be decimated if they do not add information to the polygon edge in terms of class membership (Bookstein 1991). In other words, based on some threshold distance or tolerance parameter, vertices from the edge will be decimated if their location or deviation is less than the threshold tolerance. All collinear or near-collinear vertices should be decimated, again on a curvature-based angular threshold.

For this application, a linear tolerance of 10 feet is suggested, because changes larger than 10 feet usually require some structural enhancement to the building and further, could enclose potentially large building areas. Expressed another way, bays and protrusions that are greater than the 10-foot tolerance parameter could be structurally significant and not mere artifacts of the footprint extraction or creation process. Angular tolerance for three-vertex collinearity is recommended as a $\| 10^\circ \|$ deviation at the middle vertex from the straight line connecting two outer vertices. These tolerances may be altered, depending on the structural system for the building – for instance, steel frame buildings may have a 20-foot tolerance, while Unreinforced Masonry buildings may have an 8-foot tolerance.

Despite all the variations in building footprint polygons that appear as artifacts of the extraction or manual digitizing process, a human would easily be able to identify and

judge that a particular shape as belonging to a specific class. Consider the input polygon shown in the left panel of Figure 3.14 – despite the various deviations of the polygon edge from a straight line, it is readily apparent that the polygon should be assigned to the class of “L” shapes. It is also readily apparent that the “L” shape may be defined by 6 landmarks that specify its corners, as seen in the right panel of Figure 3.14.



Figure 3.14 -- Desired landmark outputs that mimic human judgment

The automated building footprint polygon recognition process needs to mimic human judgment for standard cases, and therefore for each standard class of polygon shapes, we identify a particular set of landmarks. The task is generally made easier because buildings are generally characterized by straight line segments and orthogonal corners. Undoubtedly, there will still exist a few cases that are ambiguous even for a human – for instance, if the stem of a “T” shape deviates more than the specified threshold tolerance from the polygon edge, but is particularly small in dimension compared to the overall extents of the polygon, it might as well be classified as a rectangle. Figure 3.15 shows two examples of ambiguous cases – the building in the panel on the left could be identified as either a rectangle or an “L” shape, while the polygon in the panel on the right could belong to either of the “H” or “C” building classes. Implementing decision rules for class readjustment is beyond the scope of this research.

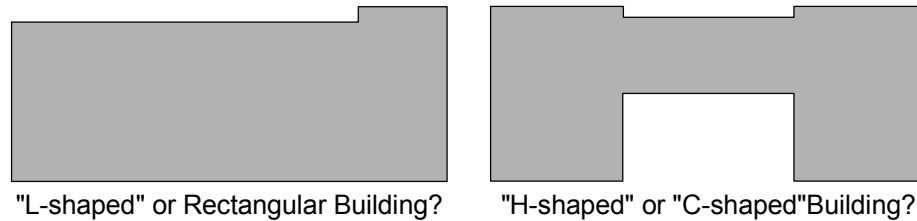


Figure 3.15 -- Ambiguity in building footprint polygon classification

3.4.7. Derivation of Landmark Sequences by Contour Traversal

The primary advantage with defining a unique sequence of landmarks with each footprint class is that both the number of landmarks and the sequence of convexity/concavity can be used for shape recognition. The preprocessing stage decimates all collinear vertices and smoothes the polygon edge in order to eliminate non-structural convexities and concavities. The simplified footprint polygon therefore contains only those linear elements that approximate the exterior-most frame of the structural system. This simplified polygon is characterized by vertices that serve as points of inflection – in other words, the segments preceding and succeeding the vertex differ significantly in curvature. Typically, in buildings, vertices will be defined at orthogonal corners or points of high curvature and may be used to define landmarks.

While the distance between landmarks will vary considerably, the sequence of landmarks will not, and therefore, the landmark representation of the footprint will be invariant to translation, rotation, scaling and even shearing. Such landmark sequences will uniquely define each class of polygons. However, for matching two “T-shaped” polygons, for instance, the landmark sequences require rotation alignment, or the same beginning point. If the starting point is correctly identified for each polygon, then each succeeding landmark in one shape will correspond to its counterpart in the other shape. Rather than use computationally expensive context-based approaches (Belongie et al.

2002) or heuristic (Gdalyahu and Weinshall 1999) or sub-optimal approaches (Adamek and O'Connor 2003), I propose a simple, computationally inexpensive algorithm that would rotationally align and match corresponding landmarks for similar shapes and classify based on successful correspondence in a manner concordant with the human judgment process for footprint classification.

Each class of footprint polygons to be identified in the research is represented by a typical polygon. For each of the polygon shapes representative of a class, a set of high-curvature landmarks in sequence is extracted by traversing the polygon edge from any arbitrary starting vertex. Each landmark set is converted into a unique sequence that represents that class of footprint polygon shapes by establishing a pattern of landmark convexities and concavities. Convexities are represented by “1” and concavities are represented by “0”. Figure 3.16 shows 9 of the 10 shape classes that need to be identified in this research along with their landmark convexity/concavity pattern sequences – the convexity patterns shown in the figure may be used as reference classes.

Each panel shows a representative polygon shape along with that shape's sequence of landmarks. Each landmark is labeled with two numbers – the first number indicates the position of the landmark in the landmark sequence, while the second number indicates whether the landmark is convex or concave. Only the octagonal shape is not shown, since it is very similar to the circular form, only with fewer landmarks. For instance, as seen in Figure 3.16, the middle-left panel shows a “T-shaped” polygon whose representative clockwise landmark convexity pattern is [1, 1, 1, 0, 1, 1, 0, 1]. This pattern describes all “T-shaped” building polygons and is invariant to the location, rotation or size of the building.

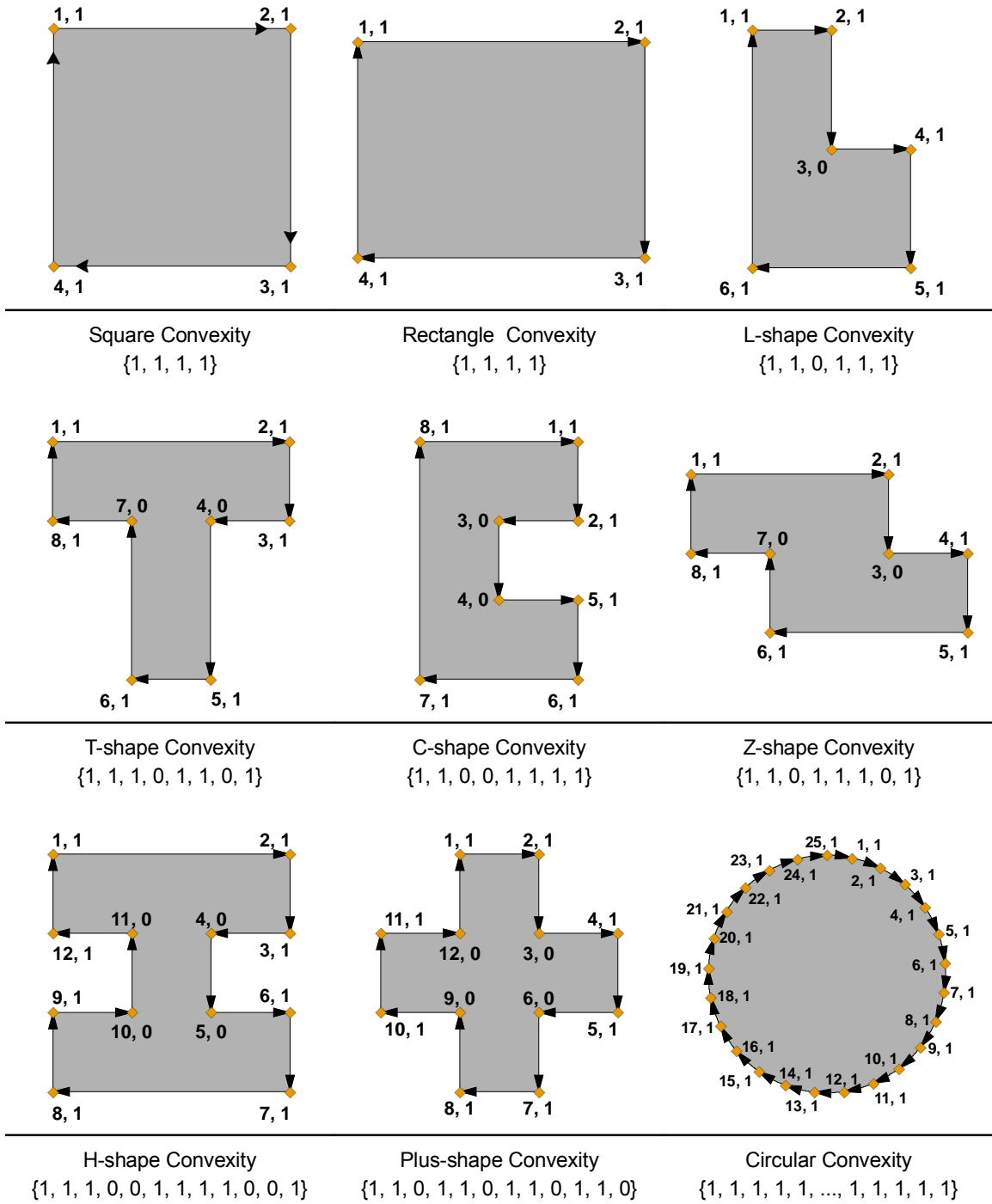


Figure 3.16 -- Landmark convexity sequences for polygon footprint classes

3.4.8. Binary Representation of Landmark Convexity

Each footprint reference class is uniquely described by a landmark convexity sequence. The convexity sequence consists of a pattern of ones and zeroes, where one indicates convexity and two concavity, as the contour is traversed from the starting point in the clockwise direction. When the sequence pattern is concatenated into a string, it is apparent that the sequence represents a number in binary format. In a variation of Bribiesca and Guzman's approach (1980), the binary sequence is altered by moving one element from the end of the sequence to the beginning until the largest number is identified. This transformed sequence now represents the largest binary number uniquely representing the shape. Effectively, the starting point for the landmark sequence has been successively rotated until the largest binary number was identified. Thus, extending the "T-shaped" polygon example, the beginning landmark convexity pattern, as a binary number was "11101101" and the sequence was successively altered till the largest binary number was identified. Finally, the "T-shaped" reference polygon class is represented by the largest binary number "11110110" or the landmark sequence [1, 1, 1, 1, 0, 1, 1, 0]. In other words, the starting landmark (for the largest binary number representation) for the "T-shaped" class is now position 8 (see Figure 3.16). Table 3.15 shows the initial representation from Figure 3.16 and the final largest binary number representation for each of the reference shapes in the analysis. The largest number representation will serve as the reference class identifier and will be used for shape classification.

Table 3.15 -- Initial and final landmark sequence binary representation

Footprint Reference Class	Initial Starting Position (Figure 3.16)	Initial Binary Representation (Figure 3.16)	Final Starting Position (Figure 3.16)	Largest Binary Representation for Classification
Square	1	1111	1	1111
Rectangle	1	1111	1	1111
L-shape	1	110111	4	111110
T-shape	1	11101101	8	11110110
C-shape	1	11001111	5	11111100
Z-shape	1	11011101	8	11101110
H-shape	1	111001111001	12	111100111100
Cruciform	1	110110110110	1	110110110110
Octagon	1	11111111	1	11111111
Circle	1	11111111...11111111	1	11111111...11111111

3.4.8.1. Determining Landmark Convexity

Since the application requires a sequence of vertex properties, the polygon edge is traversed from the specified starting point in the clockwise direction and consecutive vertices are analyzed in successive groups of three. In each set of three vertices, the first and third vertices are used analytically in order to determine the concavity or convexity of the middle point.

Adapting the Chaudhuri and Samal (2007) concept of a point and its “belongingness” relationship with any line, the middle vertex convexity or concavity property is determined by inserting its coordinate values into the straight line equation specified by the first and third points. Consider the two panels in Figure 3.17 – the panels show three landmarks encountered in clockwise order along with their coordinate values. If the total value of the function obtained by inserting the coordinates of the middle point is equal to zero, the middle point is on the line. If the value is greater than zero, the middle point is above the line and therefore the middle point is convex. If the value is less than zero, the middle point is concave.

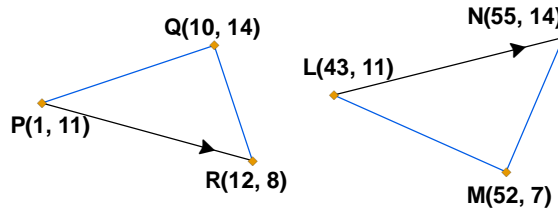


Figure 3.17 -- Determining if a polygon vertex is convex or concave

Determining the convexity or concavity property in this manner is equivalent to computing the value of the determinant of the triangle formed by taking the three points in the same traverse order. If three sequential points are specified by (x_1, y_1) , (x_2, y_2) and (x_3, y_3) , then the equation specified by the first and third points is given by

$$(y - y_1) - \left(\frac{y_3 - y_1}{x_3 - x_1} \right) (x - x_1) = 0, \text{ where slope is rise/run using the first and third points.}$$

If $f(x, y)$ is the left-hand-side of the equation, substituting the middle point in $f(x, y)$ yields

$$f(x_2, y_2) \Rightarrow (y_2 - y_1) - \left(\frac{y_3 - y_1}{x_3 - x_1} \right) (x_2 - x_1)$$

Expanding and manipulating the terms results in

$$f(x_2, y_2) \Rightarrow x_1 y_3 + x_2 y_1 + x_3 y_2 - x_1 y_2 - x_2 y_3 - x_3 y_1, \text{ which is the same as the determinant of the triangle specified by the three points in order.}$$

If $f(x_2, y_2) > 0$, then (x_2, y_2) is above the line and therefore convex. If $f(x_2, y_2) = 0$, the three points are collinear. Finally, if $f(x_2, y_2) < 0$, then (x_2, y_2) is below the line and therefore concave. Consider the left panel in Figure 3.17, with the three points P , Q and R . Substituting the coordinates of Q into the equation specified by the line PR gives $f(Q) = 60 > 0$, so Q is deemed a convex vertex. In the right panel of Figure 3.17, substituting

the coordinates of the middle point M in the equation specified by the line LN gives $f(M) = -25 < 0$, so M is deemed a concave vertex.

3.4.9. Building Footprint Classification

The building footprint algorithm is simple and computationally inexpensive. Any preprocessed sample footprint polygon shape is analyzed to yield the landmark sequence of ones and zeroes. This binary sequence is successively shifted by one position till the largest binary number is identified. The largest binary number for the sample is compared to the binary numbers specified for each of the reference classes with the same number of vertices. If the numbers are identical, the landmarks are aligned and the match is made to the appropriate class. If the equality condition fails, then the shape is assigned to the “Irregular” class. After determining the polygon footprint class, the algorithm records the class name in the footprint database.

3.5. Building Valuation

The building valuation component requires the estimation of the building replacement costs, the structural component of the replacement costs, the acceleration- and drift-sensitive components of the replacement costs and the content value. All the replacement cost components will be derived from R. S. Means 2008 Square Foot Costs (R. S. Means 2008). In the Residential section, the Means manual contains building square foot costs for seven building types (1-story, 1.5 story, 2-story, 3-story, Split bi-level, Split tri-level and Wings/Ells) in four different classes of construction (Economy, Average, Custom and Luxury). Costs per square foot are listed for various external wall types (wood frame, brick veneer and solid masonry) and basements (finished and unfinished), along with adjustments for car garages.

In the Commercial/Industrial/Institutional section, the Means manual contains building square foot costs for 2008 for 72 model buildings. Each model building has a table of square foot costs for combinations of the exterior wall and structure type for a range of areas typical for that occupancy class of buildings. The base tables do not reflect basement construction costs, but include average basement costs for that occupancy class. Further, for each model, a typical example is selected, and the various component assembly costs for that example are listed, both as line-item costs for the sections comprising the assembly and the total percentage cost of the assembly.

The square footage costs for each model type and external wall-structure type combination are derived from US National averages for 2008, and have to be adjusted for Memphis using the appropriate location factor adjustment. Residential costs in Shelby County were 0.82 of the National average residential costs, and 0.86 of the National average commercial/industrial costs.

Table 3.16 shows the 2008 national average and Memphis location-adjusted square foot costs for a 1-3 story apartment building (Model No. M.010) for combinations of exterior wall and structure type. Note that the range of a typical 1-3 story apartment varies from a minimum of 8000 sq. ft. to a maximum of 36,000 sq. ft. It is not necessary however, for all 1-3 apartment buildings to conform to this range – the Tax Records had several buildings for all model types that were less than the minimum or exceeded the maximum area for that occupancy.

Table 3.16 -- R. S. Means square foot costs - 1-3 story Apartment building (M.010)

External Wall	Frame	Area									Region	
		8000	12000	15000	19000	22500	25000	29000	32000	36000		
Brick, Concrete Block	Steel	185.55	165.95	158.10	148.30	144.50	142.20	137.50	136.10	134.10	National Average	
	Wood	182.10	161.50	153.20	142.35	138.35	135.90	130.60	129.10	126.95		
Brick Veneer	Steel	169.65	149.95	142.05	132.10	128.25	126.00	121.20	119.70	117.75		
Stucco, Concrete Block	Steel	173.65	154.95	147.50	138.80	135.15	133.05	128.95	127.55	125.75		
	Wood	160.35	141.80	134.35	125.75	122.15	120.00	116.00	114.60	112.80		
Wood Siding	Wood	159.05	140.70	133.30	124.85	121.25	119.15	115.25	113.85	112.10		
Brick, Concrete Block	Steel	159.57	142.72	135.97	127.54	124.27	122.29	118.25	117.05	115.33		Adjusted for Memphis, TN
	Wood	156.61	138.89	131.75	122.42	118.98	116.87	112.32	111.03	109.18		
Brick Veneer	Steel	145.90	128.96	122.16	113.61	110.30	108.36	104.23	102.94	101.27		
Stucco, Concrete Block	Steel	149.34	133.26	126.85	119.37	116.23	114.42	110.90	109.69	108.15		
	Wood	137.90	121.95	115.54	108.15	105.05	103.20	99.76	98.56	97.01		
Wood Siding	Wood	136.78	121.00	114.64	107.37	104.28	102.47	99.12	97.91	96.41		

The square footage costs for each combination of occupancy type, number of stories, external wall and structure type will be parameterized using standard curve fitting techniques. From the Tax Records, the corresponding combination of occupancy type, number of stories, external wall and structure type will have the replacement value estimated as the product of the building square foot cost and the building area. The building square foot costs will be estimated from the corresponding curve that captures the Means square foot cost to building area curve. Based on the literature, the replacement costs for the building will be segmented into structural and nonstructural component costs.

Since data for content value was not available, the content value will be estimated as a function of replacement costs for each occupancy category. The following sections describe the building value estimation process in greater detail.

3.5.1. Curve Fitting Routines for Model Building Square Foot Costs

The square foot costs and the area range for each combination of occupancy type, number of stories, external wall and structure type was coded into a database. Then, based on the area range and the square foot costs, curves were estimated for

each combination, setting the per square foot costs as the dependent variable and the discrete area as the independent variable. The minimum and maximum values for each combination were recorded in the database.

Four curve specifications including the linear, logarithmic, exponential and inverse models were estimated for each combination, and the parameters and the equation type of the best model was recorded for that particular combination. Figure 3.18 shows the parametric curves for the four different models for the 1-3 story Apartment supported by a steel frame structure type. Visual inspection suggests that the “Inverse” model fits the data closer than the other curves.

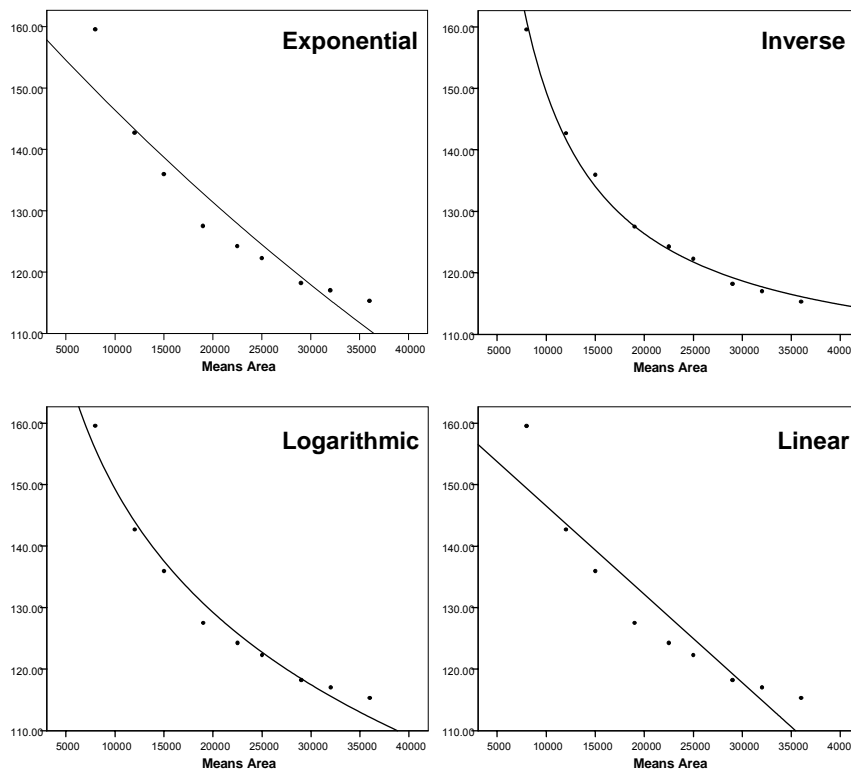


Figure 3.18 -- Parametric curves for a 1-3 story steel frame Apartment

Overall, parameters were estimated for 726 combinations. Consider the data in Table 3.16, showing the square foot costs for a 1-3 story apartment in Memphis. Note

that the minimum area is 8,000 sq. ft. and the maximum is 36,000 sq. ft. Note the corresponding square footage costs for the Face Brick with Concrete Block backup exterior wall supported by a steel frame – the minimum square foot costs (corresponding to the maximum area of 36,000 sq. ft.) is \$115.33 and the maximum square foot cost (corresponding to the minimum area of 8,000 sq. ft.) is \$159.57.

Table 3.17 shows the parameters for the four curves estimated for this line of data. Note that the best model specified by the R-squared criterion is the Inverse curve, highlighted in the table. The corresponding parameters are 103.333 for the Constant and 460843.583216 for the slope coefficient. The equation for estimating the square foot costs is $PerSqFt = 103.333 + (460843.583216 / BuildingTotSqFt)$

Table 3.17 -- Curve parameter estimates for a 1-3 story Apartment

Equation	R-squared	F-Statistic	Significance	b0	b1
Linear	0.869	46.282	0.00025	160.934	-0.001438
Logarithmic	0.974	265.774	0.00000	416.353	-28.991688
Inverse	0.995	1296.732	0.00000	103.333	460843.583216
Exponential	0.897	61.271	0.00010	163.115	-0.000011

3.5.1.1. Nomenclature for Model Buildings

The most challenging aspect of this exercise was to generate unique identifiers for each combination of occupancy type, number of stories, external wall and structure type. The Tax Records have information on the detailed and specific use, the height and external wall, while the artificial neural network will estimate the structural system of the building. The specific occupancy types in the Tax Records were standardized (for instance, “Discount Departments” and “Department Stores” were integrated into one category), and so were the number of stories and the external wall types (for instance, “Brick on Concrete Block”, “Brick on Block”, “Brick Veneer on Block”, “Brick with Block

Back-up”, “Brick with Concrete Block Back-up”, “Face Brick on Concrete Block”, “Face Brick with Concrete Block Backup” exterior wall descriptions in the Tax Records were all standardized to “Brick on Concrete Block”) – this step allowed the mapping of specific occupancy, number of stories, structure type and exterior wall to corresponding categories drawn from the Means manual. Table 3.18 shows the Means specific occupancy categories and their corresponding 3-digit codes.

Table 3.18 -- R. S. Means specific occupancy and 3-digit code

Specific Use	Code	Specific Use	Code
Apartment	001	Bank	041
College, Dormitory	001	Office	043
SF, Economy	002	Church	051
SF, Average	003	School, Elementary	053
SF, Custom	004	School, High	053
SF, Luxury	005	School, Vocational	054
Nursing Home	006	Fire Station	055
Assisted - Senior Living	007	Police Stations	055
Hotel	008	College, Classroom	056
Motel	009	Hospital	061
Store, Convenience	011	Medical Office	063
Store, Department	013	Bus Terminal	071
Store, Retail	015	Bowling Alley	081
Supermarket	017	Club, Country	083
Factory	021	Club, Social	085
Warehouse	023	Movie Theater	086
Warehouse, Mini	025	Restaurant	087
College, Laboratory	029	Restaurant, Fast Food	089
Car Wash	031	Mobile Home	099
Garage, Auto Sales	033	Garage, Parking	101
Garage, Repair	035	Garage, Underground Parking	101
Garage, Service Station	037	School, Jr High	053a

The number of stories from the Tax Records were coded using a 4-digit specification that corresponded with the Means number of stories for the specific occupancy type. For instance, a 2 story Apartment received the Number of Stories code as “0103”, since the actual building height was between 1 and 3 stories, corresponding to the Means model occupancy type of 1-3 story apartments.

Reconciling the external wall code between Means and the Tax Records occurred in two stages. First, the external walls from the Means Manual were standardized and coded, as seen from the external wall extract in Table 3.19. Then the external walls from the Tax Records, as seen in Table 3.20, were standardized and reconciled with the Means external wall codes.

The unique code identifying the particular combination was then generated by concatenating the Means occupancy code with the coded number of stories first, since the combination of occupancy and number of stories (that specifies the Means Model type) would be used for separating the structural and non-structural costs based on the assemblies for the Means model type. This Means model code was then concatenated with the structure type derived from the structure type classification model and the exterior wall code in order to generate a unique identifying code for each combination. Table 3.21 shows an extract of the replacement cost identifier code.

The raw Means data from the manual was recorded in the database using the replacement cost identifier code. Where the combinatory code and Means information did not coincide, an appropriate substitute based on judgment was chosen -- for instance, a small one-story factory building constructed in 1924, with external walls of unreinforced concrete block would be disallowed by the modern building code, and hence, the replacement value equation would substitute the combination with a one-story service garage with a galvanized steel siding external wall on a light metal frame.

Table 3.19 -- Standardization of Means External wall categories

Means External Wall Type	Means Standardized External Wall	Code
Brick Veneer	Brick Veneer	01
Face Brick Veneer	Brick Veneer	01
Face Brick on Steel Studs	Face Brick on Steel Studs	01
Face Brick Veneer on Steel Studs	Face Brick on Steel Studs	01
Face Brick w/ Structural Facing Tile	Face Brick w/ Structural Facing Tile	01
Concrete Block	Concrete Block	03
Decorative Concrete Block	Decorative Concrete Block	03
Painted Concrete Block	Painted Concrete Block	03
Precast Concrete Block	Precast Concrete Block	03
Concrete Block Stucco Face	Stucco on Concrete Block	03
Brick on Concrete Block	Brick on Concrete Block	04
Brick w/ Block Back-up	Brick on Concrete Block	04
Face Brick on Concrete Block	Brick on Concrete Block	04
Jumbo Brick on Concrete Block	Jumbo Brick on Concrete Block	04
Galvanized Steel Siding	Galvanized Steel Siding	07
Steel Siding on Steel Studs	Steel Siding on Steel Studs	07
Insulated Metal Panels	Insulated Metal Panels	08
Metal Sandwich Panel	Metal Sandwich Panel	08
Painted Reinforced Concrete	Painted Reinforced Concrete	09
Reinforced Concrete	Reinforced Concrete	09
Precast Concrete	Precast Concrete	10
Ribbed Precast Concrete Panel	Ribbed Precast Concrete Panel	10
Double Glazed Tinted Plate Glass Panels	Double Glazed Tinted Plate Glass Panels	11
Glass and Metal Curtain Wall	Glass and Metal Curtain Wall	11
Tiltup Concrete Panel	Tiltup Concrete Panel	14
Tilt-up Panels	Tiltup Panels	14
Limestone w/ Concrete Block Back-up	Limestone w/ Concrete Block Back-up	17
Stone Ashlar Veneer on Concrete Block	Stone Ashlar Veneer on Concrete Block	17
Stone w/ Concrete Block Back-up	Stone w/ Concrete Block Back-up	17
Stucco	Stucco	19
Stucco on Wood Frame	Stucco	19
Aluminum Siding	Aluminum Siding	24
Vinyl Siding	Vinyl Siding	24
Wood Shingles	Wood Shingles	25
Wood Siding	Wood Siding	25
Mobile Home	Mobile Home	29

Table 3.20 -- External walls from Tax Records reconciled with Means categories

External Wall (Tax Records)	External Wall (Means)	Code
BRICK & FRAME	Brick Veneer	01
BRICK VENEER	Brick Veneer	01
CONDO WALL	Brick Veneer	01
MASONRY & MTL	Brick Veneer	01
BLOCK	Concrete Block	03
BRICK & CONCRETE BLO	Concrete Block	04
BRICKCONCRETE BLO	Concrete Block	04
COMPOSITE	Concrete Block	03
CONCRETE BLOCK	Concrete Block	03
NATIVE STONE	Concrete Block	03
OTHER	Concrete Block	03
STONE	Concrete Block	03
METAL, LIGHT	Galvanized Steel Siding	07
METAL, SANDWICH	Metal Sandwich Panel	08
CONCRETE LOAD BEARIN	Poured Concrete	10
CONCRETE NON-LOAD BE	Poured Concrete	10
ENCLOSURE	Glass and Metal Curtain Wall	11
GLASS	Glass and Metal Curtain Wall	11
GLASS & MASONRY	Glass and Metal Curtain Wall	11
SOLAR GLASS	Glass and Metal Curtain Wall	11
CONCRETE TILT-UP	Tiltup Concrete Panel	14
MARBLE/SLATE	Stone Ashlar Veneer on Concrete Block	17
MASONRY & FRAME	Stone Ashlar Veneer on Concrete Block	17
DRYVIT	Stucco	19
STUCCO	Stucco	19
AL/VINYL	Vinyl Siding	24
ASBESTOS SHINGLE	Wood Shingles	25
ASBESTOS, COR. RIG.	Wood Siding	25
FRAME	Wood Siding	25
LOG	Wood Siding	25
MOBILE HOME	Mobile Home	29

Table 3.21 -- Generation of Replacement value identifiers

Specific Building Use		Structure Type		Stories		Exterior Wall		FullCode
Desc	Code	Desc	Code	Desc	Code	Desc	Code	
Apartment	001	Wood	AAAWF	1-3	0103	Brick on Concrete Block	04	001AAAWF010304
Apartment	001	Steel	AAASF	4-7	0407	Brick Veneer	01	001AAASF040701
Apartment	001	Concrete	AARCC	8-24	0824	Concrete Block	03	001AARCC082403
Factory	021	Concrete	AARCC	1	0001	Brick on Concrete Block	04	021AARCC000104
Factory	021	Steel	AAASF	3	0003	Tiltup Concrete Panel	14	021AAASF000314
Garage, Parking	101	Precast	AAAPC	all	0000	Precast Concrete	10	101AAAPC000010
Garage, Repair	035	Metal	AAALM	all	0000	Galvanized Steel Siding	07	035AAALM000007
Hospital	061	Concrete	AARCC	2-3	0203	Brick on Concrete Block	04	061AARCC020304
Hospital	061	Steel	AAASF	4-8	0408	Decorative Concrete Block	03	061AAASF040803
Hotel	008	Wood	AAAWF	4-7	0407	Brick Veneer	01	008AAAWF040701
Hotel	008	Steel	AAASF	8-24	0824	Glass & Metal Curtain Wall	11	008AAASF082411
SF, Economy	002	URM	URMRM	1	0001	Wood Siding	25	002URMRM000125
SF, Average	003	Wood	AAAWF	1	0001	Wood Siding	26	003AAAWF000126
SF, Custom	003	Wood	AAAWF	2	0002	Brick Veneer	01	003AAAWF000201

3.5.2. Estimating Replacement Costs for Buildings

Once the best fitting curves are parameterized for each combination of occupancy, number of stories, exterior wall and structure type, and the parameters coded to the replacement cost identifier code, estimating the replacement value is straightforward. The square foot costs of construction are estimated using the parameters of the curve in the equation type for that combination and multiplied by the total building square footage to yield the estimated replacement cost of the building. Thus, the per square foot costs for an example building whose total area is 11,800 sq. ft, 2 stories in height with exterior wall of brick on concrete block and supported by a steel frame is derived by inserting the building total area in the specified equation and is estimated as \$142.39. The replacement value for the building would be the product of the estimated square foot costs and the total building area, that is $\$142.39 * 11,800 = \$1,680,169.00$.

If the sample apartment building's area were 6,500 sq. ft. (less than the minimum area of 8,000 sq. ft. specified in Means for Apartments) with the remaining specifications

being the same, the maximum square foot costs (of \$159.57) would be used.

Correspondingly, if the area of the sample building were 45,000 sq. ft. (greater than the maximum area of 36,000 sq. ft. specified in Means for Apartments), then the minimum square foot costs (of \$115.33) would be used.

From the Tax Records, the floor area figures in square feet below ground level were extracted for each building – if the building had no basement, this would equal 0. The Means manual specified average basement construction costs of \$30.35 per square foot for 1-3 story apartments, which, locationally adjusted for Memphis, becomes \$26.10. The basement square footage is multiplied by this amount to estimate the basement construction cost. The above ground square footage is multiplied by the curve-estimated square foot costs for the above ground construction cost. The replacement cost of the building is therefore the sum of the construction costs below and above ground.

3.5.3. Structural and Non-Structural Replacement Costs

The Means Square Foot Costs also provided percentage breakdowns of costs for the various component assemblies of the building. These included the Foundation and Substructure, Superstructure, Exterior Enclosure, Roofing, Interiors, Conveyance Equipment, Water supply and Plumbing, HVAC, Fire Protection and Electrical Services and Special Construction. Based on background from Porter (2005) and Taghavi and Miranda (2003), Foundations, Superstructure, Roofing and Special Construction assemblies formed the Structural component, while Interiors, Conveyance Equipment, Water supply and Plumbing, HVAC, Fire Protection and Electrical systems assemblies were grouped under Non-structural Acceleration-sensitive components. The remaining Exterior Enclosure and Interior systems assemblies formed the drift-sensitive component. Table 3.22 shows the cost breakdown percentage for a sample of building

types. Figure 3.19 shows the percent breakdown of replacement costs into structural, nonstructural acceleration-sensitive and drift-sensitive components graphically for a subset of specific Means model types. Note that there is some variety in the percent breakdown, but the overall trend does indicate that nonstructural costs form a substantial cost component of the replacement costs.

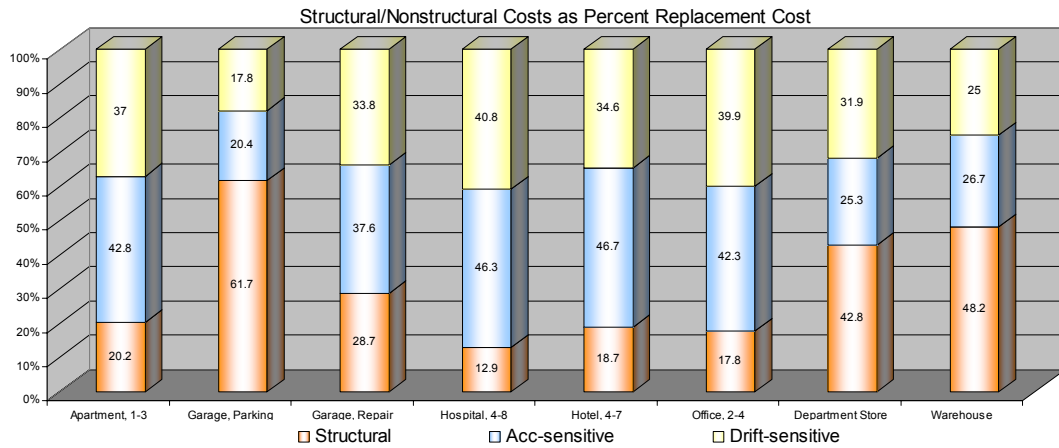


Figure 3.19 -- Structural/Nonstructural costs as percent Replacement costs

Table 3.22 -- Structural and Nonstructural cost breakdowns by Means models

Means Occupancy	Occupancy Code	Stories Code	Model Identifier	Structural	Acc-sensitive	Drift-sensitive
Apartment	001	0103	001_0103	20.2	42.8	37
Apartment	001	0407	001_0407	21.3	40.3	38.4
Apartment	001	0824	001_0824	23.1	36.8	40.1
Assisted - Senior Living	007	0000	007_0000	18.5	39.3	42.1
Auditorium	086	0000	086_0000	23.1	35.1	41.9
Bank	041	0000	041_0000	33.5	27.7	38.7
Bowling Alley	081	0000	081_0000	33.5	46	20.6
Bus Terminal	071	0000	071_0000	25.7	27.6	46.6
Car Wash	031	0000	031_0000	14.5	58.1	27.4
Church	051	0000	051_0000	31.3	28.9	39.9
Club, Country	083	0000	083_0000	15.8	48.7	35.6
Club, Social	085	0000	085_0000	21.9	37.8	40.3
College, Classroom	053	0203	053_0203	16	52.1	31.9
College, Dormitory	001	0203	001_0203	20.1	38.7	41.1
College, Dormitory	001	0408	001_0408	23	39.3	37.7
College, Laboratory	029	0000	029_0000	20.9	47.3	31.8
Factory	021	0001	021_0001	32.8	49.1	18
Factory	021	0003	021_0003	33.3	40.5	26.2
Fire Station	055	0001	055_0001	26.5	40.1	33.5
Fire Station	055	0002	055_0002	18.5	48.1	33.5
Garage, Auto Sales	033	0000	033_0000	34.3	29.8	36
Garage, Parking	101	0000	101_0000	61.7	20.4	17.8
Garage, Underground Parking	101	0000	101_0000	77	12.3	10.8
Garage, Repair	035	0000	035_0000	28.7	37.6	33.8
Garage, Service Station	037	0000	037_0000	23.3	32.3	44.4
Hospital	061	0203	061_0203	13.7	48.9	37.3
Hospital	061	0408	061_0408	12.9	46.3	40.8
Hotel	008	0407	008_0407	18.7	46.7	34.6
Hotel	008	0824	008_0824	22.7	45.7	31.7
Medical Office	063	0001	063_0001	17	40.5	42.6
Medical Office	063	0002	063_0002	14.6	45.8	39.6
Motel	009	0001	009_0001	24.1	33.1	42.9
Motel	009	0203	009_0203	13.8	38.4	47.9
Movie Theater	086	0000	086_0000	29.2	23.2	47.5
Nursing Home	006	0000	006_0000	18.1	44.8	37.3
Office	043	0001	043_0001	25.4	42.1	32.4
Office	043	0204	043_0204	17.8	42.3	39.9
Office	043	0510	043_0510	20.7	44.5	36.5
Office	043	1120	043_1120	28.4	39.1	29.8
Police Stations	055	0000	055_0000	11.4	32.4	56.1
Restaurant	087	0000	087_0000	22	49.4	28.7
Restaurant, Fast Food	089	0000	089_0000	20.7	34.5	44.8
School, Elementary	053	0000	053_0000	23.6	40.3	36.1
School, High	053	0203	053_0203	24.9	37.3	38.1
School, Jr High	053	0203	053_0203	25	35.1	40
School, Vocational	053	0000	053_0000	21	39.2	39.9
Store, Convenience	011	0000	011_0000	29.3	41.1	29.5
Store, Department	013	0001	013_0001	42.8	25.3	31.9
Store, Department	013	0003	013_0003	28.5	30.9	40.6
Store, Retail	015	0000	015_0000	28.8	39.7	31.5
Supermarket	017	0000	017_0000	28.6	33.8	37.6
Warehouse	023	0000	023_0000	48.2	26.7	25
Warehouse, Mini	025	0000	025_0000	33.7	24.9	41.4

3.5.3.1. Recording Construction Assembly Costs for Model Buildings

The specific occupancy and number of stories together generated a code that identified the Means model building type. The buildings from the Tax Records were then reconciled to follow this format. Table 3.22 also shows the nomenclature and code for recording the construction assembly cost breakdowns.

3.5.3.2. Estimating Structural, Acceleration- and Drift-Sensitive Nonstructural Costs

Again, once the model identifiers have been created, the respective assemblages (percent cost of total replacement value) are recorded with them, and grouped into structural, acceleration-sensitive and drift-sensitive components. These percentages are then multiplied by the estimated replacement cost of the building.

3.6. Estimating Content Value

The literature review identified the paucity of content value loss models for buildings and further, that claim and valuation information that exists is proprietary in nature, and rests in the private sector. The lack of available data forced the estimation model for content value to follow the existing HAZUS MR-3 model that estimates content value as a function of replacement cost and specific occupancy. Table 2.7 in the literature review highlights the percent of replacement costs by specific occupancy for estimating content value.

Chapter 4 . RESULTS AND DISCUSSION

Following the format adopted in the literature review and methodology sections, this section details the results of the various models and includes discussion of the results. The chapter begins with a discussion of structure type classification from the multinomial logistic regression and the ANN models. The following section details the results of the various subroutines for preprocessing shapes and the results of implementing the classification algorithm separately for manually digitized and automatically extracted building footprints. The next section details the results of the estimation process for replacement costs, the associated structural and nonstructural costs and the content value. The chapter concludes with a note on the creation of an integrated building inventory for Shelby County, Tennessee, using the methods described in the dissertation that may be used for loss estimation and risk assessment modeling. Appendix A shows the integrated results of the various models implemented for the MTB in the dissertation, with tabulated summaries of the Shelby County building inventory.

4.1. Structure Type from Multinomial Logistic Regression

4.1.1. Multinomial Logistic Regression Model Specification

Several specifications for the multinomial logistic regression were attempted with the input variables specified as Number of Stories, Year built, Area, Occupancy, Fire rating and Historic zone. The problem lay in the fact that for some structure types, there were no values for the input data. There were two reasons for this. First, the external wall attribute from the Tax Records were used to define Concrete tilt-up structures, and second, there were some structure-occupancy combinations that had no data. For instance, there were no Wood structures for the IND1 (heavy industrial structures), and

the specification of the multinomial logistic regression was not able to estimate parameters for such cases. In fact, 5 IND1 structures were deliberately changed to Wood from Concrete in order for the model to estimate parameters. Parking structures (COM10) did not figure in any structure type except for Precast Concrete in the sample dataset. Additionally, the full specification of 21 input variables was not used and the inputs consisted of number of stories, year built, area, occupancy (defined at 8 levels) and fire rating (defined at 3 levels). Further, in the interests of model tractability and parsimony, the structure types were collapsed into four categories including Concrete (pooled Concrete moment frame, Precast Concrete and Concrete Tilt-ups), Steel (pooled Steel frame, Light Metal frame and Reinforced Masonry), Unreinforced Masonry and Wood. Table 4.1 details the variables used in the multinomial logistic regression. The dependent variable is highlighted in the table. Overall, there were 209 Concrete, 612 Steel, 303 Unreinforced Masonry and 707 Wood structure types in the sample.

Table 4.1 -- Variable specification for the Multinomial Logistic Regression

Variable	Type	Values	Description
STORIES	numeric		Number of stories
SQ_FEET	numeric		Area of building in square feet
YEAR_BLT	numeric		Year of construction
FIRE_RTG	symbolic	FP	Fire Proof (reference)
		FR	Fire Resistant
		WJ	Wood Joists
OCC_TYPE	symbolic	COM1	Retail Trade (reference)
		COM2	Wholesale Trade
		COM4	Commercial Office (includes parking structures)
		COM5	Banks
		COM8	Restaurants and Bars
		IND1	Heavy Industrial
		IND2	Light Industrial (includes COM3 structures)
		RES3	Multi-family residential (includes group housing and hotels)
STR_TYPE	symbolic	C	Concrete (includes Concrete, Precast and Tilt-ups)
		S1	Steel (includes Steel, Light Metal, Reinforced Masonry)
		URM	Unreinforced Masonry
		W	Wood Frame (base outcome)

4.1.2. Model Performance

Table 4.2 specifies the parameter estimates for the structure types Concrete, Steel and Unreinforced Masonry, relative to the base outcome specified as Wood. The statistically significant parameters are highlighted in the table. The following text describes some of the relationships to highlight the consistencies in the relationship between the inputs and structure types.

For Concrete, square footage, wholesale trade, commercial office, multi-family residential and wood joist fire rating were statistically significant and in the expected directions. Thus, relative to the Wood structure type, as the square footage increases, the likelihood of a steel structure being used increases, as seen in the positive coefficient for square footage. The magnitude is rather small, because the square foot is a miniscule measure. Again, the likelihood of the steel structure relative to wood is reduced when the occupancy changes to multi-family residential (RES3) since most multi-family residential structures are built of wood.

For Steel, square footage, year built, commercial office, restaurants, heavy industrial, multi-family residential and fire resistant fire rating were statistically significant and in the expected directions. Just like concrete, the likelihood of steel relative to wood increases with increase in area. Like concrete, the likelihood of steel increases when the occupancy changes to commercial office (COM4) and reduces when the occupancy changes to multifamily residential (RES3) or restaurants (COM8). The heavy industrial category (IND1) is surprising, since one would expect the likelihood of steel to increase when the occupancy changes to industrial. The most likely reason is that there were no structures of Wood for heavy industrial occupancies, and 5 instances had been artificially changed for the parameter estimation. As expected, the likelihood of steel relative to wood increases when the fire rating changes to Fire Resistant.

Table 4.2 -- Parameter estimates from the Multinomial Logistic Regression

Structure Type	Input Variable	Factor Level	Coefficient	Standard Error	z-score	p-value	
Concrete	STORIES		0.2572710	0.2226969	1.16	0.248	
	SQ_FEET		0.0000266	0.0000126	2.12	0.034	
	YEAR_BLT		-0.0110742	0.0093494	-1.18	0.236	
	OCCUPANCY	COM2		1.4145040	0.6516831	2.17	0.030
		COM4		2.5708550	0.6495749	3.96	0.000
		COM5		1.6193720	0.9677616	1.67	0.094
		COM8		0.3164239	1.1907780	0.27	0.790
		IND1		-1.2324400	1.0018690	-1.23	0.219
		IND2		-0.6785393	0.7932320	-0.86	0.392
	FIRE RATING	RES3		-2.8477950	0.8361752	-3.41	0.001
		FIRE RESISTANT		-0.7231985	0.7372789	-0.98	0.327
	WOOD JOISTS		-7.4972780	0.8077755	-9.28	0.000	
CONSTANT			23.5600200	18.4574600	1.28	0.202	
Steel	STORIES		0.4156967	0.2167163	1.92	0.055	
	SQ_FEET		0.0000277	0.0000125	2.22	0.027	
	YEAR_BLT		0.0500834	0.0082343	6.08	0.000	
	OCCUPANCY	COM2		0.2412201	0.3853707	0.63	0.531
		COM4		-1.2330090	0.4262193	-2.89	0.004
		COM5		-1.2467860	0.7203546	-1.73	0.083
		COM8		-1.0649170	0.4140644	-2.57	0.010
		IND1		-1.9113810	0.9591353	-1.99	0.046
		IND2		-0.7143923	0.5019463	-1.42	0.155
	FIRE RATING	RES3		-4.5062050	0.4776803	-9.43	0.000
		FIRE RESISTANT		5.5135260	0.8225829	6.70	0.000
	WOOD JOISTS		-0.5374209	0.7321754	-0.73	0.463	
CONSTANT			-99.5938500	16.3108800	-6.11	0.000	
URM	STORIES		0.0093509	0.2283066	0.04	0.967	
	SQ_FEET		-0.0000318	0.0000149	-2.13	0.033	
	YEAR_BLT		-0.0864919	0.0073094	-11.83	0.000	
	OCCUPANCY	COM2		-0.1023053	0.4041643	-0.25	0.800
		COM4		-0.8415066	0.4267095	-1.97	0.049
		COM5		-1.0265950	0.8472171	-1.21	0.226
		COM8		-1.3625580	0.5045133	-2.70	0.007
		IND1		-2.4082760	0.9746367	-2.47	0.013
		IND2		-0.5434886	0.4815068	-1.13	0.259
	FIRE RATING	RES3		-3.9726640	0.4136092	-9.60	0.000
		FIRE RESISTANT		5.4621640	1.0763380	5.07	0.000
	WOOD JOISTS		1.1169530	1.0055040	1.11	0.267	
CONSTANT			168.8840000	14.3731100	11.75	0.000	

WOOD is the base outcome

significant at 95% confidence
 significant at 99% confidence

For Unreinforced Masonry, area, height, year built, commercial office, restaurants, heavy industrial and multi-family residential and fire resistant fire rating were

all statistically significant. The likelihood of unreinforced masonry decreases relative to wood as the square footage or year built or height increase. URM buildings were prohibited by the building code after 1974 and tend to be small 1 or 2 story structures. Similarly, the likelihood of unreinforced masonry reduces when the occupancy changes to commercial office (COM4), restaurants (COM8), heavy industrial (IND1, which is suspect), and multi-family residential (RES3).

Thus, the majority of the relationships between the inputs and structure classes are plausible and follow logical trends in construction, based on combinations of the input variables.

4.1.3. Relationships between Inputs and Structure Classes

Tables B.1 through B.6 in Appendix B detail the influence of the input variables on the factor change in the odds of structure type alternatives, specifically listing the odds comparing pairs of alternative structure types. Rows in these tables essentially comment on the input variable in the following manner – a 1 unit increase (or a factor level change from the reference level to the input variable level) results in the increased (or decreased) odds of having structure type alternative 1 relative to structure type alternative 2 by a factor of the factor change in odds (specified in the “exp(b)” column). The influence of a number of variables on structure type pairs was found to be statistically significant. This section lists some of the significant relationships.

Number of stories was found to have a significant influence on the following pairs Steel to Concrete and Steel to URM. While the magnitude for the Steel to Concrete pair was negligible, a 1-story increase results in the increased odds of having Steel relative to URM by a factor of 1.5 – this is expected, because one would expect that as the

number of stories increases, the likelihood of URM as the structure type should decrease.

The building area's influence was also consistent with logical expectations of construction practices. For instance, in the structure type pairs Concrete to Wood, Concrete to URM, Steel to Wood and Steel to URM, a 1 standard deviation increase (about 45,600 sq. ft.) in area results in the increased odds of Concrete or Steel outcomes relative to Wood and URM by a factor of approximately 3 or 15 respectively. Again, these results are expected, since the likelihood of steel and concrete would tend to dominate larger-area structures. Similarly, a 1 standard deviation increase in area results in the decreased odds of URM relative to Wood by a factor of 0.2346.

The year of construction had a lot of explanatory power and was statistically significant for almost all structure type pairings. The relationship between Steel and Concrete was statistically significant, but negligible in magnitude. A 1 standard deviation increase (about 23 years) in year of construction results in the increased odds of Concrete and Steel relative to URM by factors of 5.65 and 23 respectively. Similarly, 1 standard deviation increase in year of construction results in the increased odds of Wood relative to URM by factors of 7.3. Again, this is expected, since URM was prohibited by the building code since 1974, and URM construction had been reducing as the decades advanced.

Figure 4.1 shows some of these results graphically for Number of stories, Area and Year of construction, with structure types located along the number lines – all results are shown relative to Wood. The figure shows the factor change along the top number line and the coefficient change along the bottom number line for each variable. For instance, note that the factor change between Steel (S) and URM (U) for Stories is about 1.5 from the top number line and this corresponds exactly with the influence of the

number of stories on the Steel to URM structure pairing described earlier in this section. Note also the influence of square footage on the factor change of URM relative to Wood – relative to Wood (located at 1), URM is located at .23, and this corresponds exactly with the influence of area on the Wood to URM structure type pairing described earlier.

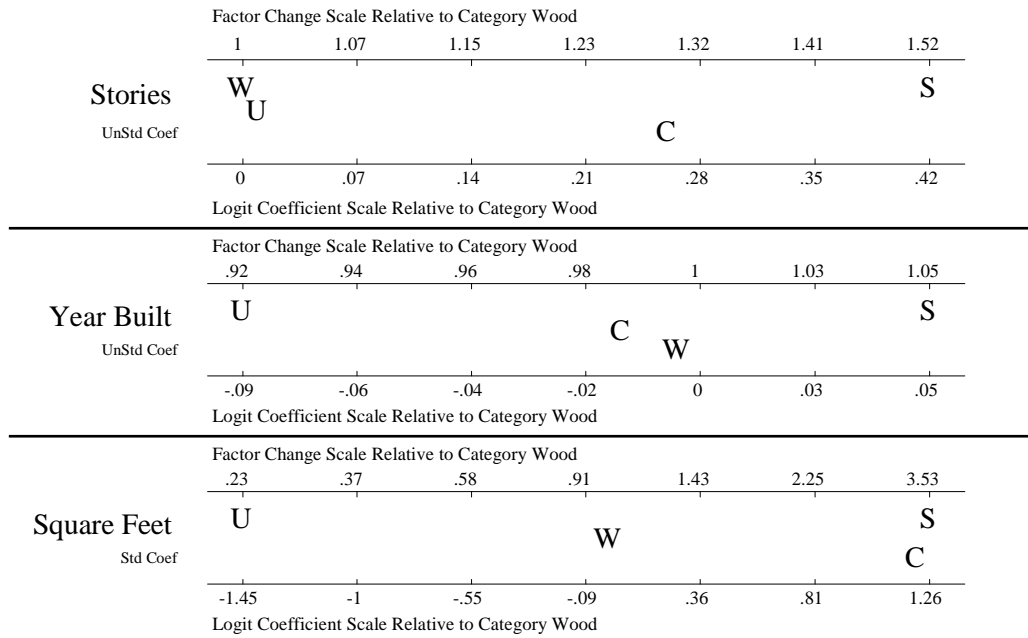


Figure 4.1 -- Influence of covariates on structure type

Similarly, Figures 4.3 and 4.4 graphically show the influence of the occupancy levels on structure type relative to wood and Figure 4.2 illustrates the influence of the fire rating variable on structure type.

		Factor Change Scale Relative to Category Wood						
		0	0	.01	.04	.17	.73	3.06
Wood Joists 0/1	C						S	U
			W				W	
		-7.5	-6.06	-4.63	-3.19	-1.75	-.32	1.12
		Logit Coefficient Scale Relative to Category Wood						
		Factor Change Scale Relative to Category Wood						
		.49	1.37	3.88	10.97	31.02	87.71	248.02
Fire Resistant 0/1	C							S
			W					U
		-.72	.32	1.36	2.4	3.43	4.47	5.51
		Logit Coefficient Scale Relative to Category Wood						

Figure 4.2 -- Influence of fire rating on structure type

		Factor Change Scale Relative to Category Wood						
		.9	1.16	1.5	1.93	2.48	3.2	4.11
Wholesale Trade 0/1	U							C
			W					
				S				
		-.1	.15	.4	.66	.91	1.16	1.41
		Logit Coefficient Scale Relative to Category Wood						
		Factor Change Scale Relative to Category Wood						
		.29	.55	1.04	1.95	3.68	6.94	13.08
Commercial Office 0/1	S							C
			U					
				W				
		-1.23	-.6	.03	.67	1.3	1.94	2.57
		Logit Coefficient Scale Relative to Category Wood						
		Factor Change Scale Relative to Category Wood						
		.29	.46	.75	1.2	1.94	3.13	5.05
Banks 0/1	S							C
			U					
				W				
		-1.25	-.77	-.29	.19	.66	1.14	1.62
		Logit Coefficient Scale Relative to Category Wood						

Figure 4.3 -- Influence of occupancy on structure type (part 1)

	Factor Change Scale Relative to Category Wood						
	.26	.34	.45	.59	.78	1.04	1.37
Food and Entertainment 0/1	U	S				W	C
	-1.36	-1.08	-.8	-.52	-.24	.04	.32
	Logit Coefficient Scale Relative to Category Wood						
	Factor Change Scale Relative to Category Wood						
	.09	.13	.2	.3	.45	.67	1
Heavy Industrial 0/1	U	S		C			W
	-2.41	-2.01	-1.61	-1.2	-.8	-.4	0
	Logit Coefficient Scale Relative to Category Wood						
	Factor Change Scale Relative to Category Wood						
	.01	.02	.05	.11	.22	.47	1
Multi-family Residential 0/1	S	U		C			W
	-4.51	-3.76	-3	-2.25	-1.5	-.75	0
	Logit Coefficient Scale Relative to Category Wood						

Figure 4.4 -- Influence of occupancy on structure type (part 2)

Note that occupancy levels COM2 (Wholesale Trade) and COM4 (Commercial Office) have considerable explanatory power for the use of Steel or Concrete buildings relative to Wood or URM. Occupancy level COM8 (Food and Entertainment) explains the increased likelihood of Wood structures over all other structure types. IND1 (Heavy Industrial) shows inconsistencies as described earlier. For instance, an occupancy level change to IND1 results in the increased odds of Wood relative to Steel and URM by factors of 6.8 and 11.1 respectively, when most Heavy Industrial structures tend to be built of URM (if they are old) or Concrete or Steel. The RES3 (Multi-family residential) occupancy level, when realized, increases the odds of Wood over Concrete, Steel and URM by factors of 17.25, 90.58 and 53.13 respectively. Similarly, Concrete was more likely than Steel by a factor of 5.25. Most multi-family apartments tend to be built of Wood, or Concrete if they are taller.

The fire rating variables also had considerable explanatory power. Fire Resistant ratings resulted in the increased odds of both Steel and URM over Concrete and Wood. The Wood Joist description obviously specified the increased odds of Wood over all other structure types.

4.2. Structure Type from Neural Networks

Based on the descriptions of the various topologies in the methodology section, five ANNs were specified including the MLP, GFF, MNN, RBF and SVM models. One additional exercise was conducted for a single hidden layer MLP network, using the final specifications from the multinomial logistic regression for comparing the results from the parametric logistic approach and the semi-parametric neural network method. The comparisons are described in Section 4.3. The following section compares the results of the five models, in terms of training and testing performance.

In all the specifications, of the 1831 sample buildings, 1284 samples (70%) were used for training the ANNs, 274 (15%) for cross-validation and 273 (15%) for testing. The cross-validation dataset was used for generalization and stopping the training process at the point where the ANN just begins to memorize (or overtrain) the data, following the process described in the literature review, Section 2.3.7.5. The testing data is used to evaluate the performance of the ANN for its classification performance against unseen data.

Table 4.3 shows the variables used in the ANN evaluation models, with the dependent variable (STR_TYPE with 8 categories) highlighted. For the purposes of comparing the ANN and multinomial specifications, the variables were the same as specified in Table 4.1.

Table 4.3 -- Variable specification for the ANN models

Variable	Type	Values	Description
STORIES	numeric		Number of stories
SQ_FEET	numeric		Area of building in square feet
YEAR_BLT	numeric		Year of construction
HIST_ZONE	symbolic	T	In the historic zone
		F	Not in the historic zone
FIRE_RTG	symbolic	FP	Fire proof
		FR	Fire Resistant
		ES	Engineered Steel
		WF	Wood Joists
		ES1	Pseudo-code for Concrete Tilt-up Wall
		ES2	Pseudo-code for Light Metal Wall
OCC_TYPE	symbolic	COM1	Retail Trade
		COM2	Wholesale Trade
		COM4	Commercial Office
		COM5	Banks
		COM8	Restaurants and Bars
		COM10	Parking structures
		IND1	Heavy Industrial
		IND2	Light Industrial (includes COM3 structures)
		RES3	Multi-family residential (includes group housing and hotels)
STR_TYPE	symbolic	C	Concrete Moment Frame (separated later to C1 and C2)
		PC1	Concrete tilt-up
		PC2	Precast Concrete
		S1	Steel Moment Frame
		S3	Light Metal Frame
		RM	Reinforced Masonry
		URM	Unreinforced Masonry
		W	Wood Frame

The OCC_TYPE variable (Building occupancy) was collapsed to include fewer categories – COM3 structures like machine shops and automobile service garages were grouped with IND2 since they resembled light industrial structures, based on preliminary examinations of their exterior wall, square footage, height and fire rating characteristics. Police and fire stations, medical offices and hospitals were grouped with COM4 (Commercial Offices). Theaters and auditoriums were grouped with IND1 (Heavy Industrial). Hotels, motels and group housing occupancies were grouped with RES3 (Multi-family residential). The FIRE_RTG variable (Structure Fire Rating) was modified

to accommodate light metal and concrete tilt-up external walls extracted from the exterior wall codes in the Tax Records, in order to prevent misclassifications between these categories and the Concrete structure type. The Concrete structure type included Concrete Moment Frame and Concrete Frame with Concrete Shear Wall categories, which would be subdivided after the ANN classification, based on height and occupancy.

4.2.1. Model Performance Evaluations

As specified in the literature review and methodology sections, there are few parametric measures for classification performance evaluation in ANNs. ANN classification performance is usually evaluated on the basis of correct and incorrect classifications implemented in the form of a confusion matrix. This section describes the accuracy of classification for the five ANN specifications and analyzes the potential sources for misclassification by structure type category.

4.2.1.1. Interpreting the Confusion Matrix

The confusion matrix, a matrix of observed against predicted classifications, is an extremely useful device to evaluate the performance of different classifiers and contains a lot of information. Interpreting the confusion matrix directly is cognitively challenging, especially for five different models. Consequently, four different performance evaluation measures are presented for each of the training and testing exercises – the first measure is the confusion matrix of raw counts, the second is a matrix showing accuracy percentages for the different structure type categories and the overall model accuracy, and is used to analyze where and why the model performs well. The third and fourth measures are matrices showing the percentage of structure types not recognized (errors) and structure types incorrectly predicted (misclassifications) respectively and is used to analyze the nature of the incorrect classifications in each model. Tables 4.4 and

4.5 list the raw count confusion matrices for the training and testing components of the five models respectively.

The lowermost row in the confusion matrix table specifies the number of known samples in each structure type column. The training sample (Table 4.4) had 11 of “PC2”, 87 of “C”, 129 of “S3”, 126 of “RM”, 172 of “S1”, 42 of “PC1”, 508 of “W” and 209 of “URM”, totaling to 1284 exemplars. The testing sample (Table 4.5) had 2 of “PC2”, 22 of “C”, 22 of “S3”, 32 of “RM”, 37 of “S1”, 9 of “PC1”, 107 of “W” and 42 of “URM”, totaling to 273 exemplars.

Predicted classes are described in rows, while the columns sum to the number of samples in each structure type. The intersecting cells between the corresponding observed and predicted classes (the diagonals) specify the accurate classifications. The aim of the classifier mechanism is to make these diagonal cells match the number of samples (well, not really, because, if that happens, one should suspect the model -- perfection is hard to achieve!).

Table 4.4 -- Training performance evaluation using a confusion matrix (counts)

Model	Code	Predicted								Predicted Totals
		PC2	C	S3	RM	S1	PC1	W	URM	
Multilayer Perceptron Neural Network	PC2	11	-	-	-	-	-	-	-	11
	C	-	66	-	-	7	-	1	7	81
	S3	-	-	126	2	9	-	-	-	137
	RM	-	2	2	74	9	-	19	6	112
	S1	-	11	1	20	123	-	1	7	163
	PC1	-	-	-	-	-	42	-	-	42
	W	-	-	-	30	9	-	468	18	525
	URM	-	8	-	-	15	-	19	171	213
Generalized Feed Forward Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM	Totals
	PC2	10	-	-	-	-	-	-	-	10
	C	-	64	-	-	7	-	1	5	77
	S3	-	-	126	1	10	-	-	-	137
	RM	-	1	2	78	8	-	19	5	113
	S1	1	13	1	22	120	-	1	12	170
	PC1	-	-	-	-	-	42	-	-	42
	W	-	1	-	25	9	-	467	17	519
	URM	-	8	-	-	18	-	20	170	216
Modular Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM	Totals
	PC2	11	-	-	-	-	-	-	-	11
	C	-	63	-	-	9	-	1	5	78
	S3	-	-	126	2	10	-	-	-	138
	RM	-	1	2	73	8	-	19	4	107
	S1	-	14	1	21	122	-	1	12	171
	PC1	-	-	-	-	-	42	-	-	42
	W	-	1	-	30	9	-	467	18	525
URM	-	8	-	-	14	-	20	170	212	
Radial Basis Functions Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM	Totals
	PC2	-	-	-	-	-	-	-	-	-
	C	-	-	-	-	-	-	-	-	-
	S3	-	-	127	2	28	33	-	-	190
	RM	-	-	-	-	-	-	-	-	-
	S1	-	38	-	75	120	6	2	73	314
	PC1	-	-	-	-	-	-	-	-	-
	W	11	20	2	49	16	2	498	41	639
URM	-	29	-	-	8	1	8	95	141	
Support Vector Machine Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM	Totals
	PC2	11	-	-	-	-	-	-	-	11
	C	-	87	-	-	-	-	-	-	87
	S3	-	-	127	-	-	-	-	-	127
	RM	-	-	-	126	-	-	-	-	126
	S1	-	-	-	-	172	-	-	-	172
	PC1	-	-	-	-	-	42	-	-	42
	W	-	-	2	-	-	-	507	-	509
URM	-	-	-	-	-	-	1	209	210	
Sample Totals		11	87	129	126	172	42	508	209	1,284

Table 4.5 -- Testing performance evaluation using a confusion matrix (counts)

Model	Code	Predicted								Predicted Totals
		PC2	C	S3	RM	S1	PC1	W	URM	
Multilayer Perceptron Neural Network	PC2	2	-	-	-	-	-	-	-	2
	C	-	16	-	-	1	-	-	3	20
	S3	-	-	22	-	2	-	-	-	24
	RM	-	-	-	21	2	-	6	-	29
	S1	-	2	-	6	28	-	-	-	36
	PC1	-	-	-	-	-	9	-	-	9
	W	-	1	-	4	-	-	96	3	104
	URM	-	3	-	1	4	-	5	36	49
Generalized Feed Forward Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM	Totals
	PC2	2	-	-	-	-	-	-	-	2
	C	-	17	-	-	1	-	-	2	20
	S3	-	-	21	-	2	-	-	-	23
	RM	-	-	-	22	2	-	6	1	31
	S1	-	2	1	7	27	-	-	1	38
	PC1	-	-	-	-	-	9	-	-	9
	W	-	1	-	2	-	-	97	3	103
	URM	-	2	-	1	5	-	4	35	47
Modular Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM	Totals
	PC2	2	-	-	-	-	-	-	-	2
	C	-	14	-	-	1	-	-	2	17
	S3	-	-	22	-	2	-	-	-	24
	RM	-	-	-	22	2	-	5	-	29
	S1	-	3	-	6	27	-	-	1	37
	PC1	-	-	-	-	-	9	-	-	9
	W	-	1	-	4	-	-	98	3	106
URM	-	4	-	-	5	-	4	36	49	
Radial Basis Functions Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM	Totals
	PC2	-	-	-	-	-	-	-	-	-
	C	-	-	-	-	-	-	-	-	-
	S3	-	-	22	-	10	8	-	-	40
	RM	-	-	-	-	-	-	-	-	-
	S1	-	11	-	23	22	1	1	19	77
	PC1	-	-	-	-	-	-	-	-	-
	W	1	5	-	9	1	-	103	7	126
URM	1	6	-	-	4	-	3	16	30	
Support Vector Machine Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM	Totals
	PC2	1	-	-	-	-	-	-	-	1
	C	-	16	-	-	3	-	-	2	21
	S3	-	-	18	-	2	-	-	-	20
	RM	-	1	1	21	1	-	2	-	26
	S1	-	-	3	8	28	-	1	5	45
	PC1	-	-	-	-	-	9	-	-	9
	W	1	1	-	1	-	-	100	3	106
URM	-	4	-	2	3	-	4	32	45	
Sample Totals		2	22	22	32	37	9	107	42	273

Let us examine the confusion matrices presented in Table 4.4 in greater detail. Note that the confusion matrix is not symmetric, because the model is not restricted to limiting the number of predictions of each class to the corresponding number of samples. For instance, in the first matrix (the MLP model) there are 87 samples of class “C” with 66 predicted correctly. Of the 87 samples of class “C”, reading the column information, 2 were wrongly classified as “RM”, 11 as “S1” and 8 as “URM”. Reading the row information, the model predicts only 81 for class “C”, of which 66 were classified correctly, 7 “S1”, 1 “W” and 7 “URM” buildings were misclassified as “C”. Effectively, the columns indicate the number of correct and wrong classifications (the model did not recognize the desired class) and the rows indicate the number of correct classifications and misclassifications (the model did not predict the desired class). Note that the classification errors along the columns and the rows are disjoint sets – this is not apparent from the confusion matrix, and requires an analysis of the raw output from the ANN model.

4.2.1.2. Comparing Accuracy of Classification

Tables 4.6 and 4.7 list the percentage of accurate classifications for the training and testing components of each of the five models.

Note that the perceptron-based models, the MLP, GFF and MNN, performed consistently and had overall accuracy in classification in the mid 80%, in both training and testing components. The accuracy was averaged over 5000 complete iterations trained 3 times in each case.

Table 4.6 -- ANN training percent accuracy of Structure type classification

Model	Code	Predicted								Overall Accuracy
		PC2	C	S3	RM	S1	PC1	W	URM	
Multilayer Perceptron Neural Network	PC2	100%	-	-	-	-	-	-	-	84.19%
	C	-	75.86%	-	-	-	-	-	-	
	S3	-	-	97.67%	-	-	-	-	-	
	RM	-	-	-	58.73%	-	-	-	-	
	S1	-	-	-	-	71.51%	-	-	-	
	PC1	-	-	-	-	-	100%	-	-	
	W	-	-	-	-	-	-	92.13%	-	
	URM	-	-	-	-	-	-	-	81.82%	
Generalized Feed Forward Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM	Totals
	PC2	90.91%	-	-	-	-	-	-	-	83.88%
	C	-	73.56%	-	-	-	-	-	-	
	S3	-	-	97.67%	-	-	-	-	-	
	RM	-	-	-	61.90%	-	-	-	-	
	S1	-	-	-	-	69.77%	-	-	-	
	PC1	-	-	-	-	-	100%	-	-	
	W	-	-	-	-	-	-	91.93%	-	
URM	-	-	-	-	-	-	-	81.34%		
Modular Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM	Totals
	PC2	100%	-	-	-	-	-	-	-	83.64%
	C	-	72.41%	-	-	-	-	-	-	
	S3	-	-	97.67%	-	-	-	-	-	
	RM	-	-	-	57.94%	-	-	-	-	
	S1	-	-	-	-	70.93%	-	-	-	
	PC1	-	-	-	-	-	100%	-	-	
	W	-	-	-	-	-	-	91.93%	-	
URM	-	-	-	-	-	-	-	81.34%		
Radial Basis Functions Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM	Totals
	PC2	0%	-	-	-	-	-	-	-	65.42%
	C	-	0%	-	-	-	-	-	-	
	S3	-	-	98.45%	-	-	-	-	-	
	RM	-	-	-	0%	-	-	-	-	
	S1	-	-	-	-	69.77%	-	-	-	
	PC1	-	-	-	-	-	0%	-	-	
	W	-	-	-	-	-	-	98.03%	-	
URM	-	-	-	-	-	-	-	45.45%		
Support Vector Machines Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM	Totals
	PC2	100%	-	-	-	-	-	-	-	99.77%
	C	-	100%	-	-	-	-	-	-	
	S3	-	-	98.45%	-	-	-	-	-	
	RM	-	-	-	100%	-	-	-	-	
	S1	-	-	-	-	100%	-	-	-	
	PC1	-	-	-	-	-	100%	-	-	
	W	-	-	-	-	-	-	99.80%	-	
URM	-	-	-	-	-	-	-	100%		

Table 4.7 -- ANN testing percent accuracy of Structure type classification

Model	Code	Predicted								Overall Accuracy
		PC2	C	S3	RM	S1	PC1	W	URM	
Multilayer Perceptron Neural Network	PC2	100%	-	-	-	-	-	-	-	84.25%
	C	-	72.73%	-	-	-	-	-	-	
	S3	-	-	100%	-	-	-	-	-	
	RM	-	-	-	65.63%	-	-	-	-	
	S1	-	-	-	-	75.68%	-	-	-	
	PC1	-	-	-	-	-	100%	-	-	
	W	-	-	-	-	-	-	89.72%	-	
	URM	-	-	-	-	-	-	-	85.71%	
Generalized Feed Forward Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM	Totals
	PC2	100%	-	-	-	-	-	-	-	84.25%
	C	-	77.27%	-	-	-	-	-	-	
	S3	-	-	95.45%	-	-	-	-	-	
	RM	-	-	-	68.75%	-	-	-	-	
	S1	-	-	-	-	72.97%	-	-	-	
	PC1	-	-	-	-	-	100%	-	-	
	W	-	-	-	-	-	-	90.65%	-	
	URM	-	-	-	-	-	-	-	83.33%	
Modular Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM	Totals
	PC2	100%	-	-	-	-	-	-	-	84.25%
	C	-	63.64%	-	-	-	-	-	-	
	S3	-	-	100%	-	-	-	-	-	
	RM	-	-	-	68.75%	-	-	-	-	
	S1	-	-	-	-	72.97%	-	-	-	
	PC1	-	-	-	-	-	100%	-	-	
	W	-	-	-	-	-	-	91.59%	-	
URM	-	-	-	-	-	-	-	85.71%		
Radial Basis Functions Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM	Totals
	PC2	0%	-	-	-	-	-	-	-	59.71%
	C	-	0%	-	-	-	-	-	-	
	S3	-	-	100%	-	-	-	-	-	
	RM	-	-	-	0%	-	-	-	-	
	S1	-	-	-	-	59.46%	-	-	-	
	PC1	-	-	-	-	-	0%	-	-	
	W	-	-	-	-	-	-	96.26%	-	
URM	-	-	-	-	-	-	-	38.10%		
Support Vector Machines Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM	Totals
	PC2	50.00%	-	-	-	-	-	-	-	82.42%
	C	-	72.73%	-	-	-	-	-	-	
	S3	-	-	81.82%	-	-	-	-	-	
	RM	-	-	-	65.63%	-	-	-	-	
	S1	-	-	-	-	75.68%	-	-	-	
	PC1	-	-	-	-	-	100%	-	-	
	W	-	-	-	-	-	-	93.46%	-	
URM	-	-	-	-	-	-	-	76.19%		

The RBF performed relatively poorly – the RBF ANN requires a priori specification about the number of Gaussians that would be fitted over the input space.

Too few Gaussians and the model's classification accuracy suffers; too many and despite long training times, there is a danger of overtraining with poor accuracy in classification of unseen data. This a priori specification is arrived at by trial and error, and while there are methods that enable educated guesses regarding the number of Gaussians, the gains over the relatively accurate perceptron-based models would not be substantial. The best results for 70 Gaussians are reported here. Note that the training accuracy for the RBF ANN is about 65% and the testing accuracy drops to 60%. The SVM ANN had near perfect classification in the training component, but dropped to about 82% in the testing dataset. The near perfect training result suggests overtraining, and essentially proves Cover's theorem (1965) that increasing the inputs artificially to a higher dimension space will enable linear classification – the classification results are strongly suspected to be spurious. Further, the SVM specification is generally used with data that has many input variables (covariates and factor levels) with few exemplars, and the thumb rule in neural literature suggests that if the number of exemplars exceeds 1000, the results may be spurious (Principe et al. 2000). The best models are generally those that exhibit consistent classification performances in training and testing. Figure 4.5 shows overall accuracy in training and testing for the five models.

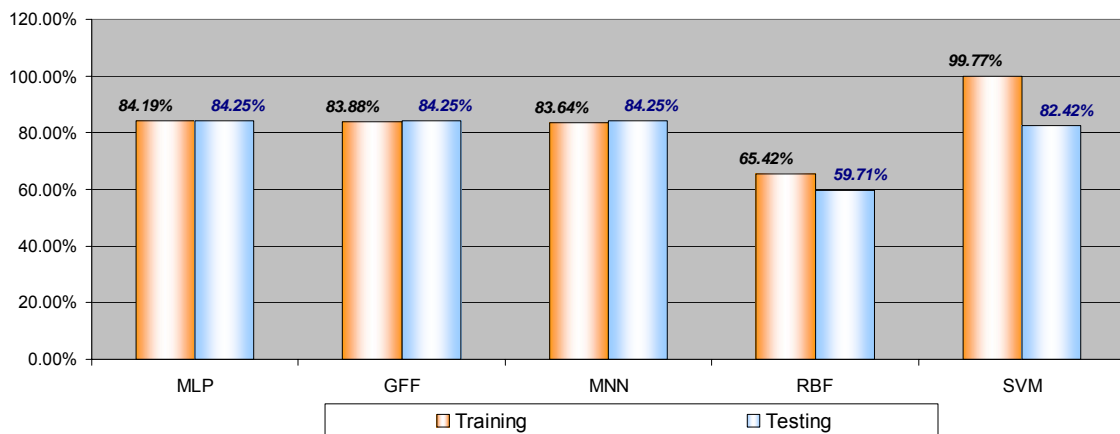


Figure 4.5 -- Structure type classification accuracy by ANN model type

4.2.1.3. Analysis of Misclassifications in the Models

Table 4.8 shows a summary of the cases where the desired structure type was not recognized (column errors in the confusion matrix) for training and testing components respectively in the five ANN models. Tables 4.9 and 4.10 show details by ANN model type for the training and testing components respectively, where the structure type was not recognized and therefore classified wrongly. These tables should be analyzed along columns. Cases of wrong classifications greater than 10% have been highlighted in both the detail tables. Diagonals highlighted in blue show accurate classifications. For the total structures predicted as the structure type described in the column, the diagonals highlight model accuracy. In other words, for the training case of the MLP ANN, there were 87 samples of “C” (see Table 4.4). Of these 87 “C” structures, 66 (or 75.86% as highlighted in the diagonal for the MLP ANN in Table 4.9) were recognized accurately, while 2 (2.3%) were wrongly classified as “RM”, 11 (12.64%) were predicted as “S1” and 8 (9.2%) were predicted as “URM”. A total of 21 (24.14%) “C” structures were not recognized and erroneously classified into other classes, which is listed in Table 4.8 for the MLP ANN training row.

Table 4.8 -- Summary of errors (Structure type not recognized) by ANN model

Model	Part	C	PC1	PC2	S1	S3	RM	URM	W	Overall
MLP	Training	24.14%	0%	0%	28.49%	2.33%	41.27%	18.18%	7.87%	15.81%
	Testing	27.27%	0%	0%	24.32%	0%	34.38%	14.29%	10.28%	15.75%
GFF	Training	26.44%	0%	9.09%	30.23%	2.33%	38.10%	18.66%	8.07%	16.12%
	Testing	22.73%	0%	0%	27.03%	4.55%	31.25%	16.67%	9.35%	15.75%
MNN	Training	27.59%	0%	0%	29.07%	2.33%	42.06%	18.66%	8.07%	16.36%
	Testing	36.36%	0%	0%	27.03%	0%	31.25%	14.29%	8.41%	15.75%
RBF	Training	100%	100%	100%	30.23%	1.55%	100%	54.55%	1.97%	34.58%
	Testing	100%	100%	100%	40.54%	0%	100%	61.90%	3.74%	40.29%
SVM	Training	0%	0%	0%	0%	1.55%	0%	0%	0.20%	0.23%
	Testing	27.27%	0%	50.00%	24.32%	18.18%	34.38%	23.81%	6.54%	17.58%

Table 4.9 – ANN training errors (Structure type not recognized) by model

Model	Code	Predicted							
		PC2	C	S3	RM	S1	PC1	W	URM
Multilayer Perceptron Neural Network	PC2	100%	-	-	-	-	-	-	-
	C	-	75.86%	-	-	4.07%	-	0.20%	3.35%
	S3	-	-	97.67%	1.59%	5.23%	-	-	-
	RM	-	2.30%	1.55%	58.73%	5.23%	-	3.74%	2.87%
	S1	-	12.64%	0.78%	15.87%	71.51%	-	0.20%	3.35%
	PC1	-	-	-	-	-	100%	-	-
	W	-	-	-	23.81%	5.23%	-	92.13%	8.61%
	URM	-	9.20%	-	-	8.72%	-	3.74%	81.82%
Generalized Feed Forward Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM
	PC2	90.91%	-	-	-	-	-	-	-
	C	-	73.56%	-	-	4.07%	-	0.20%	2.39%
	S3	-	-	97.67%	0.79%	5.81%	-	-	-
	RM	-	1.15%	1.55%	61.90%	4.65%	-	3.74%	2.39%
	S1	9.09%	14.94%	0.78%	17.46%	69.77%	-	0.20%	5.74%
	PC1	-	-	-	-	-	100%	-	-
	W	-	1.15%	-	19.84%	5.23%	-	91.93%	8.13%
URM	-	9.20%	-	-	10.47%	-	3.94%	81.34%	
Modular Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM
	PC2	100%	-	-	-	-	-	-	-
	C	-	72.41%	-	-	5.23%	-	0.20%	2.39%
	S3	-	-	98%	1.59%	5.81%	-	-	-
	RM	-	1.15%	1.55%	57.94%	4.65%	-	3.74%	1.91%
	S1	-	16.09%	0.78%	16.67%	70.93%	-	0.20%	5.74%
	PC1	-	-	-	-	-	100%	-	-
	W	-	1.15%	-	23.81%	5.23%	-	91.93%	8.61%
URM	-	9.20%	-	-	8.14%	-	3.94%	81.34%	
Radial Basis Functions Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM
	PC2	0%	-	-	-	-	-	-	-
	C	-	0%	-	-	-	-	-	-
	S3	-	-	98%	1.59%	16.28%	78.57%	-	-
	RM	-	-	-	0%	-	-	-	-
	S1	-	43.68%	-	59.52%	69.77%	14.29%	0.39%	34.93%
	PC1	-	-	-	-	-	0%	-	-
	W	100.00%	22.99%	1.55%	38.89%	9.30%	4.76%	98.03%	19.62%
URM	-	33.33%	-	-	4.65%	2.38%	1.57%	45.45%	
Support Vector Machines Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM
	PC2	100.00%	-	-	-	-	-	-	-
	C	-	100.00%	-	-	-	-	-	-
	S3	-	-	98.45%	-	-	-	-	-
	RM	-	-	-	100.00%	-	-	-	-
	S1	-	-	-	-	100.00%	-	-	-
	PC1	-	-	-	-	-	100%	-	-
	W	-	-	1.55%	-	-	-	99.80%	-
URM	-	-	-	-	-	-	0	100.00%	

Table 4.10 -- ANN testing errors (Structure type not recognized) by model

Model	Code	Predicted							
		PC2	C	S3	RM	S1	PC1	W	URM
Multilayer Perceptron Neural Network	PC2	100%	-	-	-	-	-	-	-
	C	-	72.73%	-	-	2.70%	-	0.00%	7.14%
	S3	-	-	100.00%	0.00%	5.41%	-	-	-
	RM	-	0.00%	0.00%	65.63%	5.41%	-	5.61%	0.00%
	S1	-	9.09%	0.00%	18.75%	75.68%	-	0.00%	0.00%
	PC1	-	-	-	-	-	100%	-	-
	W	-	0	-	12.50%	0.00%	-	89.72%	7.14%
	URM	-	13.64%	-	0	10.81%	-	0	85.71%
Generalized Feed Forward Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM
	PC2	100%	-	-	-	-	-	-	-
	C	-	77.27%	-	-	2.70%	-	0.00%	4.76%
	S3	-	-	95.45%	0.00%	5.41%	-	-	-
	RM	-	0.00%	0.00%	68.75%	5.41%	-	5.61%	2.38%
	S1	0.00%	9.09%	4.55%	21.88%	72.97%	-	0.00%	2.38%
	PC1	-	-	-	-	-	100%	-	-
	W	-	4.55%	-	6.25%	0.00%	-	90.65%	7.14%
URM	-	9.09%	-	0	13.51%	-	3.74%	83.33%	
Modular Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM
	PC2	100%	-	-	-	-	-	-	-
	C	-	63.64%	-	-	2.70%	-	0.00%	4.76%
	S3	-	-	100%	0.00%	5.41%	-	-	-
	RM	-	0.00%	0.00%	68.75%	5.41%	-	4.67%	0.00%
	S1	-	13.64%	0.00%	18.75%	72.97%	-	0.00%	2.38%
	PC1	-	-	-	-	-	100%	-	-
	W	-	4.55%	-	12.50%	0.00%	-	91.59%	7.14%
URM	-	18.18%	-	-	13.51%	-	3.74%	85.71%	
Radial Basis Functions Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM
	PC2	0%	-	-	-	-	-	-	-
	C	-	0%	-	-	-	-	-	-
	S3	-	-	100%	0.00%	27.03%	88.89%	-	-
	RM	-	-	-	0%	-	-	-	-
	S1	-	50.00%	-	71.88%	59.46%	11.11%	0.93%	45.24%
	PC1	-	-	-	-	-	0%	-	-
	W	50.00%	22.73%	0.00%	28.13%	2.70%	0.00%	96.26%	16.67%
URM	1	27.27%	-	-	10.81%	0.00%	2.80%	38.10%	
Support Vector Machines Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM
	PC2	50.00%	-	-	-	-	-	-	-
	C	-	72.73%	-	-	8.11%	-	-	4.76%
	S3	-	-	81.82%	-	5.41%	-	-	-
	RM	-	4.55%	4.55%	65.63%	2.70%	-	1.87%	-
	S1	-	-	13.64%	25.00%	75.68%	-	0.93%	11.90%
	PC1	-	-	-	-	-	100%	-	-
	W	50.00%	4.55%	-	3.13%	-	-	93.46%	7.14%
URM	-	18.18%	-	6.25%	8.11%	-	3.74%	76.19%	

Again, the errors in the three MLP-based models (the MLP, GFF and MNN) are relatively consistent in both training and testing. The RBF does not perform as well with several cases where the structure type errors are 100%. As mentioned before, the SVM shows spurious results. In the three MLP-based models, “C” buildings not recognized are largely distributed between “S1” and “URM”. The classification performance of “RM” structures is low, with unrecognized “RM” being largely distributed between “S1” and “W”. “S1” unrecognized structures are uniformly distributed among all categories, except “PC1” and “PC2”. The performance in the “W” class is very good, with unrecognized structures being largely distributed between “RM” and “URM”, with “RM” dominating in the testing component. Finally, unrecognized “URM” structures are distributed more or less uniformly over several categories in the training component, and between “C” and “W” in the testing component.

Table 4.11 shows a summary of the cases where the desired structure type was misclassified (row errors in the confusion matrix) for training and testing components in respectively in the five ANN models. Tables 4.12 and 4.13 show details by ANN model type for the training and testing components respectively, where the structure type was misclassified. These tables should be analyzed along rows. Cases of wrong classifications greater than 10% have been highlighted in both the detail tables. For the total structures predicted as the structure type described in the row, the diagonals highlight prediction accuracy (not model accuracy as in the previous detail tables). In other words, for the training case in the MLP ANN, the model predicted a total of 81 “C” structures (see Table 4.4). Of these 81 “C” structures, 66 (or 81.48% as highlighted in the diagonal for the MLP ANN in Table 4.12) were classified correctly, 7 (8.64%) were misclassified as “S1”, 1 (1.23%) was misclassified as “W” and 7 (8.64%) were misclassified as “URM”. A total of 15 (18.52%) “C” structures were therefore

misclassified into other categories, which is listed in Table 4.11 for the MLP ANN training row. Figure 4.6 shows the overall distribution of classification errors by ANN model type.

Table 4.11 -- Summary of errors (Structure type misclassified) by ANN model

Model	Part	C	PC1	PC2	S1	S3	RM	URM	W	Overall
MLP	Training	18.52%	0%	0%	24.54%	8.03%	33.93%	19.72%	10.86%	15.81%
	Testing	20.00%	0%	0%	22.22%	8.33%	27.59%	26.53%	7.69%	15.75%
GFF	Training	16.88%	0%	0%	29.41%	8.03%	30.97%	21.30%	10.02%	16.12%
	Testing	15.00%	0%	0%	28.95%	8.70%	29.03%	25.53%	5.83%	15.75%
MNN	Training	19.23%	0%	0%	28.65%	8.70%	31.78%	19.81%	11.05%	16.36%
	Testing	17.65%	0%	0%	27.03%	8.33%	24.14%	26.53%	7.55%	15.75%
RBF	Training	0%	0%	0%	61.78%	33.16%	0%	32.62%	22.07%	34.58%
	Testing	0%	0%	0%	71.43%	45.00%	0%	46.67%	18.25%	40.29%
SVM	Training	0%	0%	0%	0%	0%	0%	0%	0.39%	0.23%
	Testing	23.81%	0%	0%	37.78%	10.00%	19.23%	28.89%	5.66%	17.58%

Table 4.12 -- ANN training errors (Structure type misclassified) by model

Model	Code	Predicted							
		PC2	C	S3	RM	S1	PC1	W	URM
Multilayer Perceptron Neural Network	PC2	100%	-	-	-	-	-	-	-
	C	-	81.48%	-	-	8.64%	-	1.23%	8.64%
	S3	-	-	91.97%	1.46%	6.57%	-	-	-
	RM	-	1.79%	1.79%	66.07%	8.04%	-	16.96%	5.36%
	S1	-	6.75%	0.61%	12.27%	75.46%	-	0.61%	4.29%
	PC1	-	-	-	-	-	100%	-	-
	W	-	-	-	5.71%	1.71%	-	89.14%	3.43%
	URM	-	3.76%	-	-	7.04%	-	8.92%	80.28%
Generalized Feed Forward Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM
	PC2	100%	-	-	-	-	-	-	-
	C	-	83.12%	-	-	9.09%	-	1.30%	6.49%
	S3	-	-	91.97%	0.73%	7.30%	-	-	-
	RM	-	0.88%	1.77%	69.03%	7.08%	-	16.81%	4.42%
	S1	0.59%	7.65%	0.59%	12.94%	70.59%	-	0.59%	7.06%
	PC1	-	-	-	-	-	100%	-	-
	W	-	0.19%	-	4.82%	1.73%	-	89.98%	3.28%
URM	-	3.70%	-	-	8.33%	-	9.26%	78.70%	
Modular Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM
	PC2	100%	-	-	-	-	-	-	-
	C	-	80.77%	-	-	11.54%	-	1.28%	6.41%
	S3	-	-	91%	1.45%	7.25%	-	-	-
	RM	-	0.93%	1.87%	68.22%	7.48%	-	17.76%	3.74%
	S1	-	8.19%	0.58%	12.28%	71.35%	-	0.58%	7.02%
	PC1	-	-	-	-	-	100%	-	-
	W	-	0.19%	-	5.71%	1.71%	-	88.95%	3.43%
URM	-	3.77%	-	-	6.60%	-	9.43%	80.19%	
Radial Basis Functions Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM
	PC2	0%	-	-	-	-	-	-	-
	C	-	0%	-	-	-	-	-	-
	S3	-	-	67%	1.05%	14.74%	17.37%	-	-
	RM	-	-	-	-	-	-	-	-
	S1	-	12.10%	-	23.89%	38.22%	1.91%	0.64%	23.25%
	PC1	-	-	-	-	-	0%	-	-
	W	1.72%	3.13%	0.31%	7.67%	2.50%	0.31%	77.93%	6.42%
URM	-	20.57%	-	-	5.67%	0.71%	5.67%	67.38%	
Support Vector Machines Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM
	PC2	100.00%	-	-	-	-	-	-	-
	C	-	100.00%	-	-	-	-	-	-
	S3	-	-	100.00%	-	-	-	-	-
	RM	-	-	-	100.00%	-	-	-	-
	S1	-	-	-	-	100.00%	-	-	-
	PC1	-	-	-	-	-	100%	-	-
	W	-	-	0.39%	-	-	-	99.61%	-
URM	-	-	-	-	-	-	0	99.52%	

Table 4.13 -- ANN testing errors (Structure type misclassified) by model

Model	Code	Predicted							
		PC2	C	S3	RM	S1	PC1	W	URM
Multilayer Perceptron Neural Network	PC2	100%	-	-	-	-	-	-	-
	C	-	80.00%	-	-	5.00%	-	-	15.00%
	S3	-	-	91.67%	-	8.33%	-	-	-
	RM	-	-	-	72.41%	6.90%	-	20.69%	-
	S1	-	5.56%	-	16.67%	77.78%	-	-	-
	PC1	-	-	-	-	-	100%	-	-
	W	-	0.96%	-	3.85%	-	-	92.31%	2.88%
	URM	-	6.12%	-	2.04%	8.16%	-	10.20%	73.47%
Generalized Feed Forward Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM
	PC2	100.00%	-	-	-	-	-	-	-
	C	-	85.00%	-	-	5.00%	-	-	10.00%
	S3	-	-	91.30%	-	8.70%	-	-	-
	RM	-	-	-	70.97%	6.45%	-	19.35%	3.23%
	S1	-	5.26%	2.63%	18.42%	71.05%	-	-	2.63%
	PC1	-	-	-	-	-	100%	-	-
	W	-	0.97%	-	1.94%	-	-	94.17%	2.91%
URM	-	4.26%	-	2.13%	10.64%	-	8.51%	74.47%	
Modular Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM
	PC2	100%	-	-	-	-	-	-	-
	C	-	82.35%	-	-	5.88%	-	-	11.76%
	S3	-	-	91.67%	-	8.33%	-	-	-
	RM	-	-	-	75.86%	6.90%	-	17.24%	-
	S1	-	8.11%	-	16.22%	72.97%	-	-	2.70%
	PC1	-	-	-	-	-	100%	-	-
	W	-	0.94%	-	3.77%	-	-	92.45%	2.83%
URM	-	8.16%	-	-	10.20%	-	8.16%	73.47%	
Radial Basis Functions Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM
	PC2	0%	-	-	-	-	-	-	-
	C	-	0%	-	-	-	-	-	-
	S3	-	-	55%	-	25.00%	20.00%	-	-
	RM	-	-	-	0%	-	-	-	-
	S1	-	14.29%	-	29.87%	28.57%	1.30%	1.30%	24.68%
	PC1	-	-	-	-	-	0%	-	-
	W	0.79%	3.97%	-	7.14%	0.79%	-	81.75%	5.56%
URM	3.33%	20.00%	-	-	13.33%	-	10.00%	53.33%	
Support Vector Machines Neural Network	Code	PC2	C	S3	RM	S1	PC1	W	URM
	PC2	100.00%	-	-	-	-	-	-	-
	C	-	76.19%	-	-	14.29%	-	-	9.52%
	S3	-	-	90.00%	-	10.00%	-	-	-
	RM	-	3.85%	3.85%	80.77%	3.85%	-	7.69%	-
	S1	-	-	6.67%	17.78%	62.22%	-	2.22%	11.11%
	PC1	-	-	-	-	-	100%	-	-
	W	0.94%	0.94%	-	0.94%	-	-	94.34%	2.83%
URM	-	8.89%	-	4.44%	6.67%	-	0	71.11%	

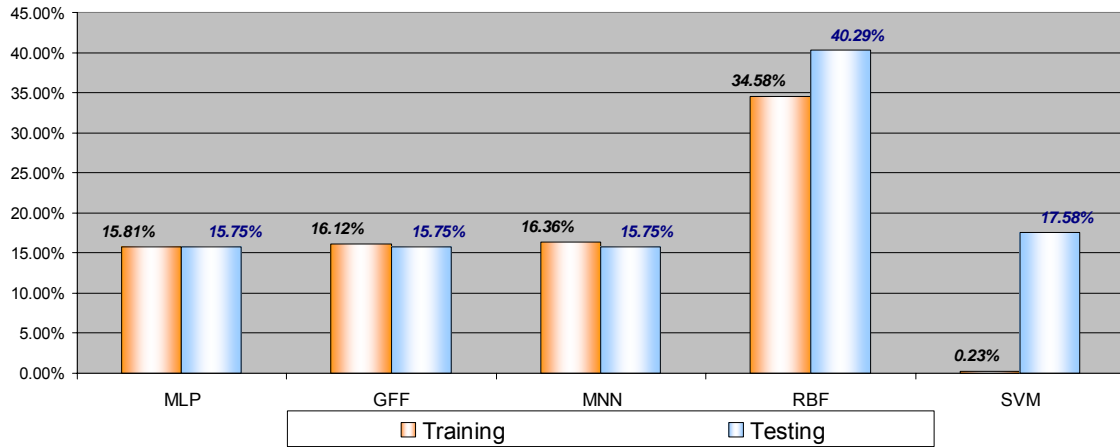


Figure 4.6 -- Structure type classification errors by ANN model type

4.2.1.4. Analysis of Classification Errors

Analyzing the tables presented above, clearly the MLP-based models have several problems, both in terms of classification errors and misclassifications. There is considerable confusion between “C”, “S1” and “URM”. The models also have problems discriminating among “RM”, “S1” and “W”. Further, “S1” is distributed among several categories particularly among “C”, “RM”, “S3” and “URM”. “W” is classified well, but is sometimes confused between “RM” and “URM”.

An analysis of the input variables revealed that similar combinations of the input variables (stories, area, year of construction, occupancy and fire rating) had substantial numbers of differing structure types, where even a human cannot accurately classify the structure type based on the input combination. For instance, in the sample data, 15 instances of Concrete buildings built mostly between 1915 and 1950 in the historic zone ranged from 1 to 5 stories and used for warehouses or retail or commercial offices with fire resistant fire rating were confused with Unreinforced Masonry because there were several URM buildings with similar characteristics. In another example, over 30 Unreinforced Masonry buildings in the sample of 1-2 stories, moderately sized, built

between 1910 and 1958 and used for retail or restaurants or offices or small apartments were confused with Wood structure types because there were several instances of Wood frame buildings sharing these same characteristics. In other words, the models (or I) could not discriminate between structure types that shared common input characteristics.

Overall, the models perform well for Precast Concrete, Light Metal, Concrete Tilt-ups and Wood frame buildings. Precast Concrete and Concrete Tilt-ups are not relevant to the model performance, because they were coded from the exterior wall documentation in the Tax Records. Concrete buildings have classification errors with Unreinforced Masonry and Steel buildings. Steel frames are misclassified among Concrete, Reinforced Masonry and Unreinforced Masonry structure types. Light Metal frames have some confusion with Steel frames. Reinforced Masonry is often misclassified as Steel or Wood. Unreinforced Masonry is confused with Concrete, Steel and Wood structure types.

4.2.1.5. Consequences of Classification Errors in Loss Estimation and Mitigation

Classification errors between Concrete and Steel, Steel and Light Metal, Steel and Reinforced Masonry, and Reinforced Masonry and Wood or Light Metal are not problematic, because building behavior under earthquake stresses exhibit similar damage characteristics – admittedly, there is considerable variation in behavior based on height, area and form, but the consequences are not very significant. Errors between Unreinforced Masonry and other structures are particularly problematic, because Unreinforced Masonry buildings are susceptible to heavy damage in earthquakes and pose grave danger to humans with high potential for severe injuries and deaths.

Let us separate the types of Unreinforced Masonry classification errors for this discussion. First, different structure types could be misclassified as Unreinforced Masonry – let us term this as a Type 1 URM classification error. The other case is where Unreinforced Masonry buildings are misclassified into one of the other structure types – this is termed as a Type 2 URM classification error. In loss estimation exercises, Type 1 URM classification errors have implications in terms of overestimating damage to buildings, and consequently overestimate casualties, shelter requirements, debris generated and direct and indirect economic losses. Type 2 URM classification errors have more drastic consequences. In loss estimation exercises, these errors will underestimate damage to buildings, and consequently underestimate casualties, shelter needs, debris and direct and indirect losses. However, and this aspect is infinitely more crucial, loss estimation exercises will not recognize damage to these erroneously classified buildings, but the real world consequences can potentially be very devastating. Damage to or the collapse of Unreinforced masonry buildings poses severe threats to life safety and can potentially cause high numbers of casualties and deaths. Direct and indirect losses may also be higher than expected.

The ANN models output a probability score along with the structure type classification. An analysis of the classification errors revealed that consistently, the probability scores of at least two structure type categories competed strongly, and the model chose the structure type with the higher probability. In the context of the Unreinforced Masonry discussion, between Concrete and Unreinforced Masonry, several pairs of competing probabilities were noted. Combinations of probability scores as (C, URM) included (.06, .45), (.16, .9), (.63, .65), (.14, .62), (.39, .67), (.19, .39), etc. Note that several cases show consistently low URM levels of probability (less than .7) and correspondingly close levels of C probability.

These probability scores may be gainfully used in mitigation planning. First, URM structures are recognized as critical structures. Mitigation plans should adopt a strategic approach to retrofitting URM structures with priority for residential and commercial occupancies (those occupancies with higher occupancy rates) over warehouses and factories. Third, inventory modeling exercises using ANNs should analyze the probability for URMs for all other structure type classifications and investigate those that have competing URM probabilities (say $> .25$ URM scores). Fourth, all URM misclassifications should be highlighted and analyzed for competing probability scores with other structure types. This would enable the quantification of URM structures with a higher degree of scrutiny along with a spatial delineation of the URM structures. Mitigation plans could then communicate the heightened vulnerability of such structures to the appropriate stakeholders and arrange funding sources to retrofit URM buildings to seismically safer standards.

4.2.1.6. Model Complexity, Sample Size and Model Calibration in Neural Networks

There are several cases where the model could not discriminate between two structure types because of the extreme similarity in input combinations – this problem is similar to multicollinearity problems experienced in linear regression. Typically, multicollinear inputs are handled by collecting more samples – see Goldberger (1991) for a witty and yet enlightening explanation of the problem of multicollinearity and its solution. However, the process of collecting samples is time- and resource-consuming, especially in the case of structure types within a large city. Most local jurisdictions lack the proper mechanisms for targeted sample data collection, and more importantly, rarely have the funds necessary to develop calibration sample databases.

Artificial neural computing approaches for classification are not parametric and require decisions on the topological complexity of the connections between the

processing elements (layering), the number of exemplars (sample size) and their distribution into training, cross-validation and testing sets (model calibration for generalizability).

Model Topological Complexity (Number of Processing Elements and Layers)

Generally, classification problems based on a finite set of factor levels and covariates are nonlinear in the input space and network topology usually requires at least one hidden layer to account for the nonlinearity. Additional hidden layers add complexity to the model and increase the number of connection weights to be estimated and could potentially overfit the data. The general approach to determining model complexity in ANNs is similar to step-wise linear regression – that is, start with the simplest topological configuration with the fewest number of processing elements and progressively increase the processing elements and then the layers while evaluating the classification performance. While there are quantitative measures that allow decision-making on model size or complexity (Akaike 1974) based on balancing decreasing errors and increasing penalties as a function of model size, most practitioners (Principe et al. 2000) advocate a performance-based evaluation of model complexity.

Sample Size for Model Calibration

There are very real benefits in modeling structure type using calibration samples drawn from the local region of interest, rather than the commonly used top-down approach of imposing a structure type distribution derived at large geographic levels. Collection of calibration data is therefore vital to the structure type modeling advocated in this research. If this method is expected to be replicated in other regions, the question of the number of samples required for model calibration needs to be addressed. Determining sample size is critical, since too large a sample may waste resources with

little modeling gains, while prediction inaccuracies result from small sample sizes. The problem is heightened in classification models, where an adequate number of samples have to be drawn for each class, and the proportion of each class in the general population may not be known.

In typical statistical designs for estimating population proportions based on the assumption of normality, sample sizes are influenced by the confidence level, the confidence interval and the prevalence of a class. Any standard introductory statistical textbook would explain these concepts. In simple terms, the confidence level is the amount of uncertainty that can be tolerated. The confidence interval, also termed margin of error, represents the upper and lower bounds of error that may be tolerated. In other words, the margin of error is the maximum difference between the proportion estimated from the samples and the true proportion of the class in the population. The confidence level therefore represents how often the true value or proportion of the population lies within the confidence interval. Confidence level for a standard normal distribution is implemented by the “Z” score, a critical value determined by the area under the standard normal curve. A confidence level of 95% is represented by a “Z” score of 1.96, and 99% by 2.85. The prevalence of a class (or a response) is the expectation of the proportion of the class within the total population, and is usually established from prior research. For instance, in the context of structure type classification, one could be 95% confident that the true proportion of concrete buildings is between 9% and 13%, using a confidence interval of 2%. Thus, sample size n , may be estimated by the following formula

$$n = (z_c^2 * p(1-p)) / e^2 ,$$

where z_c is the Z score corresponding to the confidence level, p is the proportion of the class in the population and e is the confidence interval.

Assuming a population size of 20,000 composed of 3 classes whose proportions are estimated by prior studies at 20%, 30% and 50%, for a confidence level of 95% with a margin of error of 5%, the minimum number of samples for the 3 classes would be 246, 323 and 384 respectively. For a confidence level of 99% with a margin of error of 5%, the corresponding sample sizes would increase to 520, 682 and 812. Similarly, for a confidence level of 95%, but with a confidence interval of 10%, the sample sizes would decrease to 61, 80 and 96. Other methods for determining sample size use the total population of buildings with a chi-squared distribution, as seen in the formula

$$n = \chi^2 * N * p(1-p) / e^2 (N-1) + \chi^2 * p(1-p) ,$$

where χ^2 is the value of chi-square for one degree of freedom and the desired confidence interval, N is the population size or the size of the smallest sub-group to be proportionately represented, p is the proportion of the class in the population and e is the confidence interval.

The resulting sample size calculations yield similar results. It should be noted that the sample size calculations assume that the samples are genuinely randomly distributed, and if this assumption is violated (owing to some structural stratification mechanism in the population, or selection bias), confidence intervals or sample size calculations may not be reliable. In reality, this is often a problem – in the context of buildings, one rarely knows the proportion of buildings of structural classes a priori. Further, structure types are not distributed randomly across the region – buildings follow a historical trajectory of development, based on previously settled areas, or follow

particular arterial transportation connections, or may be influenced by building code and enforcement. The sample size calculations may however be used as rough guidelines.

In addition to the guidelines suggested above, Principe et al (2000) suggest a set of thumb rules for determining sample sizes based on the topological complexity of the neural network, or the number of input attribute columns that influence the dependent variable. The topological complexity is measured by the number of connection weights in the network, and the total number of exemplars should be between 5 and 10 times the number of connection weights. While generating such sizes is possible in simulations, the sample size determined here is generally of an order higher than is feasible for field-based surveys. The MLP ANN model specified in Figure 3.4 has 673 connection weights, resulting in a minimum sample size of over 3,300. Another practical rule relates the total number of attribute columns, suggesting that the total sample size be at least 50 times the number of attribute columns. Again, applying this rule for the MLP ANN of Figure 3.4, the minimum sample size is about 1050. A final rule relates the number of attribute columns and the smallest class – the minimum number of exemplars for the smallest class in the population should be between 5 and 10 times the number of attribute columns.

Model Calibration for Generalization

Small training sets result in inadequate estimation of network weights and poor classification performance, while overly large training sets with small cross-validation and testing datasets may result in memorization of data patterns and poor generalization to unseen data. In addition to good classification performance, consistency in model performance over training, cross-validation and testing datasets is highly desired for adequate generalization to unseen data. In other words, if the training samples account for 90% of the sample data, while cross-validation and testing account for the remaining

10%, training classification performance may be very high, but testing classification may be poor. Similarly, if training samples account for only 50% of the sample data, while the remaining samples are distributed equally between cross-validation and testing, the training classification performance may be erratic and inconsistent with the testing performance. Further, learning in ANNs is a stochastic process and weights estimated by training the network should be estimated over several runs in order to establish consistency and reliability.

4.3. Comparison of Multinomial Logistic Regression and Neural Networks

For the purposes of comparing the logistic regression with ANNs, the same specification that was used for the logistic regression was run with an MLP-based ANN with one hidden layer. Table 4.14 shows the confusion matrices of raw counts derived from each model. Table 4.15 details the percentage of accurate classifications. Tables 4.16 and 4.17 show the percentages of structure type not recognized and misclassifications respectively for the two specifications.

Table 4.14 -- Comparison of confusion matrices (Logistic vs. ANN)

Model	Structure Type	URM	S	C	W	Totals
Neural Network	URM	222	17	11	42	292
	S	46	539	23	38	646
	C	13	15	172	3	203
	W	22	41	3	624	690
	Totals	303	612	209	707	1,831
Multinomial Logistic Regression	URM	208	30	14	28	280
	S	42	519	31	28	620
	C	15	13	161	5	194
	W	38	50	3	646	737
	Totals	303	612	209	707	1,831

The performance of the two models was strikingly similar – the multinomial logistic regression correctly classified 1534 samples (83.78%), while the ANN correctly

classified 1557 samples (85.04%) of the total 1831 available samples. Over 90% (1412) correspondence between the models was noted in successfully classified samples. The confusion matrices and the corresponding percentages of accuracy and errors reveal that the ANN performed marginally better, particularly for the Unreinforced Masonry and Concrete structure categories. The multinomial logistic regression performed slightly better than the ANN for Wood structure types.

Table 4.15 -- Classification accuracy (Logistic vs. ANN)

Model	Structure Type	URM	S	C	W	Overall
Neural Network	URM	73.27%				85.04%
	S		88.07%			
	C			82.30%		
	W				88.26%	
Multinomial Logistic Regression	URM	68.65%				83.78%
	S		84.80%			
	C			77.03%		
	W				91.37%	

As derived from Table 4.16, the ANN does not recognize 81 (26.73%) of the 303 Unreinforced Masonry structure samples, and is most likely to erroneously categorize Unreinforced Masonry into Steel (15.18%) structure types, while the multinomial logistic regression fails to recognize 95 (31.35%) of the 303 Unreinforced Masonry structures and tends to distribute the Unreinforced Masonry recognition errors largely between Steel (13.86%) and Wood (12.54%) structure types. From Table 4.17, of the 292 Unreinforced Masonry predictions by the ANN, 70 (23.97%) of the Unreinforced Masonry structures are misclassified largely as Wood (14.38%), while the multinomial logistic regression distributes its 72 (25.71%) errors predominantly between Steel (10.71%) and Wood (10.00%) structures.

Table 4.16 -- Percent Structure type not recognized (Logistic vs. ANN)

Model	Structure Type	URM	S	C	W
Neural Network	URM	73.27%	2.78%	5.26%	5.94%
	S	15.18%	88.07%	11.00%	5.37%
	C	4.29%	2.45%	82.30%	0.42%
	W	7.26%	6.70%	1.44%	88.26%
Multinomial Logistic Regression	URM	68.65%	4.90%	6.70%	3.96%
	S	13.86%	84.80%	14.83%	3.96%
	C	4.95%	2.12%	77.03%	0.71%
	W	12.54%	8.17%	1.44%	91.37%

Table 4.17 -- Percent Structure type not predicted (Logistic vs. ANN)

Model	Structure Type	URM	S	C	W
Neural Network	URM	76.03%	5.82%	3.77%	14.38%
	S	7.12%	83.44%	3.56%	5.88%
	C	6.40%	7.39%	84.73%	1.48%
	W	3.19%	5.94%	0.43%	90.43%
Multinomial Logistic Regression	URM	74.29%	10.71%	5.00%	10.00%
	S	6.77%	83.71%	5.00%	4.52%
	C	7.73%	6.70%	82.99%	2.58%
	W	5.16%	6.78%	0.41%	87.65%

4.3.1. Differences and Relative Advantages of Multinomial Logistic Regression and Artificial Neural Networks

Deviations between the two model performances are relatively minor, with consistent patterns of recognition and misclassification errors. Further, sensitivity tests in the ANN environment mirror significant relationships between the structure type (dependent) and the independent covariates and factors specified in the multinomial logistic regression. The relatively close performances agree with what has been found in surveys of classification literature. Dreiseitl and Ohno-Machado (2002) reviewed over 70 publications from medical classification literature, and noted that both logistic

regression and ANN approaches perform similarly with the increased flexibility offered by ANNs being the primary reason for their preferred use. The differences and relative advantages of each approach are described in further detail below.

Unlike Support Vector Machine approaches that estimate linear (and therefore dichotomous) separations between classes, both multinomial logistic regressions and ANNs attempt to model or approximate the posterior probability of the dependent variable given the specific combination of inputs. Multinomial logistic regression models are parametric, based on a clearly specified functional form described earlier in Section 2.2, while ANNs are classified as semi-parametric or non-parametric. Multinomial logistic regression models therefore have substantially more explanatory power than ANNs and permit interpretation and evaluation of the effects of the input variables on the dependent. In particular, the odds of successful outcomes between pairs of the dependent variable alternatives (given the specific input combination) are quantified clearly in logistic regressions. While ANNs have analytical procedures to examine the sensitivity of outcomes to inputs heuristically (Zurada et al. 1994), the sensitivity measures tend to be unit-less and are not interpretable in a quantitative sense. Further, ANNs do not present quantitative measures between outcome pairs. Additionally, ANNs are sensitive to starting values and the methods are difficult to replicate in a mathematical sense, while logistic parameters are determined by more pleasing (in a statistical sense) maximum likelihood estimation methods. If the performances between logistic and ANN approaches are so similar, is there any advantage to sacrificing the explanatory and evaluative power of a parametric logistic regression specification for an ANN model?

The parametric specification of multinomial logistic regressions require a minimum number of samples in each cell of a cross-tabulation between the dependent

variable alternatives and the input covariates and factor levels for estimation. Very often, it may not always be possible to collect sample sets that are complete with respect to all cells of a cross-tabulation between the dependent variable alternative and the inputs, because of inadequate sampling, or the expense of collecting an underrepresented category or because that category may not exist in the general population vis-à-vis a particular input. In this research for instance, no samples were observed for Wood structures that were used in Heavy Industrial occupancies, consistent with construction practices. Such gaps reduce the tractability of the multinomial logistic regression and the relative inflexibility frequently prompts the simplification of both the number of alternatives in the dependent variable and the number of levels for input factors. ANNs are far more flexible, and their semi-parametric nature is far more tolerant to noise or gaps in the input combinations, and less likely to drastically fluctuate in classification performance efficiency (Rojas 1995).

Full effects multinomial logistic regression specifications may overtrain the sample data, while training in ANN procedures may be stopped at recognizable points, specifically to prevent memorization of training data patterns. Thus, ANNs use cross-validation routines explicitly (Principe et al. 2000) in the process of weight estimation for generalization to unseen data (and thus prevent overtraining), while given the same limited sample data set, estimation or classification performance suffers in the logistic regression approach if samples are set aside for cross-validation.

Parametric models are generally not effective in modeling non-linear and complex relationships between inputs and outputs, while ANNs have been shown to learn complex patterns by example efficiently without requiring large numbers of samples (Makhfi 2007). In addition, since ANNs learn complex relationships based on examples derived directly from the representative population without human intervention

(StatSoft 2003), ANN engines may be specifically designed for automated calibration, adaptation and response and embedded in application environments familiar to the end-user. In the case of multinomial logistic regression, the inherent lack of flexibility and inadequate fault-tolerance greatly reduces automation potential.

ANNs may be designed and optimized for extraordinary parallel processing, greatly enhancing the speed of training and outcome prediction, especially for large and complex datasets (Rumelhart et al. 1986; Rumelhart and McClelland 1986).

Finally, while ANNs require some user knowledge for variable selection, network topology and result interpretation (Nilsson 1996; Patterson 1996; Principe et al. 2000) the level of such knowledge is considerably lower than traditional non-linear statistical methods, particularly when model performance is emphasized (Anderson and McNeil 1992).

4.3.2. Using Artificial Neural Networks for Structure Type Classification

In this research, the lack of data values in all cells of cross tabulations of structure type and occupancy reduced the tractability of the multinomial logistic regression model, prompting the simplification of both the dependent variable and the factor levels. The inherent flexibility of the ANN and its semi-parametric approach is far more forgiving of gaps in the data. On the other hand, the black-box nature of ANNs makes it more difficult to explain good performance or convince doubters about the efficiency of the approach. Multinomial logistic regression approaches, by their parametric nature, allow for the validation of a model's plausibility by comparing similar studies or surveying experts in the field. The choice in this dissertation to use ANNs for structure type classification was prompted more by the needs of classification performance than interpretation or explanation – after all, the output building data

inventory would be used by a variety of downstream models and applications and accuracy in accounting of the built environment was the primary motivating factor.

In any case, the close similarity in performance between similar specifications for both the multinomial logistic regression and the ANN models legitimizes the use of ANNs. While the ANN models do not have statistical measures for analyzing the performance of the model or for describing the relationship between the inputs and the outcome, the results of input variable sensitivity tests within the ANN framework show patterns that are consistent with the quantitative and statistically significant relationships between inputs and structure type outcomes derived in the multinomial logistic regression model. Figure 4.7 shows the results of the sensitivity of structure type to the various input variables in the ANN models. Note that Year of construction, Retail and Wholesale Trade and the Fire rating variables seem to be good explainers for Steel, Wood and URM structure classes – these relationships were noted in the multinomial logistic regression results also – see the Tables in Appendix B.

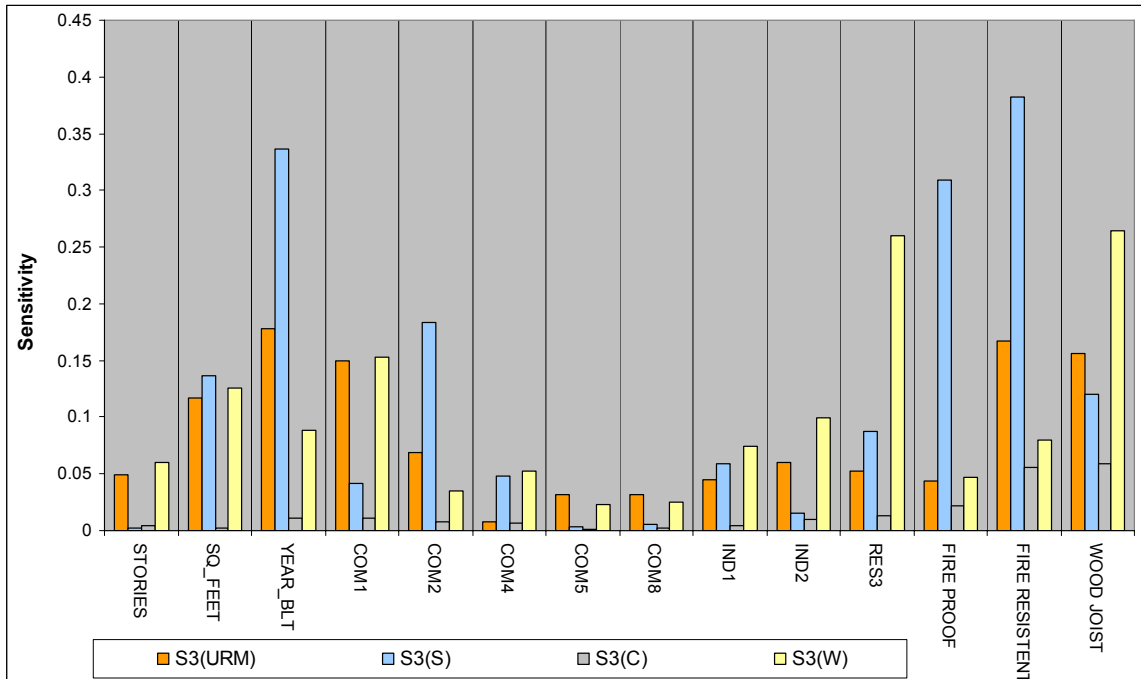


Figure 4.7 -- Structure type sensitivity to input variables (ANNs)

4.4. Recognizing Building Footprints

The 2D building configuration classification is implemented in two stages. In the first stage, the building footprint is preprocessed in the GIS environment using the following routines in sequence – collinear vertex removal, enforcing orthogonalization, edge-smoothing by removing spikes, small concavities and convexities, simplification and final adjustments. For testing and validation, several building footprint libraries were created. Over 5,000 building footprints for all classes were created through manual digitization. Three separate libraries were tested for manually digitized footprints and one for automatically extracted footprints.

4.4.1. Classification on Manually-digitized Building Footprints

Figure 4.8 shows the various steps in the sequence of preprocessing the building footprints, using the example of a single, casually digitized L-shaped polygon. Note the

level of extraneous detail, the collinear vertices and the lack of orthogonal corners in the base input polygon.

As the panels indicate, after collinear vertices are decimated, corners are made orthogonal based on a user-specified threshold of 20 degrees. Then, minor protrusions and intrusions are removed, after which the building is simplified and approximated. The simplified building is then processed by the classification routine. In the validation experiments, the algorithm classified all the input footprints with a success rate of over 97% -- the only errors occurred in boundary conditions between classes, such as between C-shaped and Rectangular buildings, where the intrusion into the rectangle was of a minor dimension relative to the breadth of the rectangle.

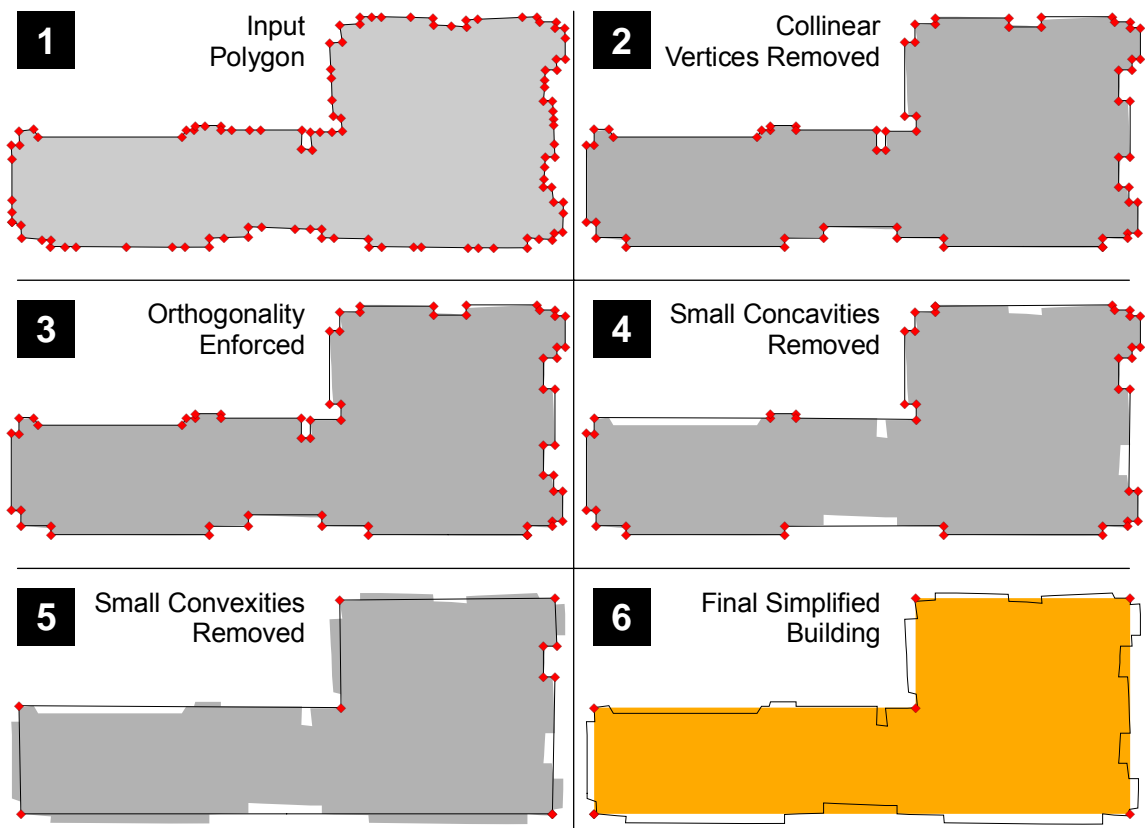


Figure 4.8 -- Preprocessing manually-digitized footprints

Against an initial library of about 166 “clean” buildings of all 10 footprint classes, the shape recognition algorithm achieved 100% classification accuracy. This library was created solely for testing the shape classifier, and consisted of all footprint classes with atomic vertices and orthogonal edges where appropriate.

A second library of about 5,600 footprints was created, consisting of all 10 footprint classes at various scales. This library represented real world conditions and contained buildings with non-orthogonal corners and collinear vertices and was created to test the preprocessing and classification algorithms. Again, the classifier performed excellently, but the preprocessing routines created polygons that were somewhat different from the original inputs. In several cases, the errors represented boundary conditions, reflecting ambiguity between pairs of footprint classes. Table 4.18 shows the performance of the shape recognition module for this library.

Table 4.18 -- Performance of shape recognition (Manual digitization)

Reference Type	Total Samples	Errors	Percent Errors
CIRCULAR	90	-	-
CRUCIFORM	455	-	-
C-SHAPED	1,449	114	7.87%
H-SHAPED	1,113	-	-
IRREGULAR	180	13	7.22%
L-SHAPED	800	33	4.13%
OCTAGON	153	16	10.46%
RECTANGULAR	650	241	37.08%
T-SHAPED	490	-	-
Z-SHAPED	220	-	-
Totals	5,600	417	7.45%

The module achieved about 93% accuracy, with the largest amount errors found in ambiguous boundary conditions between rectangles and other classes.

Figure 4.9 shows some examples of incorrect classifications for manually digitized buildings. The preprocessing routines use absolute values for threshold

distances and angles for all sizes of buildings. As a result, there is some sensitivity to geographic scale, or effectively, the size of the building footprint.

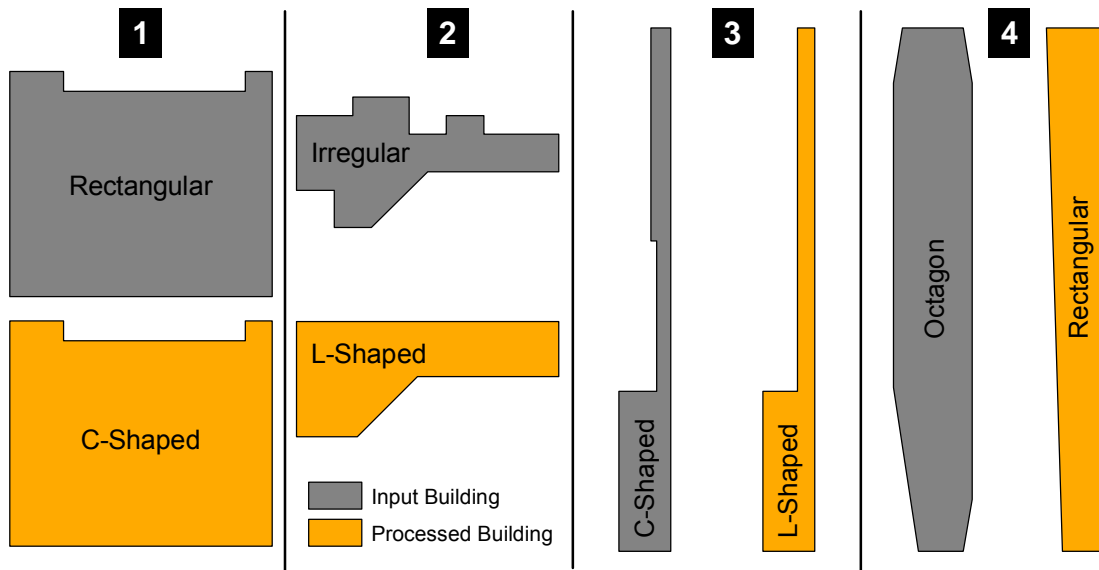


Figure 4.9 -- Example classification errors for manually digitized buildings

As seen in Figure 4.9, in panel 1, the input shape has been classified as “Rectangular” even though it is ostensibly “C-shaped” since the intrusions into the rectangle are relatively insignificant. This represents a typical boundary condition between C-shaped and Rectangular buildings. The shape recognition could not identify the footprint as a rectangle because the intrusions were greater than the specified simplification threshold distance. This ambiguity and consequent misclassification is seen again in panels 3 and 4. The error shown in panel 2 is a genuine error – when the threshold tolerance (especially for small buildings) is greater than one intrusion or protrusion but is less than the distance of successive protrusions or intrusions, simplification behavior is unstable. Consequently, the irregular building was misclassified as L-shaped. The same error did not occur when the building was enlarged, suggesting that the simplification thresholds are somewhat sensitive to scale.

4.4.2. Classification of Automatically-extracted Building Footprints

The overall performance with automatically extracted footprints from aerial photographs was less than that of processing manually digitized footprints. The input automatically extracted footprints are characterized by extremely noisy contours that make preprocessing and simplification unstable. Shapes with longer linear dimensions tend to be recognized easily, while those with several contour vertices that change slope rapidly and are of magnitudes greater than the user-specified threshold tend to be classified as irregular. Figure 4.10 shows the various steps in the sequence of preprocessing the building footprints, using the example of a poorly extracted and noisy T-shaped polygon. Again, noise could include collinear vertices, non-orthogonal corners, spikes, intrusions and protrusions, etc. that are artifacts of the extraction process. The preprocessing routines remove collinear vertices, orthogonalize corners, remove spikes, minor concavities and convexities and repeat collinearity removal and orthogonalization after simplification.

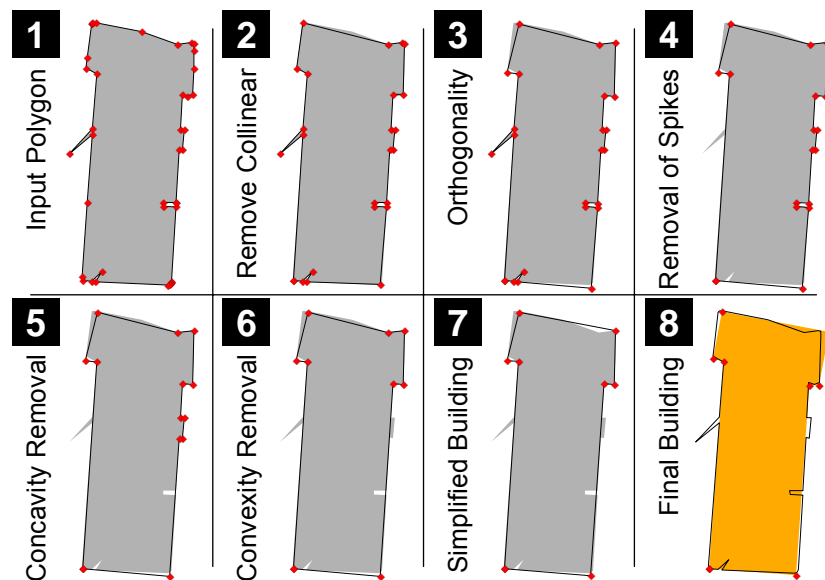


Figure 4.10 -- Preprocessing automatically extracted building footprints

As the successive panels indicate, after collinearity removal and orthogonalization, spikes, minor intrusions and protrusions are removed before the polygon is simplified. Collinearity removal and orthogonalization routines are run on the simplified building footprint, since collinear and non-orthogonal vertices may result as artifacts of simplification. When tested against a library of automatically extracted building footprints, the routines achieved success rates ranging from 70% to 82%. Table 4.19 quantifies the performance for a sample of 99 automatically extracted building footprints. The high percentage errors for C-, H- and Z-shaped footprints are not that serious because of the small sample size. However, a large number of rectangular buildings are misclassified.

Table 4.19 -- Example classification errors for automatically extracted footprints

Reference Type	Total Samples	Errors	Percent Errors
C-SHAPED	4	2	50.00%
H-SHAPED	2	2	100.00%
IRREGULAR	16	1	6.25%
L-SHAPED	17	2	11.76%
RECTANGULAR	57	19	33.33%
T-SHAPED	1	-	-
Z-SHAPED	2	1	50.00%
Totals	99	27	27.27%

Figure 4.11 depicts examples of misclassifications for building footprints automatically extracted from aerial photographs. Note that the input buildings are humanly intuited classifications and are often ambiguous because of the extreme noise in the contours of the input polygons. The preprocessing routines do remove collinear vertices and orthogonalize corners, but very often, the variation in the segments of the exterior contour of the input polygon results in poor landmark vertex definition.

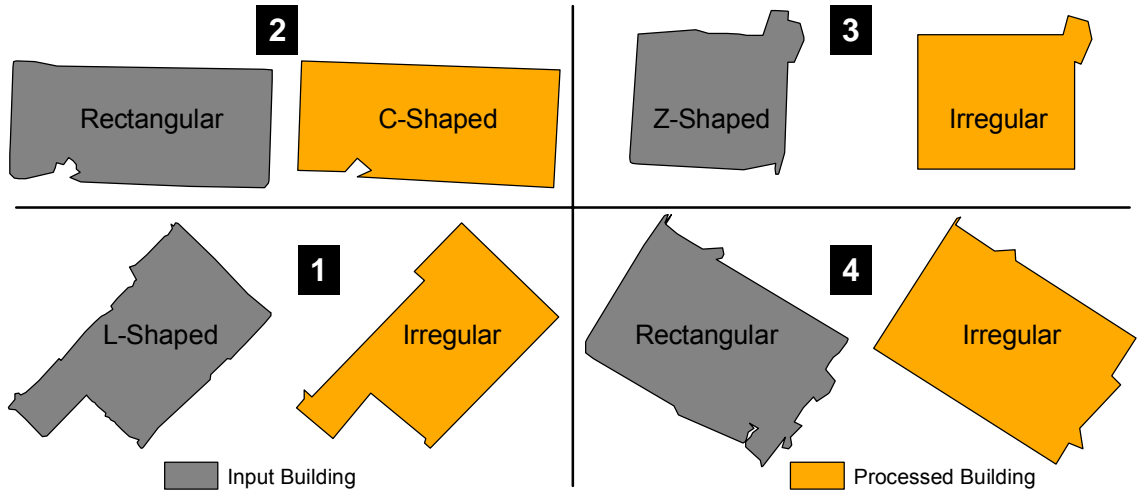


Figure 4.11 -- Examples of misclassifications for automatically extracted footprints

Note the high ambiguity in the input polygons in all four panels and the consequent results of the processed buildings. Again, using a single absolute threshold measure for simplification results in some sensitivity to footprint size. For instance, the building in panel 2 could have been correctly classified but for a few minor deviations – adjusting the simplification threshold would solve this problem for this particular building, but create classification errors elsewhere. Figure 4.12 shows two examples of successful L-shaped classifications from the sample library.

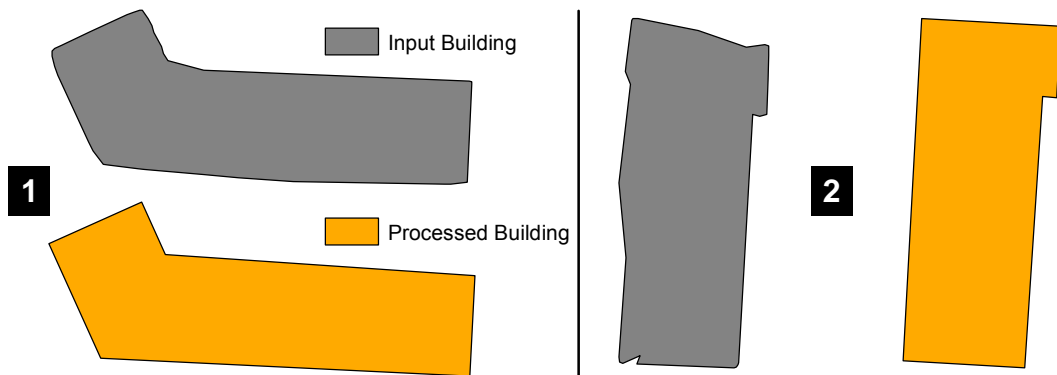


Figure 4.12 -- Successful classifications for automatically extracted footprints

4.4.3. Notes on the Classification Algorithm

In general, input footprints should be separated based on their methods of extraction. Manually digitized footprints tend to be less noisy in the edge contours and so, classification results demonstrate very high success rates. By contrast, the footprints generated from aerial photographs through automatic extraction routines tend to have extreme noise in their edge contours. In some cases, overlapping pixels may cause vectorized footprints to cross polygon boundaries, creating topologically inconsistent input footprint polygons.

Figure 4.13 shows examples of extremely noisy contours. In fact the noise level may be so great that it would be challenging for even humans to classify the footprints into the correct classes. Despite the noise, the algorithms produce accurate classifications of about 75%. In terms of improving the performance, either the preprocessing routines should be modified so as to include fuzzy generalization in the GIS environment, or alternately, the contours of the extracted footprint may be generalized at the point of extraction.

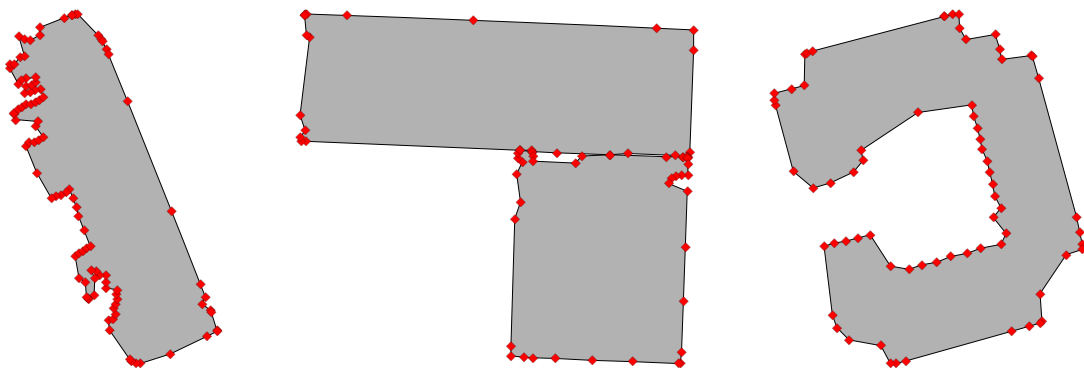


Figure 4.13 -- Examples of edge noise in automatically extracted footprints

4.5. Building Valuation

This section briefly covers the results of implementing the methodology for estimating the building replacement costs and content value for Shelby County. The building valuation component required the estimation of the building replacement costs, the structural component of the replacement costs, the acceleration- and drift-sensitive components of the replacement costs and the content value. All replacement costs were derived from the R. S. Means 2008 Square Foot Costs (2008) location-adjusted for Shelby County.

4.5.1. Replacement Costs for Shelby County

Following the flowchart described in Figure 3.1 of the methodology chapter, building valuation components required the determination of structure type as one of the inputs. After predicting the structure type using the ANN, the structure type was recorded into the building database. The occupancy code and replacement value code identifiers (nomenclature for which is described in Section 3.5.1.1. of the methodology chapter) were also recorded for each building along with the average square foot costs and the minimum and maximum area ranges and square foot costs from the Means manual. Replacement costs were then computed as the total above ground and below ground costs – if the building had no basement, then the below ground costs would amount to \$ 0.00. Table 4.20 shows the replacement cost by structure type in Shelby County. Note that the total replacement cost is dominated by Wood structures, amounting to over 59% of the total replacement cost in Shelby County, followed by Steel, Concrete and Masonry. This is expected, since Wood structures are dominated by residential occupancies that amount to almost 90% of the GBS. Note however, the low average of replacement cost for Light Wood structures (W1), largely comprising the

single-family stock. The total replacement cost for all buildings in Shelby County is estimated at just over \$ 87 billion.

Table 4.20 -- Replacement cost by Structure type in Shelby County

Structure Type	Count	Average Replacement Cost in \$ thousands	Row Total in \$ millions	Replacement Cost as Percent of Total
C1	913	4264.06	3,893.09	4.46%
C2	81	17649.44	1,429.60	1.64%
MH	43	4053.47	174.30	0.20%
PC1	1,110	6682.28	7,417.33	8.50%
PC2	35	6136.24	214.77	0.25%
RM	1,600	638.39	1,021.42	1.17%
S1	3,608	3717.73	13,413.57	15.37%
S3	3,522	1121.21	3,948.90	4.52%
URM	11,141	355.09	3,956.10	4.53%
W1	271,853	142.65	38,778.59	44.43%
W2	12,097	1077.26	13,031.61	14.93%
Totals	306,003	285.22	87,279.29	100.00%

For the square footage and dwelling unit imputation procedures described in the methodology chapter, the appropriate records were marked as imputed in the building inventory database. Table 4.21 describes the distribution of replacement costs for imputed structures and occupancy. Note that the total estimated replacement costs for the imputed structures amounts to about \$ 11 billion (or 13.62%), dominated by Multi-family apartments, Schools, Colleges and Universities, Churches and Emergency Response. Imputations were necessary for the Multi-family residential apartments, particularly for the 1960-1980 period, since the Tax Assessor's database had large gaps for these periods. Imputations in the other occupancy categories were necessary, since they are non-taxable (non-profit or governmental functions) and therefore, not recorded in the Tax Assessor's database.

Table 4.21 -- Imputation of Replacement costs (in millions of dollars) by HAZUS

MH MR-3 occupancy category

Occupancy Type	Imputed Replacement Costs	Total Replacement Costs	Percent by Occupancy	Percent of Imputed Total	Percent of Total
COM1	1.78	4,197.05	0.04%	0.01%	0.00%
COM2	29.22	10,762.05	0.27%	0.25%	0.03%
COM3	4.96	856.50	0.58%	0.04%	0.01%
COM4	10.09	5,904.48	0.17%	0.08%	0.01%
COM5	0.31	191.83	0.16%	0.00%	0.00%
COM6	-	727.20	0.00%	0.00%	0.00%
COM7	-	1,090.94	0.00%	0.00%	0.00%
COM8	7.15	752.99	0.95%	0.06%	0.01%
COM9	-	77.12	0.00%	0.00%	0.00%
COM10	-	302.23	0.00%	0.00%	0.00%
EDU1	1,690.31	1,744.28	96.91%	14.22%	1.94%
EDU2	2,741.27	2,741.27	100.00%	23.05%	3.14%
GOV2	178.39	204.86	87.08%	1.50%	0.20%
IND1	7.91	2,643.49	0.30%	0.07%	0.01%
IND2	9.70	658.13	1.47%	0.08%	0.01%
IND4	-	9.19	0.00%	0.00%	0.00%
IND5	-	74.13	0.00%	0.00%	0.00%
REL1	841.26	1,108.44	75.90%	7.07%	0.96%
RES1	21.39	38,187.72	0.06%	0.18%	0.02%
RES2	22.01	174.30	12.63%	0.19%	0.03%
RES3	6,077.07	13,308.68	45.66%	51.11%	6.96%
RES4	233.71	1,235.18	18.92%	1.97%	0.27%
RES5	9.98	34.39	29.02%	0.08%	0.01%
RES6	4.24	292.83	1.45%	0.04%	0.00%
Totals	11,890.76	87,279.29	13.62%	100.00%	13.62%

Table 4.22 shows the distribution of replacement costs for imputed structures by decade of construction. Note that replacement costs for imputed structures rise in the 1960s and then remain fairly steady by decade. The largest fraction of replacement costs for imputed structures occurred in the 1970-1979 decade. Note however, that the Tax Assessor's database has large omissions in Multi-family residential since the 1960s, coinciding with the period when the City of Memphis began to grow rapidly.

Table 4.22 -- Imputation of Replacement costs (in millions of dollars) by decade

Decade	Imputed Replacement Costs	Not Imputed Replacement Costs	Total Replacement Costs	Percent by Decade	Percent by Imputed Total	Percent by Total
Pre-1939	74.97	762.15	837.12	8.96%	1.23%	0.56%
40-49	86.70	263.13	349.84	24.78%	1.43%	0.65%
50-59	146.52	385.08	531.60	27.56%	2.41%	1.10%
60-69	1,007.74	1,686.11	2,693.85	37.41%	16.58%	7.57%
70-79	2,007.64	2,180.30	4,187.94	47.94%	33.04%	15.09%
80-89	1,117.52	913.43	2,030.95	55.02%	18.39%	8.40%
90-99	751.70	411.50	1,163.21	64.62%	12.37%	5.65%
Post-2000	884.27	629.91	1,514.19	58.40%	14.55%	6.64%
Totals	6,077	7,232	13,309	45.66%	100.00%	45.66%

4.5.2. Structural and Nonstructural Replacement Costs

Based on the percentage decomposition of replacement costs into structural, nonstructural acceleration- and drift-sensitive components derived from Means assembly costs (described in Section 3.5.3), the process estimated and recorded the various costs in the building inventory database. Table 4.23 describes the total nonstructural costs by component category and structure type for Shelby County in millions of dollars. Apart from the Concrete Tilt-ups, Precast Concrete and Light Metal structures where structural cost proportions are over 40%, in all other structure type categories, structural costs form only between 20% and 30% of the total replacement costs for the structure. The bulk of the total replacement costs is derived from the non-structural acceleration- and drift-sensitive costs. Table 4.24 lists the average structural, nonstructural acceleration- and drift-sensitive costs by structure type for Shelby County in thousands of dollars.

Table 4.23 – Total Nonstructural Replacement costs by Structure type for Shelby County (thousands of dollars)

Structure Type	Count	Structural Replacement Costs	Nonstructural Acceleration-sensitive Costs	Nonstructural Drift sensitive Costs
C1	913	974.27	1,486.97	1,431.85
C2	81	347.87	577.85	503.88
MH	43	42.53	65.89	65.89
PC1	1,110	3,365.45	2,137.06	1,914.82
PC2	35	132.43	44.09	38.25
RM	1,600	331.37	335.95	354.10
S1	3,608	3,258.66	5,582.98	4,571.93
S3	3,522	1,562.99	1,303.24	1,082.67
URM	11,141	1,219.32	1,333.14	1,403.63
W1	271,853	9,034.88	10,635.42	19,108.29
W2	12,097	2,819.34	5,127.99	5,084.29
Totals	306,003	23,089.11	28,630.59	35,559.60

Table 4.24 -- Average Nonstructural Replacement costs by Structure type for Shelby County (thousands of dollars)

Structure Type	Average Structural Replacement Costs	Average Nonstructural Acceleration-sensitive Costs	Average Nonstructural Drift-sensitive Cost
C1	1,067.11	1,628.67	1,568.29
C2	4,294.69	7,133.97	6,220.78
MH	989.05	1,532.21	1,532.21
PC1	3,031.94	1,925.28	1,725.06
PC2	3,783.72	1,259.80	1,092.73
RM	207.11	209.97	221.32
S1	903.18	1,547.39	1,267.17
S3	443.78	370.03	307.40
URM	109.44	119.66	125.99
W1	33.23	39.12	70.29
W2	233.06	423.91	420.29
Averages	75.45	93.56	116.21

4.5.3. Content Value

Content value was estimated as a function of replacement costs and specific occupancy – Table 4.25 describes the total and average content values by structure type in Shelby County. As expected, over 40% of the content value is concentrated in Wood structures, followed by significant amounts in Steel and Concrete structures.

Table 4.25 -- Content value by Structure type in Shelby County

Structure Type	Count	Average Content Value (\$ thousands)	Content Value (\$ millions)	Content Value as Percent of Total
C1	913	4624.72	4,222.37	6.54%
C2	81	15062.00	1,220.02	1.89%
MH	43	2026.74	87.15	0.13%
PC1	1,110	7068.46	7,845.99	12.15%
PC2	35	3070.60	107.47	0.17%
RM	1,600	658.41	1,053.46	1.63%
S1	3,608	4146.02	14,958.85	23.16%
S3	3,522	1250.60	4,404.61	6.82%
URM	11,141	306.90	3,419.13	5.29%
W1	271,853	72.82	19,795.39	30.65%
W2	12,097	618.18	7,478.14	11.58%
Totals	303,006	211.08	64,592.57	100.00%

4.6. The Shelby County Building Inventory Database

Based on implementing the methodology for the various components of the dissertation, a comprehensive building inventory database for Shelby County was created. The database is extensively described with tabulated summaries in Appendix A. The building inventory was successfully applied in various loss estimation exercises for the MTB and in some structural class sensitivity analyses.

Chapter 5 . CONCLUSION AND VALIDATION

Various cutting-edge technologies, techniques and innovative methods from several sources were used in the course of this dissertation. The research methods were drawn from city and regional planning, mitigation planning, earthquake hazard risk assessment and loss estimation, computer science, pattern recognition, valuation, GIS technologies, software engineering and advanced statistics. The substantive parts of this dissertation are all components of earthquake risk assessment and loss estimation modeling and include models that classify buildings in a region by structure type, classify buildings by shape and estimate various aspects of building value. The earthquake risk assessment and loss estimation modeling process is rife with uncertainty, and the focus of this dissertation was to reduce the “factual’ uncertainty in the description of the at-risk building inventory, without which there can be no modeling effort. The artificial neural computing approach to structure type determination and the implementation of the detailed valuation methodology substantially reduce uncertainty in the description of the built environment. The shape recognition module is somewhat ahead of the current state-of-the-art in loss estimation modeling, since shape parameters have not yet been implemented in risk assessment studies at a region level. The methods were implemented in order to produce a building inventory database for Shelby County. This chapter summarizes the methods used in the research and includes a section on the validation of the building inventory produced for Shelby County. Limitations of the research are also discussed and the chapter concludes with this dissertation’s implications for future research.

5.1. Validation of Shelby County Building Inventory Data

Several methods were used for validating the building inventory dataset for Shelby County developed in this dissertation in order to demonstrate the application of the suggested models and for loss estimation exercises. Housing units and housing counts are validated by comparing the residential stock accounts between the building inventory and external sources, including the US Census. The structural classification is compared with earlier studies (including some performed for Shelby County). Since building footprint data did not exist for all structures in Shelby County, building class types were not included in the building inventory. However, the algorithm was validated by performance against digitally created building footprint libraries, described in the Results chapter. Building costs are validated by comparing the building inventory account to datasets derived from HAZUS MH MR-3.

5.1.1. Validating Residential Data using Dwelling Unit Comparisons

Several sources were used to judge the quality of the inventory produced using the methods developed in this dissertation. The gaps in square footage and/or dwelling unit information in the Tax Records was particular cause for concern – the limitations of the data required imputation procedures for estimating and accounting for the gaps.

The Tax Records had substantially fewer gaps in the non-residential and single-family residential portions of the database. Most of the gaps were found in multi-family residential parcels. When dwelling unit information was missing for multi-family residential buildings, but square footage was available, the number of dwelling units in the structure was imputed based on the average square footage per dwelling unit for the decade, and included quality control checks for similar (in terms of size and age) multi-family residential buildings in the vicinity of the imputed building. Instances with dwelling

unit information where square footage was missing were relatively rare. Where both dwelling unit and square footage information were missing, aerial photographs were inspected and footprints crudely digitized in order to extract footprint square footage. The length of shadows was used to determine the number of stories and therefore the total square footage. Then, the number of dwelling units was imputed using the process outlined above.

If imputation procedures are not appropriate, the resulting housing unit information could deviate significantly from established counts and projects. Accordingly, the counts of residential housing units were extracted from the building inventory database and compared with estimates from the US Census Bureau (US Census Bureau 2008), the American Community Survey (American Community Survey Office 2006, 2007) and other sources (City-Data.com 2008).

The sources for the US Census, City-Data and the American Community Survey data include Census of Population and Housing Population Estimates, Small Area Income and Poverty Estimates, State and County Housing Unit Estimates, County Business Patterns, Nonemployer Statistics, Economic Census, Survey of Business Owners, Building Permits and the Consolidated Federal Funds Report (US Census Bureau 2008).

Table 5.1 shows the comparison of housing units by residential occupancy from three different sources. While there is some discrepancy in the Duplexes and Triplexes/Quads residential occupancy classes, the numbers generally agree and follow increasing year trends.

Table 5.1 -- Validation of dwelling units by residential occupancy classes

<i>Residential Occupancies</i>		<i>Data Sources</i>		
Occupancy Description	Occupancy	Building Inventory 2008	US Census 2006	City-Data 2006
Single-family Residential	RES1	269,223	276,968	255,584
Multi-family Residential (2 units)	RES3A	15,245	9,815	10,617
Multi-family Residential (3-4 units)	RES3B	8,133	17,952	19,565
Multi-family Residential (5-9 units)	RES3C	30,782	32,643	28,297
Multi-family Residential (10-19 units)	RES3D	38,727	29,730	17,082
Multi-family Residential (20-59 units)	RES3E	25,097	9,060	27,351
Multi-family Residential (50+ units)	RES3F	11,328	13,904	
Mobile Homes	RES2	4,136	4,065	4,235
Total Multi-family Residential Housing Units		129,312	113,104	102,912
Total Housing Units		402,671	394,137	362,731
Single-family Residential	RES1	66.86%	70.27%	70.46%
Multi-family Residential (2 units)	RES3A	3.79%	2.49%	2.93%
Multi-family Residential (3-4 units)	RES3B	2.02%	4.55%	5.39%
Multi-family Residential (5-9 units)	RES3C	7.64%	8.28%	7.80%
Multi-family Residential (10-19 units)	RES3D	9.62%	7.54%	4.71%
Multi-family Residential (20-59 units)	RES3E	6.23%	2.30%	7.54%
Multi-family Residential (50+ units)	RES3F	2.81%	3.53%	
Mobile Homes	RES2	1.03%	1.03%	1.17%
Total Multi-family Residential Housing Units		32.11%	28.70%	28.37%
Total Housing Units		100.00%	100.00%	100.00%

Note in particular that the single-family to multi-family dwelling units proportion follows a trend towards a slightly higher number of multi-family residential, but are relatively consistent. In fact, the American Community Survey Office estimated the 2003 proportion between single-family and multi-family residential dwelling unit ratios at 63% to 26% (American Community Survey Office 2007).

Table 5.2 shows the estimated or recorded numbers of single-family units (based on building permits) that have been added to the Shelby County single-family stock since 2000. Note that the total post-2000 count of single-family construction for the building inventory is 28,612 while the City-Data.com data amounts to 29,357. The proportions by year are relatively consistent, especially in the 2003-2007 period.

Table 5.2 -- Validation using post-2000 single-family residential construction

Year	Building Inventory 2008		City-Data 2006	
	Counts	Percent	Counts	Percent
2000	4,171	14.58%	3,583	12.20%
2001	3,485	12.18%	3,450	11.75%
2002	3,680	12.86%	4,147	14.13%
2003	4,367	15.26%	4,587	15.62%
2004	4,490	15.69%	4,736	16.13%
2005	4,425	15.47%	4,769	16.24%
2006	3,994	13.96%	4,085	13.91%
Totals	28,612	100.00%	29,357	100.00%

Table 5.3 compares residential housing unit counts by decade of construction between the inventory produced by this dissertation against information derived from City-Data.com (2008). Both sources show a remarkable degree of consistency, being off in the counts or percentages by relatively small amounts, and the differences seem to decrease in more recent decades. All the tables are used for validating the building inventory developed in this dissertation for Shelby County, by comparing the building inventory from this dissertation to estimates from external sources (City-Data.com 2008). Figure 5.1 shows the percent of residential housing units constructed by decade for the tax-based building inventory and that of City-Data.com (ibid).

Table 5.3 -- Validation of residential housing units by decade

DECADE	Building Inventory 2008		City-Data 2006	
	Counts	Percent	Counts	Percent
1939 or earlier	37,613	9.68%	25,924	7.30%
1940 to 1949	26,425	6.80%	27,197	7.65%
1950 to 1959	56,143	14.46%	55,302	15.56%
1960 to 1969	62,693	16.14%	62,321	17.54%
1970 to 1979	82,582	21.26%	72,400	20.38%
1980 to 1989	60,824	15.66%	57,082	16.07%
1990 to 1999	62,115	15.99%	55,077	15.50%
Totals	388,395	100.00%	355,303	100.00%

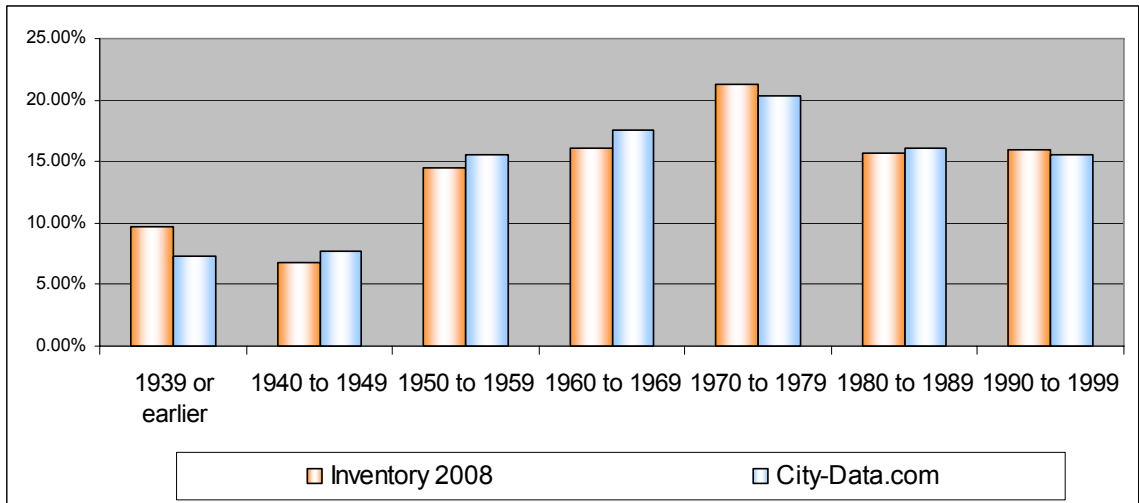


Figure 5.1 -- Residential housing units by decade

5.1.2. Validation of General Building Stock Characteristics

The distribution of structure type from the building inventory was compared with the results of earlier works on structure type classification (Malik 1995; Jones and Malik 1997). The Jones and Malik study results are compared with the building inventory, as seen in Table 5.4. Since the earlier study had only five structure classes, structure type counts were collapsed from the building inventory to the same classes to facilitate comparisons. As expected, Wood structures dominate the general building stock, increasing from 89% in 1994 to 93% in 2008. Light metal, Masonry and Concrete structure counts reduced to some extent, but not significantly, while Steel structures increased in counts significantly. The comparison is not made to explain a trend – rather, it is used to validate the classification results. The Jones and Malik study inferred structures directly from the tax records “supplemented with other information gathered from various sources” (Jones and Malik 1997, pp.13), while this dissertation modeled structure type based on primary and surveyed (calibration) data, so there are bound to

be significant differences, not attributable to temporal change. Second, the quality of Tax Records has significantly improved since 1997, with more details and complete records in a relational format – this again could cause significant differences in the general building stock. However, even with the differences, consistent patterns in structure type are apparent.

Table 5.4 -- Comparison with NCEER report (Jones and Malik 1997)

Structure Type	Codes	NCEER Study 1997	Building Inventory 2008
Wood	W	227,099	283,950
Light Metal	S3	9,427	3,522
Masonry	RM/URM	13,974	12,741
RCC	C1/C2/PC2/PC1	2,734	2,139
Prot. Steel	S1	463	3,608
Miscellaneous	Unknown/MH	2,377	43
Totals		256,074	306,003
Wood	W	88.68%	92.79%
Light Metal	S3	3.68%	1.15%
Masonry	RM/URM	5.46%	4.16%
RCC	C1/C2/PC2/PC1	1.07%	0.70%
Prot. Steel	S1	0.18%	1.18%
Miscellaneous	Unknown/MH	0.93%	0.01%
Totals		100.00%	100.00%

NCEER study by Barclay G. Jones and Ajay M. Malik (1994)

Table 5.5 compares the building inventory generated in this research with the results of an earlier Earthquake Engineering Research Institute study (Jones and Chang 1994). The table compares building counts, areas and replacement costs for residential and non-residential occupancies. While it is possible that significant numbers of buildings have been demolished and new ones built, the earlier study may have overestimated the counts. However, in terms of the relative proportions of residential and non-residential building counts, areas and replacement values, consistent patterns are seen in all three categories. Note in particular, while residential accounts for over

90% of the building stock, it accounts for only about 60% in area and replacement costs in both studies.

Table 5.5 -- Inventory validation by broad occupancy (Jones and Chang 1994)

		EERI Study 1994		Building Inventory 2008	
Type	Broad Use	Millions of Sq. ft.	Percent	Millions of Sq. ft.	Percent
Area	Residential	348.8372	57.68%	630.8122	62.58%
	Non-residential	255.9812	42.32%	377.2400	37.42%
	Total Buildings	604.8184	100.00%	1,008.0522	100.00%
Type	Broad Use	Number	Percent	Number	Percent
Counts	Residential	283,781	91.59%	288,107	94.15%
	Non-residential	26,074	8.41%	17,896	5.85%
	Total Buildings	309,855	100.00%	306,003	100.00%
Type	Broad Use	Millions of Dollars	Percent	Millions of Dollars	Percent
Repl. Cost	Residential	24.4151	56.69%	53.2384	61.00%
	Non-residential	18.6548	43.31%	34.0409	39.00%
	Total Buildings	43.0699	100.00%	87.2793	100.00%

EERI study by Barclay G. Jones and Stephanie E. Chang (1994)

Table 5.6 compares the general building stock characteristics generated in HAZUS MH MR-3 with those estimated in the building inventory. The HAZUS MR-3 data is current to 2002, while the building inventory is based on the Shelby County Tax Digest of 2007. Accordingly, all comparisons made between the HAZUS MR-3 and this building inventory datasets will be relevant for the period up to 2002. Note also that HAZUS MR-3 replacement costs are based on averages for one of 36 specific model types, while the building inventory uses a parameterized specification of additional model types, heights, external wall type and structural system for the estimation of replacement costs. The building inventory in this table includes only those structures built before 2002 and the replacement costs have been adjusted to 2002 costs, using the period adjustment specified in the Historical Cost Indexes section of the Means manual (R. S. Means 2008). Significant deviations between the two databases have been highlighted in the table.

Table 5.6 -- GBS characteristics from HAZUS MH MR-3 and study inventory

Specific Occupancy	Building Count		Area (thousands of sq. ft.)		Replacement Cost (billions)	
	Inventory	HAZUS	Inventory	HAZUS	Inventory	HAZUS
RES1	252,130	256,335	455,023.60	434,162.03	31.84	40.58
RES2	4,050	4,140	4,357.48	4,549.27	0.16	0.16
RES3A	7,594	5,298	13,394.65	15,883.50	0.95	1.05
RES3B	2,073	4,861	7,298.67	14,580.90	0.90	1.05
RES3C	4,171	2,812	26,838.39	22,475.69	3.41	2.88
RES3D	2,799	1,072	31,470.93	12,750.97	3.50	1.46
RES3E	856	186	17,651.84	5,425.43	1.73	0.61
RES3F	91	273	9,677.71	13,580.89	1.08	1.49
RES4	326	99	9,994.21	4,440.77	1.11	0.49
RES5	57	367	262.85	8,250.30	0.03	1.03
RES6	84	118	2,428.59	2,082.27	0.26	0.22
COM1	3,780	316	47,866.64	26,247.34	3.50	1.87
COM2	4,732	1,156	150,083.22	33,484.91	9.39	2.19
COM3	1,507	1,499	8,116.81	14,919.54	0.75	1.31
COM4	2,790	516	41,653.85	36,752.17	5.27	4.22
COM5	196	509	981.18	1,972.69	0.16	0.32
COM6	19	95	3,556.84	4,268.55	0.63	0.82
COM7	384	919	5,391.27	6,135.91	0.94	0.87
COM8	1,241	1,924	5,933.53	9,582.28	0.65	1.41
COM9	28	62	665.69	363.54	0.07	0.04
COM10	46	-	6,480.63	-	0.25	0.00
IND1	702	429	31,009.77	10,714.43	2.42	0.81
IND2	308	365	8,989.70	8,212.34	0.58	0.54
IND3	-	206	-	6,212.21	0.00	0.78
IND4	27	53	102.14	1,053.78	0.01	0.13
IND5	12	9	586.53	93.21	0.07	0.01
IND6	-	372	-	9,360.87	0.00	0.61
AGR1	-	210	-	2,784.84	0.00	0.18
REL1	1,000	863	8,261.81	14,691.63	0.97	1.75
GOV1	-	494	-	5,058.35	0.00	0.47
GOV2	48	54	1,381.06	526.09	0.19	0.08
EDU1	276	196	14,617.74	7,918.46	1.60	0.79
EDU2	16	40	23,200.00	1,387.35	2.52	0.17
Totals	287,336	285,848	937,277.34	739,922.53	74.93	70.37

For the most part, the data agrees between the two datasets, particularly in building counts (some differences are seen for some multi-family residential, retail trade, wholesale trade, commercial offices, hospitals, restaurants and medical offices). The Tax Records had little information on Institutional Dormitories (RES5) and Government Offices (GOV1). This results in a significant undercounting of these categories in the Tax inventory. In addition, HAZUS MR-3 models distributions of square footage for all occupancy types based on default mapping schemes, and there are bound to be large

deviations from reality, particularly in the analyses of small areas and by specific occupancy classes.

The replacement costs between the two databases are similar for all occupancies except Residential Single-family (RES1) and Wholesale Trade (COM2). Wholesale trade differences arise from the significant differences in square footage (and therefore building counts in HAZUS MR-3) between the two databases. In the single-family category, the difference in replacement costs amounts to about \$ 8.75 billions. HAZUS MR-3 models the distribution of basements based on a crude mapping scheme and sets 25% of all single-family residences as having full basements. The Tax Records indicated that most of the single-family buildings were built at grade on slab, and only about 87,729 buildings had basements – further, this was dominated by part-basements or crawl-spaces, not full basements. When the difference in basement costs were applied to the Tax-based inventory, the differences amounted to nearly \$ 3 billion. The second reason for the difference in replacement costs stems from the distribution of single-family buildings into Economy, Average, Custom and Luxury construction classes. HAZUS MR-3 does not provide documentation on the distribution, but analyses of the data suggest that the distribution is another example of a top-down mapping scheme based on census blockgroup geography-based income ratios developed at the state level. The Tax-based inventory based the distribution on the condition, desirability and utility classifications of the residence. Differences in the relative distributions could lead to significant differences in replacement costs, although it was not possible to quantify this difference. The third reason for the difference stems from the fact that the replacement cost in the dissertation was derived from the assumption that the building would be built in 2008 to the same specifications (of occupancy, height, structure and external wall type) as the original. A large number of the buildings in the Tax Records

had external walls of wood siding on wood frame or brick veneer on wood frame, resulting in lower per square foot costs when compared with the average per square foot costs that HAZUS MR-3 uses.

In terms of total building counts, the databases are almost exact. There is a 20% difference in square footage and a 6% difference in replacement costs for the entire inventory from the two databases. Considering that HAZUS MR-3 uses a top-down modeling approach, while the inventory is based on a bottom-up aggregation (and accounting for the gaps in the Tax Records), the estimates are fairly close.

Table 5.7 compares the cost per square foot for all the specific occupancy classes between the HAZUS MR-3 and the Tax-based inventory databases. Note that the per square foot costs are comparable (given the difference in years) between the two databases.

Table 5.7 -- Replacement costs per square foot comparisons

Specific Occupancy	Replacement Costs/sq. ft.		Specific Occupancy	Replacement Costs/sq. ft.	
	Inventory	HAZUS		Inventory	HAZUS
RES1 - ECONOMY	\$ 71.88	\$ 59.58	COM3	\$ 101.10	\$ 100.89
RES1 - AVERAGE	\$ 77.11	\$ 79.29	COM4	\$ 137.02	\$ 102.69
RES1 - CUSTOM	\$ 77.91	\$ 99.63	COM5	\$ 172.65	\$ 153.97
RES1 - LUXURY	\$ 85.00	\$ 117.55	COM6	\$ 192.32	\$ 144.60
RES2	\$ 40.00	\$ 30.90	COM7	\$ 184.39	\$ 129.82
RES3A	\$ 77.83	\$ 67.24	COM8	\$ 118.22	\$ 101.57
RES3B	\$ 134.50	\$ 73.08	COM9	\$ 115.85	\$ 102.35
RES3C	\$ 138.40	\$ 125.63	COM10	\$ 40.98	\$ 34.78
RES3D	\$ 121.19	\$ 112.73	IND1	\$ 84.80	\$ 73.82
RES3E	\$ 106.24	\$ 108.86	IND2	\$ 70.04	\$ 61.91
RES3F	\$ 120.50	\$ 111.69	IND4	\$ 90.02	\$ 78.61
RES4	\$ 121.22	\$ 104.63	IND5	\$ 124.06	\$ 119.51
RES5	\$ 124.60	\$ 113.31	REL1	\$ 127.77	\$ 114.08
RES6	\$ 116.10	\$ 104.62	GOV2	\$ 148.33	\$ 117.32
COM1	\$ 79.05	\$ 77.17	EDU1	\$ 119.07	\$ 95.21
COM2	\$ 68.01	\$ 59.24	EDU2	\$ 118.16	\$ 114.68

The square footages used in the building inventory fall comfortably in the ranges specified in the Means manual. Note that the per square foot costs for 3 of the 4 single-family categories are significantly higher in HAZUS MR-3, because of the exterior wall specification and the distribution of basements.

5.2. Applicability of Research Methods to Other Fields

ANNs offer great potential for application in several disciplines. While there are several parametric models to choose for a particular application, the very nature of parameterization renders them somewhat inflexible. However, the parameters are associated with quantitative measures and tests that enable detailed explanations of the relationships between the dependent variable and the explanatory ones. In contrast, ANNs are semi-parametric in nature, and are extremely flexible. Further, ANNs are far more forgiving of noisy or faulty or partly missing data than parametric models. However, the evaluation of ANN models is generally based on their performance, rather than on a quantitative relationship between inputs and outputs. Hence in applications where performance is emphasized over explanation, ANNs may be used with extremely good results.

ANNs may be used in a variety of applications that require function approximation, classification or time-series modeling. Thus, ANNs may be used in transportation modeling to determine mode choice, or in traditional urban planning to model land use change using historical data in combination with aerial imagery, or to approximate the process of land transformation and building conversion in an urban setting, or even in scenario-based urban development modeling for decision support. In fact, ANNs lend themselves to excellent applications in cellular automata that may be used with raster-based GIS for urban growth modeling. ANNs are increasingly being used in remote sensing and photogrammetric applications for land use and land cover

classification. While this dissertation uses ANNs in order to develop a classifier mechanism, ANNs have also been used effectively for function approximation and time-series modeling. Based on historic data, ANNs have been used to predict future trends and may be used in business applications. ANNs can model housing price using location, housing characteristics and spatial autocorrelation. In short, there are few substantive areas that will not benefit from using ANNs.

The shape recognition routines developed in this research may be extended to more complex shapes and for the development of building databases for seismic and wind hazards. Shape recognition routines may be used in combination with footprints generated through automated routines from aerial imagery in applications to predict the occupancy characteristics of buildings – this would particularly be useful in developing base data for regions that have no existing digital data. In other words, the relationship between spatial layouts of buildings and their usage may be tailored in a building shape recognition application to classify building features by occupancy.

Building valuation routines used in this research may be used to quantify and model public and private immovable assets and for portfolio management. In addition, the results of building valuation may be used to detect patterns in the spatial configuration of assets for a variety of applications.

5.3. Implications of the Research

An accurate accounting of the physical assets of a community is necessary for risk assessment and damage estimation modeling, but such accuracy is wanting. Much of the responsibility for mitigation planning and the reduction of community vulnerability lies with local governments, and if they are to play a greater role in mitigating hazard risks, a complete description of the at-risk built environment is a prerequisite. Clearly,

mitigation planning needs factual information about the built environment, but if the inputs to mitigation planning are suspect, then the estimates produced by any hazard modeling exercises in the context of mitigation planning would also be inaccurate. Additionally, research suggests that reducing uncertainty in hazard modeling has substantial benefits for mitigation through loss-avoidance regulations, code enforcement, design guidelines, directed land use planning and growth management and policy-making in general.

Bounds on the Accuracy of the Building Inventory

The process of seismic risk assessment is characterized by considerable aleatoric and epistemic uncertainty, particularly in the location of the seismic hazard, propagation of seismic energy, site response and building behavior. Given this context, is it worth incurring increasing costs in order to develop an accurate building inventory? Phrased another way, how accurate does the building inventory need to be, given that loss estimates are widely uncertain anyway? While there is no clear answer to this question, it is obvious that there are diminishing marginal returns in attempting to be completely accurate in the building inventory. To some extent, accuracy in the building inventory may be conceptualized as a function of the variable costs of collecting samples for structure type calibration (since model development and estimation are sunk costs, and data for building valuation are readily available from primary sources). Depending on the number of distinct structure types to be estimated, the total occupancy types and the total population of buildings, the costs for sample collection may vary – if external sources of funding are available for hazard mitigation planning, more samples may be acquired. Alternately, local jurisdictions facing seismic hazards could implement specific policies related to data collection, particularly in the context of the building permit approval and code enforcement process. Even without considering the uncertainty in

risk assessment, there are direct benefits in increasing the factual accuracy of the general building stock. Planners of local jurisdictions may get a clearer understanding of the spatial distribution of capital assets and more importantly, the distribution of structures particularly vulnerable to seismic hazards, such as old concrete or unreinforced masonry buildings. Analyzing these spatial distributions against the locations of hazardous areas would enable planners to direct land uses, manage growth, enforce seismic building codes for new construction and develop strategic measures for retrofitting vulnerable structures.

Funding for developing such accurate descriptions of the built assets is low, at present, and local governments are forced to make do with sub-optimal procedures or inaccurate data for their mitigation planning needs. Complete and comprehensive inventories of the building inventory are cumbersome to create and are time and resource intensive, primarily because the attributes of buildings for hazard modeling and loss estimation are not easily obtained. There exists therefore a significant need for new technologies and innovative methods aimed at both reducing uncertainty and costs in the modeling process. This research has demonstrated the development of an accurate and reliable building inventory at relatively low cost.

A second implication of this research is the utilization of a combination of primary and derived data in a bottom-up approach to inventory development. Primary data usually exists in some form with the Tax Assessor, while local planning jurisdictions often have aerial images, road, hydrography, rail and other planimetric and cadastral datasets. Non-taxable properties are not recorded by the Tax Assessor and require compilation from other sources (churches, schools, and other government buildings). Previous studies that model the building inventory use indirect methods, relying on consistent and systematic regularities in patterns of location, distribution of building

occupancies and relative sizes, and other associated characteristics like structural systems and height, etc. These approaches assume that while there may be large variation over individual buildings or in small areas, the regional or macro characteristics will remain constant. In general, these studies are performed for one or two specific regions in order to examine the pattern and then the pattern is replicated to model inventories for other regions. While the consistency is definitely observed at the regional scale, local and individual variation can potentially cause tremendous uncertainty in estimates of damage in loss estimation exercises. HAZUS MR-3 uses such indirect methods and models square footage distributions using a top-down approach (found in the various “mapping” schemes) and distributes the square footages to various occupancies and structure types. Of course, exogenous controls such as the census of demographics and housing and business accounts are used to guide the disaggregation process, but these controls are static over long time periods, while local growth trajectories lead to significantly different distributions of occupancies and structure types. This research argues that a bottom-up approach that uses primary data derived at the local level would be far more consistent and accurate and further, maintain that consistency and accuracy over time.

Geographic Bounds for Generalization of Building Inventory Models

While the bottom-up inventory modeling approach will produce more reliable accounts for the region, is there a geographic size limit to generalizing from the sample? In other words, if samples are collected from the City of Memphis, are the models accurate and valid for Shelby County, South-west Tennessee, the entire state of Tennessee, or a generally large continuous region around Memphis that includes parts of other states? At what scale does the bottom-up approach then become a top-down approach? Posed another way, could the model parameters estimated from samples

drawn from the City of Memphis be used to develop the building inventory data for the City of St. Louis, or for East Arkansas or Southern Illinois?

In general, if base conditions remain the same, model parameters estimated from local samples may be used for large regions or even other small regions elsewhere in the country. But there are too many local and regional variations for general use of models calibrated from local sample data to permit complete generalization. The valuation models for buildings are based on parameters such as occupancy, height, area and structure type as related to costs of construction – these patterns are observed at the individual building level (not for the region) and the models are parameterized for particular combinations of the input variables. Similarly, the shape recognition application depends on the shape of the building footprint that clearly does not change across regions. Consequently, valuation models and shape recognition applications may easily be generalized to other regions or scales without difficulty or concern over accuracy, after making regionally based adjustments to reflect transportation and constructed-related cost variations for the valuation component. Structure type models are more difficult to generalize. There are three primary factors limiting such generalization, including construction practices, building codes and geographic scales.

First, construction patterns and practices vary in response to climate, topography, availability and cost of raw materials and labor, etc. Thus, the types of structures one might see along a beachfront would be different from those in mountains or in dry-desert areas, and consequently, structure type models developed for Miami may not be easily generalized to Phoenix.

The adoption of building codes and standards of construction at the local level influences the types of structures and the resulting costs of protecting the population against specific disaster risks. For instance, the State of California has mandated the

necessity of mitigating against earthquake risk through adoption of the International Building Code (ICC 2000) for individual building standards and the incorporation of the seismic safety element in comprehensive planning for general planning. Shelby County and the City of Memphis, TN have been resisting the adoption of the International Building Code, citing associated increases in costs of construction that could drive potential investors away and inhibit local economic development. While economic development is indeed always a concern, mitigation measures that influence the life-safety of local populations are heavily reliant on risk attitudes and understanding of risk on the part of local decision-makers. In mid-America, the last major seismic event occurred almost two centuries ago, and local decision-makers tend to adopt pro-risk attitudes by comparing the present-day tangible costs of incorporating seismic safety against the benefits of an event in the future (ranging over 3000 years) that might not happen, or has a very low probability of occurrence or that might occur elsewhere in the vicinity, causing little damage in their jurisdiction. This results in an uneven distribution of structures over time across regions. Consequently, model parameters derived from samples in San Fernando, CA may not be effectively generalized to produce inventory estimates for Shelby County, TN.

Closely associated with legislative mandates for seismic safety are how laws and codes are enforced. This aspect is harder to measure – for instance, the definition of Unreinforced Masonry varied across cities and regions (unreinforced infill walls within a reinforced frame or unreinforced bearing walls or even unreinforced exterior veneers on wood frames), so when legislation disallowed the construction of these structures (based on date of adoption), the momentum of traditional construction practices coupled with nomenclature confusion prevented the consistent enforcement of masonry building codes across regions. Differences in the code adoption dates, code enforcement

training and local attitudes to code enforcement also result in regional variations of construction types. Again, these reasons could potentially prevent easy generalization of structure type models across regions.

Another aspect reflecting choice in structure type is related to the scale of the local region and its growth patterns. There are clear differences in the building stock for large cities like Memphis, TN and Los Angeles, CA compared with isolated urban places, cities or towns with corresponding smaller populations or rural regions. Similarly, there are differences in building stock by structure type and occupancy based on the speed of growth – a rapidly growing region would be characterized by more residential stock first and then commercial-industrial stock, while slow-growth areas demonstrate stable patterns of building types. Thus, the region's growth rate would reflect occupancy and structure type patterns that differ from other rates, and generalization could be implemented across regions after adjusting for growth rates. Overall, this research recommends developing structure type inventory based on samples drawn from the local region being modeled.

5.3.1. Specific Implications

This study derives three major components that may be used for loss estimation modeling – models for the classification of buildings by structural system, algorithms for recognizing the shape of buildings from their footprints and estimation techniques for building valuation.

For the structure type classification, the research established the consistency between a traditional parametric approach using a multinomial logistic regression specification and a numerical basis approach using artificial neural network models. While ANNs do show marginally better performance results than multinomial logistic

regression approaches, the logistic models have far greater explanatory power and enable greater understanding, interpretation and evaluation of the relative contributions of the independent variables to particular outcomes. Depending on the application, the marginally higher performance of ANNs may not be worth the complete loss of explanatory power and parametric evaluation permitted by multinomial logistic regression approaches. In this research, the multinomial logistic regression model serves as a vehicle to formally express the relationships between the inputs and the structure type outcomes. The inherent flexibility and noise-forgiving nature of the semi-parametric neural computing approaches allow for slightly better performance and implementation in the production datasets, so ANN specifications were used to implement structure type classifications for the building inventory production dataset.

Thus, local communities can collect survey data on structure type of buildings within their jurisdictions, which while expensive is far less time or resource consuming than surveying all the buildings or relying on inaccurate existing data and/or sub-optimal methods. Subsequently, structure type classifications may be calibrated and validated using the survey data, and the parameters applied to the unseen buildings in order to estimate the structure type for the entire building stock.

One of the outputs of the structure type classification models is a logistic or hyperboloid tangent function magnitude that lies in the 0 to 1 range, and as such, may be interpreted as the probability of the given input combination that realize that particular structure type outcome. Thresholds may be set for the magnitude of the probability, below which further investigation may be warranted. Additionally, as described in detail in Section 4.2.1.5, combinations of low probability scores and competing probability scores should be examined in greater detail, particularly for Unreinforced Masonry structures. Since resources for retrofitting are limited, local communities should pursue

a strategic approach and identify those structures that are most vulnerable to earthshaking hazards, and particularly those that pose threats to life safety. Utilizing the probabilistic output of the classification models would enable the design of risk reduction strategies, reduce the uncertainty in structure type classification through follow-up surveys for ambiguous classifications and enhance the process of mitigation planning.

The dissertation also demonstrated the implementation of innovative spatial computation techniques for building configuration recognition from building footprints in the GIS environment. To date, this type of automated shape recognition has not been developed for building footprints in the GIS arena. The syntactic approach of landmark correspondence, whose roots are derived from pattern recognition, is less computationally intensive and more efficient, and has been automated. The performance of the algorithm for manually digitized building footprints is excellent, while the algorithm faces some difficulty in recognizing footprints that were automatically extracted from aerial imagery. In its defense, classifying these automatically extracted footprints was difficult even for humans.

Finally, the dissertation estimated the replacement value of buildings by curve fitting routines that parameterize per square foot construction costs by occupancy, height, exterior wall and structural frame type. Further, the replacement costs were decomposed into structural, nonstructural acceleration- and drift-sensitive costs, based on the different assembly costs specified in Means and on classifications uncovered in the literature review. Note that the structure type was derived from the ANN specifications and then used in combination with other primary data for estimating building replacement costs.

In general, the methods that generate the specific components required as attributes for the building in loss estimation modeling are all replicable and produce

reliable, consistent and accurate building inventory data at a fraction of the costs of traditional survey methods. Further, while the benefits of reduced uncertainty in the context of mitigation planning do not lend themselves for easy quantification, studies show that they do exist and are substantial.

5.3.2. Limitations of the Research and Future Directions

5.3.2.1. Limitations in Structure Type Classification Modeling

One primary limitation of the research was that the structure type classification module was limited to 11 categories. Several studies have parameterized the behavior of several other structure types not found in this dissertation. Specifically, Steel structures may be further classified as Steel Moment Frame, Steel Braced Frame, Steel Frame with Concrete Shear Walls, and Steel Frame with Unreinforced Masonry Infill Walls. The main reason that the study limited itself to just one type of steel frame building was that detecting the other steel frame types was not possible in the context of a windshield survey. However, a combination of the external wall data from the Tax Records and the structure type classification module could be used to subdivide steel frame into the appropriate sub-classes. The same argument holds for Concrete Frames with Unreinforced Masonry Infill Walls and Reinforced Masonry Bearing Walls with Precast Concrete Diaphragms. The structure type module could still be used with the parsimonious specification, and the Concrete, Steel and Reinforced Masonry categories could be subdivided into other categories using database searches of external wall specifications. The disadvantage of increasing the number of structure type categories is that the calibration sample counts need to be substantially higher, with at least 10 exemplars for each occupancy-structure type combination, and this cannot be determined a priori, so additional field surveys may need to be performed to fill gaps in the occupancy-structure type cross-tabulations.

Related to the structural component, another limitation of the research in terms of building behavior under earthshaking stresses is that this dissertation does not model or identify the foundation type of the building. The type of foundation is a key attribute that influences the capacity of a building to withstand ground motion stresses and is much more difficult to identify, because it is hidden from view. Modeling foundation type was beyond the scope of this dissertation, and would clearly add value to the process of estimating damage to a building under earth-shaking.

While the structure type module does classify structures adequately and demonstrate the potential for automation, a future extension of this module could embed the MLP ANN engine within a GIS framework. Local community planners could develop a preliminary inventory with the necessary variable specifications for structure type classifications along with the calibration and validation sample data, train the embedded ANN engine to develop parameters and implement them to classify the entire building inventory by structure type, all without leaving the GIS environment.

5.3.2.2. Limitations in the Shape Recognition Application

One limitation with the shape recognition module is that it fails to identify separate buildings that exist very closely. In other words, if two distinct rectangular buildings are located at the same orientation and separated by 6 inches, the shape recognition application would recognize them as a single building and not as two separate buildings. Technically, this is a limitation of the shape extraction component that is beyond the scope of this dissertation, and does not really reflect inadequacies in the shape recognition module. The building footprint is captured either through manual digitization or through automatic extraction routines from aerial photographs. Small gaps of less than 2 feet between buildings would be observed indistinctly as linear pixels, and in some cases, not observed at all, depending on the scale of the buildings, the

resolution of the input image and shadow artifacts in the input image. If such gaps are not noticeable by the human eye from the image, the extraction process would treat the two structures as a single building, even if they were of different heights. A similar argument holds for the non-recognition of expansion joints within the same structure or expanded structures. The identification of close, but separate buildings is important for two reasons. First, the response of the building to ground motion depends on its shape, and the overall shape of the composite structure would be identified in a different category than each of the buildings in close proximity. Second, if the two buildings were built at different times or to different design specifications (such as differences in interstory height, or structure type), ground motion translated to interstory drift would result in one structure impinging or pounding against others in close proximity. Additionally, failure of one building could potentially cause failure of other proximal buildings. The overall effect of not recognizing buildings in close proximity as separate structures could potentially underestimate damage in these building groups. In general, local regulations on minimum setbacks and distances between structures prevent occurrences of structures built in close proximity. However, in older areas of cities, several “row” type structures are observed as typical commercial/retail establishments – these structures may have been built at different times or to different specifications and are rarely captured or extracted as individual building segments. Typical expansions to existing structures like hospitals also follow similar patterns – the expanded structure may be built to different specifications and separated from the original structure by a thin expansion joint. In summary, the shape recognition module does not identify separate buildings that exist in very close proximity.

The shape recognition algorithms did not perform as well for buildings automatically extracted from aerial photographs, owing to noise in the exterior contour of

the building footprint. One area for future research could be to achieve a higher degree of generalization in the building footprint extraction process using smoothing routines. Another potential area for future research could be to use the extracted footprints in vector form in Fast Fourier or Wavelet Transformation routines or even design specific generalization and smoothing routines to smooth the external contour and then apply the shape recognition routines to classify the building. A third study could be directed at designing simplification algorithms to use relative measures such as line segment length ratios or subtended polygon area ratios, in order to lessen the sensitivity of generalization to scale.

The preprocessing components for the shape recognition module have been designed to eliminate very small and well specified protrusions and intrusions in the contour of the building footprint. Currently, the preprocessing routines do not eliminate successive convexities or concavities or combinations of successive convexities and concavities, nor do they eliminate concavities or convexities at building corners. Additional routines could be designed and written to comprehensively eliminate all types of convexities and concavities along any region of the contour. The geometry of the final simplified building footprint configuration could then be used to derive the exact locations of the centers of gravity and shear for the building in order to determine loading eccentricity.

A third area of future research in shape recognition that is not covered in this dissertation is a process to reconcile classes based on generalization thresholds (areal percentages or length-based measures) in order to resolve class ambiguities. For instance, if the stem of a T-shaped building forms less than 10% of the building's total footprint area, then that building's class could be adjusted from a T-shape to a rectangular shape. This would require input from structural engineers in order to

determine boundary thresholds and rules to adjust classes between pairs of shapes (L-shape to rectangle, Z-shape to T-shape, etc.).

The shape recognition module of this research limits itself to the two-dimensional classification of building footprint configurations. For damage assessment and modeling, the three-dimensional massing is equally important. Future research could be directed to develop a typology of massing characteristics for all the two-dimensional classes (symmetric L-shape, minor asymmetric L-shape along longer/shorter dimension, major asymmetric L-shape along longer-shorter dimension, etc.) and then develop methods to classify buildings using a combination of shape configuration and massing.

Shape configurations of buildings have been analyzed in terms of concavities affecting the behavior under shaking stresses for individual buildings, but these studies are few in number. A key area of damage modeling could estimate the behavior of various building shape classes by relating massing, shape, the number of concavities and loading eccentricities to induced damage. The results could be simplified and parameterized for efficient loss estimation at a regional level.

5.3.2.3. Limitations in Building Valuation Modeling

The separation of replacement costs into structural and nonstructural acceleration- and drift-sensitive components in this research was based on broad construction assemblies. Substantial research has been directed in the recent past (Porter 2005) that go beyond the primary assemblies into very specific and detailed component sub-assemblies that may be parameterized by occupancy class. Additional areas of research could include surveys of several occupancy classes for detailed sub-assemblies in order to develop combined fragilities for sub-assembly groups. This approach would not only enable the modeling of damage to nonstructural components

with reduced uncertainty, but also help in developing better engineering design guidelines for connections and anchorages between sub-assemblies. Finally, the lack of available data precluded the estimation of content value by specific occupancy class in this dissertation. Future research could therefore be directed to procure content value data, calibrated against damage to contents in the context of a real earthquake.

Despite all these limitations, the research methods were extensively validated and demonstrate effectively that advanced technologies and methods may be effectively and innovatively applied on combinations of primary and derived data and replicated in order to produce a bottom-up, reliable, accurate and cost-effective building inventory.

APPENDIX A . Tabulated Summaries and Descriptions of the Shelby County Building Inventory

This appendix outlines the structure of the Memphis Test Bed building inventory. The inventory covers all of Shelby County, TN and contains details on the mapping of “building use” categories from the most specific uses through HAZUS occupancy categories to “broad occupancy” categories that are used for presenting summaries in this appendix. The tabulations also contain frequency tables for the building counts by structure type and broad occupancy. Other variables are then summarized as two-way cross-tabulations, usually by structure type and/or occupancy. Each cross tabulation was also provided as a separate worksheet in a workbook for dissemination.

There are a total of 346,393 parcels in the Tax Records. Of these, 54,841 parcels did not have any structures (as derived from the Tax Records). These vacant parcels comprise mainly of parcels designated by the Shelby County Tax Assessor as “Accessory Improvements”, “Cell Tower sites”, “Cemeteries”, “Common Areas” and “Parking” for Multi-family or Condominium parcels, “Tax-exempt” and “Vacant Land” – 39,657 parcels classified as “Vacant Land” in the Tax Records did not have any built structures, and therefore may be regarded as undeveloped.

The building inventory database contains a total of 291,552 land parcels with built structures ranging from a maximum of 202 to a minimum of 1. Since some land parcels have more than one building, there are 306,003 building records in the dataset. Each building inventory database record corresponds to a single building. Each building may be uniquely identified by the field BLDG_ID, designated as a primary key for the building database.

The inventory is described by the following tables:

Table	Description
Table A.1	Attribute Schema for the Shelby County Building Inventory
Table A.2	Mapping Specific Occupancy to HAZUS Occupancy Classes
Table A.3	Mapping HAZUS Occupancy Categories to General Occupancy Types
Table A.4	General Structure type (Frequency Table)
Table A.5	General Occupancy (Frequency Table)
Table A.6	Cross-tab of Structure type and General occupancy (counts)
Table A.7	Cross-tab of Structure type and General occupancy (percentages)
Table A.8	Cross-tab of Structure type and Number of Stories
Table A.9	Cross-tab of Structure type and Year Built
Table A.10	Cross-tab of Structure type and Basement class
Table A.11	Cross-tab of General occupancy and Basement class
Table A.12	Cross-tab of Structure type and Square Footage class
Table A.13	Cross-tab of Structure type and Replacement cost class
Table A.14	Building Replacement Costs (in millions) by Structure type and General occupancy
Table A.15	Cross-tab of Structure type and Content Value class
Table A.16	Cross-tab of Structure type and Essential Facility designation
Table A.17	Cross-tab of General occupancy and Number of Dwellings in structure
Table A.18	Building Counts by Structure Type and Number of Dwellings in Building

The descriptions of the inventory are specific for the Memphis Test Bed, Shelby County, Tennessee. The building inventory database shown here contains the integrated results of the implementation of the various modules described in the research. Specifically, the structure type was determined using primary Tax Assessor's records in an ANN framework with production parameters calibrated and validated by field surveys. The final structure type is classified as eleven different types, derived from the ANN and analyses of the Tax Records. Inputs to the structure type classification model included building area, number of stories, year of construction, presence in a historic zone, occupancy and fire rating category. Since we did not have a spatial dataset of building footprints, building shape could not be recorded as part of the inventory database. Using R.S. Means Square Foot Costs for 2008 for a variety of occupancies, heights, external wall and structure type combinations, parametric curves were estimated in order to determine the replacement costs of each building. The

replacement costs were further decomposed into structural, nonstructural acceleration- and drift-sensitive components and content value, and recorded in the same database.

Table A.1 -- Attribute schema for building inventory dataset

FieldName	Description
PAR_ID	Parcel Identifier (Duplicates allowed for multiple buildings in the same parcel)
PARID_CARD	Improvement Identifier (Duplicates allowed for identical buildings in the same parcel)
BLDG_ID	Building Identifier (Unique, Primary Key Constraint -- No Duplicates allowed for this field)
LAT	Latitude of Parcel Centroid in Geographic Coordinate System, NAD 1983
LON	Longitude of Parcel Centroid in Geographic Coordinate System, NAD 1984
STR_TYPE	General Structure Type (used for summarized tabulations in this workbook)
STR_PROB	Structure Type Probability score derived from the Artificial Neural Network Model
YEAR_BLT	Year of building construction
STORIES	Total number of stories for the building
A_STORIES	Total number of above-ground stories for the building
B_STORIES	Total number of below-ground stories for the building
BSMT_TYPE	Basement type
SQ_FEET	Total building area in square feet
GSQ_FEET	Total ground floor area for the building in square feet (computed)
NO_DU	Total number of dwelling units in the building
EF	Essential Facility designation
APPR_VAL	Appraised value for the building in dollars, inherited from Tax Records (incomplete)
REPL_CST	Replacement cost in dollars for the building from R.S.Means Square Foot Costs 2008
STR_CST	Structural component of the replacement cost in dollars
NSTRA_CST	Acceleration-sensitive component of the replacement cost in dollars
NSTRD_CST	Drift-sensitive component of the replacement cost in dollars
CONT_VAL	Value of building contents in dollars
DGN_LVL	Design-level for the building as per HAZUS MR-3 specifications
OCC_TYPE	Broad HAZUS Occupancy Category -- Multi-family Residential specified by "RES3" only
OCC_DETAIL	Specific Occupancy Category, describing the detailed use of the building
MAJOR_OCC	Major Occupancy category for the parcel in which the building is sited
BROAD_OCC	General Occupancy categories (used for summarized tabulations in this workbook)
IMPUTED	Imputed record designator, used to complete the building database
XCOORD	X-Coordinate of the building in Tennessee State Plane, NAD 1983, feet
YCOORD	Y-Coordinate of the building in Tennessee State Plane, NAD 1983, feet
STR_TYP2	Detailed Structure Type as per HAZUS MR-3 specifications
OCC_TYPE2	Detailed HAZUS Occupancy Category for the building
TRACT_ID	Census Tract Identifier in which the building is located
CT_LAT	Latitude of Census tract in which the building is located
CT_LON	Longitude of Census tract in which the building is located

**Table A.2 -- Mapping Tax Record-based specific occupancy to HAZUS MH MR-3
specific occupancy categories**

S. No.	Detailed Use of Building	No. of Bldgs	HAZUS Occupancy	General Occupancy
1	APT <100 UNITS	2,048	RES3	Multi-family Residential
2	APT >100 UNITS	6,323	RES3	Multi-family Residential
3	APT HI-RISE	36	RES3	Multi-family Residential
4	AUTO DEALER/F-SERVICE	170	COM2	Wholesale Trade
5	AUTO SERVICE GARAGE	972	COM3	Light Industrial
6	BANK	220	COM5	Office Commercial
7	BAR/LOUNGE	96	COM8	Food and Entertainment
8	BOWLING ALLEY	5	COM8	Food and Entertainment
9	BRDING-ROOMING HOUSE	45	RES5	Multi-family Residential
10	CAR WASH - AUTOMATIC	120	COM3	Light Industrial
11	CAR WASH - MANUAL	107	COM3	Light Industrial
12	CINEMA/THEATER	26	COM9	Light Industrial
13	CLUB HOUSE	367	COM8	Food and Entertainment
14	COLD STORAGE	15	IND2	Light Industrial
15	COLLEGES	16	EDU2	Education
16	COMM SHOPPING CENTER	105	COM1	Retail Trade
17	CONDO UNIT	1,096	RES3	Multi-family Residential
18	CONVENIENCE FOOD MKT	446	COM1	Retail Trade
19	COUNTRY CLUB	32	COM8	Food and Entertainment
20	CULTURAL FACILITIES	2	COM9	Light Industrial
21	DAY CARE CENTER	181	COM3	Light Industrial
22	DEPARTMENT STORES	106	COM1	Retail Trade
23	DOWNTOWN ROW TYPE	254	COM1	Retail Trade
24	DUPLEX	6,608	RES3	Multi-family Residential
25	ECONOMY APTS	858	RES3	Multi-family Residential
26	FIRE STATIONS	13	GOV2	Office Commercial
27	FLEX WAREHOUSE	197	IND2	Light Industrial
28	FRANCHISE FOOD	494	COM8	Food and Entertainment
29	FUNERAL HOME	37	COM4	Office Commercial
30	HANGAR	14	IND2	Light Industrial
31	HEALTH SPA	11	COM8	Food and Entertainment
32	HOSPITALS	22	COM6	Health Care
33	HOTEL/MOTEL HI RISE	38	RES4	Multi-family Residential
34	HOTEL/MOTEL LO RISE	293	RES4	Multi-family Residential
35	KWIK LUBE	43	COM3	Light Industrial
36	LIBRARY	6	COM4	Office Commercial
37	LUMBER STORAGE	3	COM2	Wholesale Trade
38	MFG/PROCESSING	736	IND1	Heavy Industrial
39	MINI WAREHOUSE	1,016	COM2	Wholesale Trade
40	MOBILE HOME PARK	43	RES2	Multi-family Residential
41	NBHD SHOPPING CENTER	153	COM1	Retail Trade

**Table A.3 -- (cont'd from previous) Mapping Tax Record-based specific occupancy
to HAZUS MH MR-3 specific occupancy categories**

S. No.	Detailed Use of Building	No. of Bldgs	HAZUS Occupancy	General Occupancy
42	NIGHT/CLUB/DNR THEATER	17	COM8	Food and Entertainment
43	NURSING HOME	65	RES6	Multi-family Residential
44	OFFICE BLDG H-R 5ST	86	COM4	Office Commercial
45	OFFICE BLDG L/R 1-4S	1,994	COM4	Office Commercial
46	OFFICE CONDOMINIUM	781	COM4	Office Commercial
47	OFFICE MEDICAL	358	COM7	Office Commercial
48	PARKING GARAGE/DECK	50	COM10	Parking Garage
49	POLICE STATIONS	35	GOV2	Office Commercial
50	PREFAB WAREHOUSE	1,493	COM2	Wholesale Trade
51	RADIO/TV TRANSMITTER BLD	13	IND2	Light Industrial
52	RADIO/TV/MIN PIC STUDIO	2	IND2	Light Industrial
53	RAIL/BUS/AIR TERMINAL	2	IND2	Light Industrial
54	RECREATIONAL/HEALTH	29	COM8	Food and Entertainment
55	REGIONAL SHPMALL/CNT	17	COM1	Retail Trade
56	RELIGIOUS	1,021	REL1	Places of Worship
57	RES ON COMM LAND	1,082	RES1	Single-family Residential
58	RESEARCH & DEVELOPMENT	14	IND5	Light Industrial
59	RESTAURANT	248	COM8	Food and Entertainment
60	RETAIL CONDOMINIUM	15	COM1	Retail Trade
61	RETAIL MULTI OCCUP	503	COM1	Retail Trade
62	RETAIL SINGLE OCCUP	1,932	COM1	Retail Trade
63	RETIREMENT CENTER	21	RES6	Multi-family Residential
64	SCHOOL	280	EDU1	Education
65	SERVICE STATION FULL SERVICE	179	COM3	Light Industrial
66	SINGLE-FAMILY RESIDENTIAL1	118,140	RES1	Single-family Residential
67	SINGLE-FAMILY RESIDENTIAL2	128,273	RES1	Single-family Residential
68	SINGLE-FAMILY RESIDENTIAL3	15,149	RES1	Single-family Residential
69	SINGLE-FAMILY RESIDENTIAL4	6,801	RES1	Single-family Residential
70	SKATING RINK	10	COM8	Food and Entertainment
71	SOCIAL/FRATERNAL HALL	14	RES5	Multi-family Residential
72	STORE-RETAIL	6	COM1	Retail Trade
73	STRIP SHOPPING CNTR	415	COM1	Retail Trade
74	SUPERMARKET	68	COM1	Retail Trade
75	SWIMMING-INDOOR POOL	3	COM8	Food and Entertainment
76	TELEPHONE EQUIPMENT BLDG	5	IND2	Light Industrial
77	TENNIS CLUB - INDOOR	10	COM8	Food and Entertainment
78	TOWNHOUSE	946	RES3	Multi-family Residential
79	TRIPLEX	218	RES3	Multi-family Residential
80	TRUCK TERMINAL	76	IND2	Light Industrial
81	VETERINARY CLINIC	50	COM7	Office Commercial
82	WAREHOUSE	2,209	COM2	Wholesale Trade
Total		306,003		

**Table A.4 -- Mapping HAZUS MH MR-3 occupancy categories to general
occupancy classes**

S. No.	HAZUS Occupancy	HAZUS Occupancy Description	Building Count	Percent Buildings	General Occupancy
1	COM1	Retail Trade	4,020	1.31%	Retail Trade
3	COM2	Wholesale Trade	4,891	1.60%	Wholesale Trade
4	COM3	Personal and Repair Services	1,576	0.52%	Light Industrial
5	COM4	Professional/Technical Services	2,930	0.96%	Office Commercial
6	COM5	Banks	220	0.07%	Office Commercial
7	COM6	Hospital	22	0.01%	Health Care
8	COM7	Medical Office/Clinic	408	0.13%	Office Commercial
9	COM8	Restaurants and Bars	1,322	0.43%	Food and Entertainment
10	COM9	Theaters	28	0.01%	Light Industrial
2	COM10	Parking Garages	50	0.02%	Parking Garage
11	EDU1	Education (Graded Schools)	280	0.09%	Education
12	EDU2	Education (Colleges)	16	0.01%	Education
13	GOV2	Emergency Services (Police/Fire/EOC)	48	0.02%	Office Commercial
14	IND1	Heavy Industrial	709	0.23%	Heavy Industrial
15	IND2	Light Industrial	324	0.11%	Light Industrial
16	IND4	Food/Drugs/Chemicals	27	0.01%	Light Industrial
17	IND5	High Technology	14	0.00%	Light Industrial
18	REL1	Place of Worship	1,021	0.33%	Places of Worship
19	RES1	Single-family Residential	269,442	88.05%	Single-family Residential
20	RES2	Mobile Home	43	0.01%	Multi-family Residential
21	RES3A	Multi-family Residential (2 units)	7,026	2.30%	Multi-family Residential
22	RES3B	Multi-family Residential (3-4 units)	1,441	0.47%	Multi-family Residential
23	RES3C	Multi-family Residential (5-9 units)	1,972	0.64%	Multi-family Residential
24	RES3D	Multi-family Residential (10-19 units)	2,100	0.69%	Multi-family Residential
25	RES3E	Multi-family Residential (20-59 units)	3,132	1.02%	Multi-family Residential
26	RES3F	Multi-family Residential (50+ units)	2,464	0.81%	Multi-family Residential
27	RES4	Temporary Lodging (Hotel/Motel)	331	0.11%	Multi-family Residential
28	RES5	Institutional Dormitory	59	0.02%	Multi-family Residential
29	RES6	Nursing Home	87	0.03%	Multi-family Residential
Totals			306,003	100.00%	

n.b. Total Multi-family residential units = 18,135 or 5.93%

Table A.5 -- General Structure type (Frequency table)

General Structure Type	Code	No. of Buildings	Percent
Concrete Moment Resisting Frame	C1	913	0.30%
Concrete Frame with Concrete Shear Wall	C2	81	0.03%
Manufactured Home	MH	43	0.01%
Concrete Tilt-up	PC1	1,110	0.36%
Precast Concrete Frame	PC2	35	0.01%
Reinforced Masonry	RM	1,600	0.52%
Steel Frame	S1	3,608	1.18%
Light Metal Frame	S3	3,522	1.15%
Unreinforced Masonry	URM	11,141	3.64%
Light Wood Frame	W1	271,853	88.84%
Commercial Wood Frame	W2	12,097	3.95%
Totals		306,003	100.00%

Table A.6 -- General occupancy classes (Frequency table)

General Occupancy Category	No. of Buildings	Percent Buildings
Education	296	0.10%
Food and Entertainment	1,322	0.43%
Health Care	22	0.01%
Heavy Industrial	709	0.23%
Light Industrial	1,969	0.64%
Multi-family Residential	18,643	6.09%
Office Commercial	3,605	1.18%
Parking Garage	50	0.02%
Places of Worship	1,021	0.33%
Retail Trade	4,013	1.31%
Single-family Residential	269,464	88.06%
Wholesale Trade	4,889	1.60%
Totals	306,003	100.00%

Table A.7 -- Building counts and percentages by Structure type and General occupancy

Structure Type	Schools and Colleges	Food and Entertainment	Health Care	Retail Trade	Wholesale Trade	Office Commercial
C1	268	10	7	35	51	436
C2	-	-	1	-	1	45
MH	-	-	-	-	-	-
PC1	-	6	-	73	739	50
PC2	-	-	-	-	-	-
RM	-	95	-	252	758	22
S1	13	109	11	1,271	210	1,063
S3	2	35	-	249	2,206	133
URM	4	157	1	1,668	788	384
W1	3	744	-	344	79	1,096
W2	6	166	2	121	57	376
Totals	296	1,322	22	4,013	4,889	3,605
Percent	0.10%	0.43%	0.01%	1.31%	1.60%	1.18%
Structure Type	Heavy Industrial	Light Industrial	Places of Worship	Parking Garage	Multi-family Residential	Single-family Residential
C1	63	5	2	1	35	-
C2	-	-	-	-	34	-
MH	-	-	-	-	43	-
PC1	86	142	-	-	14	-
PC2	-	1	-	34	-	-
RM	-	471	1	-	1	-
S1	108	211	74	15	523	-
S3	342	510	11	-	34	-
URM	105	609	36	-	1,068	6,321
W1	5	15	68	-	9,731	259,768
W2	-	5	829	-	7,160	3,375
Totals	709	1,969	1,021	50	18,643	269,464
Percent	0.23%	0.64%	0.33%	0.02%	6.09%	88.06%

Table A.8 -- Building percentages by Structure type and General occupancy

Structure Type	Schools and Colleges	Food and Entertainment	Health Care	Retail Trade	Wholesale Trade	Office Commercial
C1	0.09%	0.00%	0.00%	0.01%	0.02%	0.14%
C2	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%
MH	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
PC1	0.00%	0.00%	0.00%	0.02%	0.00%	0.02%
PC2	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
RM	0.00%	0.03%	0.00%	0.08%	0.25%	0.01%
S1	0.00%	0.04%	0.00%	0.42%	0.07%	0.35%
S3	0.00%	0.01%	0.00%	0.08%	0.72%	0.04%
URM	0.00%	0.05%	0.00%	0.55%	0.26%	0.13%
W1	0.00%	0.24%	0.00%	0.11%	0.03%	0.36%
W2	0.00%	0.05%	0.00%	0.04%	0.02%	0.12%
Percent	0.10%	0.43%	0.01%	1.31%	1.60%	1.18%
Structure Type	Heavy Industrial	Light Industrial	Places of Worship	Parking Garage	Multi-family Residential	Single-family Residential
C1	0.02%	0.00%	0.00%	0.00%	0.01%	0.00%
C2	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%
MH	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%
PC1	0.03%	0.05%	0.00%	0.00%	0.00%	0.00%
PC2	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%
RM	0.00%	0.15%	0.00%	0.00%	0.00%	0.00%
S1	0.04%	0.07%	0.02%	0.00%	0.17%	0.00%
S3	0.11%	0.17%	0.00%	0.00%	0.01%	0.00%
URM	0.03%	0.20%	0.01%	0.00%	0.35%	2.07%
W1	0.00%	0.00%	0.02%	0.00%	3.18%	84.89%
W2	0.00%	0.00%	0.27%	0.00%	2.34%	1.10%
Percent	0.23%	0.64%	0.33%	0.02%	6.09%	88.06%

Table A.9 -- Building counts and percentages by Structure type and Number of stories

Structure Type	Number of Stories						Row Totals
	1 Story	2 - 3 Stories	4 - 7 Stories	8 - 10 Stories	11 - 20 Stories	Over 21 Stories	
C1	410	429	74	-	-	-	913
C2	-	-	17	26	32	6	81
MH	43	-	-	-	-	-	43
PC1	1,020	83	4	3	-	-	1,110
PC2	1	14	18	1	1	-	35
RM	1,530	70	-	-	-	-	1,600
S1	2,579	825	158	14	26	6	3,608
S3	3,341	168	12	1	-	-	3,522
URM	9,083	1,962	91	4	1	-	11,141
W1	194,839	77,013	1	-	-	-	271,853
W2	1,975	10,081	41	-	-	-	12,097
Totals	214,821	90,645	416	49	60	12	306,003
Structure Type	Number of Stories						Row Percent
	1 Story	2 - 3 Stories	4 - 7 Stories	8 - 10 Stories	11 - 20 Stories	Over 21 Stories	
C1	0.1340%	0.1402%	0.0242%	-	-	-	0.30%
C2	-	-	0.0056%	0.0085%	0.0105%	0.0020%	0.03%
MH	0.0141%	-	-	-	-	-	0.01%
PC1	0.3333%	0.0271%	0.0013%	0.0010%	-	-	0.36%
PC2	0.0003%	0.0046%	0.0059%	0.0003%	0.0003%	-	0.01%
RM	0.5000%	0.0229%	-	-	-	-	0.52%
S1	0.8428%	0.2696%	0.0516%	0.0046%	0.0085%	0.0020%	1.18%
S3	1.0918%	0.0549%	0.0039%	0.0003%	-	-	1.15%
URM	2.9683%	0.6412%	0.0297%	0.0013%	0.0003%	-	3.64%
W1	63.6723%	25.1674%	0.0003%	-	-	-	88.84%
W2	0.6454%	3.2944%	0.0134%	-	-	-	3.95%
Percent	70.20%	29.62%	0.14%	0.02%	0.02%	0.00%	100.00%

Table A.10 -- Building counts by Structure type and Year of construction (Decade)

Structure Type	Year of Construction by Decade								Row Totals
	Pre-1939	40-49	50-59	60-69	70-79	80-89	90-99	Post-2000	
C1	74	37	136	472	121	56	16	1	913
C2	18	1	7	19	21	13	2	-	81
MH	4	18	6	4	7	2	2	-	43
PC1	5	18	50	170	283	229	244	111	1,110
PC2	-	-	2	4	2	15	2	10	35
RM	-	-	-	9	251	354	738	248	1,600
S1	46	51	122	516	626	948	703	596	3,608
S3	103	171	239	432	657	746	829	345	3,522
URM	5,462	1,769	3,027	742	141	-	-	-	11,141
W1	24,033	22,193	47,244	34,464	40,328	34,201	41,338	28,052	271,853
W2	199	108	293	2,488	3,021	2,172	1,849	1,967	12,097
Totals	29,944	24,366	51,126	39,320	45,458	38,736	45,723	31,330	306,003
Structure Type	Year of Construction by Decade								Row Percent
	Pre-1939	40-49	50-59	60-69	70-79	80-89	90-99	Post-2000	
C1	0.02%	0.01%	0.04%	0.15%	0.04%	0.02%	0.01%	0.00%	0.30%
C2	0.01%	0.00%	0.00%	0.01%	0.01%	0.00%	0.00%	-	0.03%
MH	0.00%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	-	0.01%
PC1	0.00%	0.01%	0.02%	0.06%	0.09%	0.07%	0.08%	0.04%	0.36%
PC2	-	-	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%
RM	-	-	-	0.00%	0.08%	0.12%	0.24%	0.08%	0.52%
S1	0.02%	0.02%	0.04%	0.17%	0.20%	0.31%	0.23%	0.19%	1.18%
S3	0.03%	0.06%	0.08%	0.14%	0.21%	0.24%	0.27%	0.11%	1.15%
URM	1.78%	0.58%	0.99%	0.24%	0.05%	-	-	-	3.64%
W1	7.85%	7.25%	15.44%	11.26%	13.18%	11.18%	13.51%	9.17%	88.84%
W2	0.07%	0.04%	0.10%	0.81%	0.99%	0.71%	0.60%	0.64%	3.95%
Percent	9.79%	7.96%	16.71%	12.85%	14.86%	12.66%	14.94%	10.24%	100.00%

Table A.11 -- Building counts by Structure type and Basement class

Structure Type	Commercial Basement	Residential Full Basement	Residential Part Basement	Residential Crawl Space	None	Row Total	Row Percent
C1	77	-	-	-	836	913	0.30%
C2	49	-	-	-	32	81	0.03%
MH	-	-	-	-	43	43	0.01%
PC1	11	-	-	-	1,099	1,110	0.36%
PC2	6	-	-	-	29	35	0.01%
RM	8	-	-	-	1,592	1,600	0.52%
S1	98	-	-	-	3,510	3,608	1.18%
S3	10	-	-	-	3,512	3,522	1.15%
URM	233	12	228	5,150	5,518	11,141	3.64%
W1	1	175	2,427	83,078	186,172	271,853	88.84%
W2	52	16	105	562	11,362	12,097	3.95%
Totals	545	203	2,760	88,790	213,705	306,003	100.00%
Percent	0.18%	0.07%	0.90%	29.02%	69.84%	100.00%	

Table A.12 -- Building counts by General occupancy and Basement class

General Occupancy	Commercial Basement	Residential Full Basement	Residential Part Basement	Residential Crawl Space	None	Row Total	Row Percent
Education	4	-	-	-	292	296	0.10%
Food and Entertainment	19	-	-	-	1,303	1,322	0.43%
Health Care	8	-	-	-	14	22	0.01%
Heavy Industrial	19	-	-	-	690	709	0.23%
Light Industrial	24	-	-	-	1,945	1,969	0.64%
Multi-family Residential	90	-	138	3,864	14,551	18,643	6.09%
Office Commercial	164	-	4	13	3,424	3,605	1.18%
Parking Garage	9	-	-	-	41	50	0.02%
Places of Worship	10	-	-	-	1,011	1,021	0.33%
Retail Trade	136	-	-	-	3,877	4,013	1.31%
Single-family Residential	2	198	2,618	84,913	181,733	269,464	88.06%
Wholesale Trade	65	-	-	-	4,824	4,889	1.60%
Totals	550	198	2,760	88,790	213,705	306,003	100.00%
Percent	0.18%	0.06%	0.90%	29.02%	69.84%	100.00%	

Table A.13 -- Building counts and percentages by Structure type and Area class

Structure Type	Square Footage					Row Total
	Less than 2500 sq. ft.	2,500 - 5,000 sq. ft.	5,000 - 10,000 sq. ft.	10,000 - 50,000 sq. ft.	More than 50,000 sq. ft.	
C1	89	130	114	311	269	913
C2	-	-	-	4	77	81
MH	17	1	-	3	22	43
PC1	23	56	75	423	533	1,110
PC2	1	-	-	6	28	35
RM	478	567	327	199	29	1,600
S1	644	551	792	1,118	503	3,608
S3	651	721	932	1,064	154	3,522
URM	7,635	1,814	911	685	96	11,141
W1	218,859	52,994	-	-	-	271,853
W2	-	-	9,058	2,990	49	12,097
Total	228,397	56,834	12,209	6,803	1,760	306,003
Structure Type	Square Footage					Row Percent
	Less than 2500 sq. ft.	2,500 - 5,000 sq. ft.	5,000 - 10,000 sq. ft.	10,000 - 50,000 sq. ft.	More than 50,000 sq. ft.	
C1	0	0.04%	0.04%	0.10%	0.09%	0.30%
C2	-	-	-	0.00%	0.03%	0.03%
MH	0.01%	0.00%	-	0.00%	0.01%	0.01%
PC1	0.01%	0.02%	0.02%	0.14%	0.17%	0.36%
PC2	0.00%	-	-	0.00%	0.01%	0.01%
RM	0.16%	0.19%	0.11%	0.07%	0.01%	0.52%
S1	0.21%	0.18%	0.26%	0.37%	0.16%	1.18%
S3	0.21%	0.24%	0.30%	0.35%	0.05%	1.15%
URM	2.50%	0.59%	0.30%	0.22%	0.03%	3.64%
W1	71.52%	17.32%	-	-	-	88.84%
W2	-	-	2.96%	0.98%	0.02%	3.95%
Percent	74.64%	18.57%	3.99%	2.22%	0.58%	100.00%

Table A.14 -- Building counts and percentages by Structure type and Replacement cost categories

Structure Type	Replacement Cost in Dollars							Row Total
	Less than \$50 K	\$50K to \$100 K	\$100 K to \$300 K	\$300 K to \$500 K	\$500 K to \$1,000 K	\$ 1 Mill to \$ 5 Mill	More than \$ 5 Mill	
C1	3	1	51	67	148	364	279	913
C2	-	-	-	-	-	2	79	81
MH	15	2	1	-	-	10	15	43
PC1	7	3	17	52	121	489	421	1,110
PC2	-	-	1	-	1	17	16	35
RM	26	42	596	327	409	180	20	1,600
S1	18	60	564	365	888	1,262	451	3,608
S3	48	110	718	647	923	992	84	3,522
URM	88	4,403	3,739	1,064	1,078	708	61	11,141
W1	303	85,753	177,059	7,117	1,621	-	-	271,853
W2	-	-	31	2,047	4,246	5,722	51	12,097
Totals	508	90,374	182,777	11,686	9,435	9,746	1,477	306,003
Structure Type	Replacement Cost in Dollars							Row Percent
	Less than \$50 K	\$50K to \$100 K	\$100 K to \$300 K	\$300 K to \$500 K	\$500 K to \$1,000 K	\$ 1 Mill to \$ 5 Mill	More than \$ 5 Mill	
C1	0.0010%	0.0003%	0.0167%	0.0219%	0.0484%	0.1190%	0.0912%	0.30%
C2	-	-	-	-	-	0.0007%	0.0258%	0.03%
MH	0.0049%	0.0007%	0.0003%	-	-	0.0033%	0.0049%	0.01%
PC1	0.0023%	0.0010%	0.0056%	0.0170%	0.0395%	0.1598%	0.1376%	0.36%
PC2	-	-	0.0003%	-	0.0003%	0.0056%	0.0052%	0.01%
RM	0.0085%	0.0137%	0.1948%	0.1069%	0.1337%	0.0588%	0.0065%	0.52%
S1	0.0059%	0.0196%	0.1843%	0.1193%	0.2902%	0.4124%	0.1474%	1.18%
S3	0.0157%	0.0359%	0.2346%	0.2114%	0.3016%	0.3242%	0.0275%	1.15%
URM	0.0288%	1.4389%	1.2219%	0.3477%	0.3523%	0.2314%	0.0199%	3.64%
W1	0.0990%	28.0236%	57.8619%	2.3258%	0.5297%	-	-	88.84%
W2	-	-	0.0101%	0.6689%	1.3876%	1.8699%	0.0167%	3.95%
Percent	0.17%	29.53%	59.73%	3.82%	3.08%	3.18%	0.48%	100.00%

Table A.15 -- Building Replacement costs (in millions of dollars) by Structure type and General occupancy

Structure Type	Schools Colleges	Food Entertainment	Health Care	Retail Trade	Wholesale Trade	Office Commercial
C1	1,786.50	17.09	79.07	256.07	179.53	1,128.86
C2	-	-	28.93	-	10.05	750.31
MH	-	-	-	-	-	-
PC1	-	9.63	-	366.53	5,950.98	158.77
PC2	-	-	-	-	-	-
RM	-	43.99	-	140.57	503.87	17.83
S1	2,681.83	108.66	585.81	2,235.74	681.18	3,962.59
S3	2.65	52.88	-	182.68	2,243.70	203.14
URM	6.64	117.85	14.45	809.61	1,051.40	311.10
W1	1.45	231.61	-	83.35	18.17	392.27
W2	6.49	171.28	18.95	119.60	122.46	465.55
Totals	4,485.55	752.99	727.20	4,194.16	10,761.34	7,390.42
Percent	5.14%	0.86%	0.83%	4.81%	12.33%	8.47%
Structure Type	Heavy Industrial	Light Industrial	Places of Worship	Parking Garage	Multi-family Residential	Single-family Residential
C1	345.92	4.58	3.47	9.66	82.34	-
C2	-	-	-	-	640.32	-
MH	-	-	-	-	174.30	-
PC1	570.69	346.73	-	-	14.01	-
PC2	-	0.17	-	214.59	-	-
RM	-	313.78	1.09	-	0.30	-
S1	828.92	309.70	122.61	77.97	1,818.56	-
S3	812.30	407.60	16.78	-	27.18	-
URM	85.19	283.87	29.17	-	580.39	666.44
W1	0.48	4.30	31.72	-	2,186.28	35,828.95
W2	-	4.34	903.60	-	9,515.73	1,703.61
Totals	2,643.49	1,675.07	1,108.44	302.23	15,039.39	38,199.00
Percent	3.03%	1.92%	1.27%	0.35%	17.23%	43.77%

Table A.16 -- Building counts by Structure type and Content Value category

Structure Type	Content Value in Dollars							Row Total	Row Percent
	Less than \$50 K	\$50K to \$100 K	\$100 K to \$300 K	\$300 K to \$500 K	\$500 K to \$1,000 K	\$ 1 Mill to \$ 5 Mill	More than \$ 5 Mill		
C1	3	1	45	83	123	368	290	913	0.30%
C2	-	-	-	-	-	5	76	81	0.03%
MH	17	1	-	-	3	17	5	43	0.01%
PC1	4	5	31	33	114	478	445	1,110	0.36%
PC2	-	-	1	1	3	23	7	35	0.01%
RM	26	43	595	323	408	185	20	1,600	0.52%
S1	12	61	597	456	898	1,184	400	3,608	1.18%
S3	44	108	711	642	916	972	129	3,522	1.15%
URM	4,297	2,382	1,894	971	900	635	62	11,141	3.64%
W1	85,934	149,961	34,201	1,316	433	8	-	271,853	88.84%
W2	-	-	2,934	2,551	5,268	1,320	24	12,097	3.95%
Totals	90,337	152,562	41,009	6,376	9,066	5,195	1,458	306,003	100.00%
Percent	29.52%	49.86%	13.40%	2.08%	2.96%	1.70%	0.48%	100.00%	

Table A.17 -- Building counts by Structure type and Essential Facility designation

Structure Type	Essential Facility Type									Row Total	Row Percent
	EFFS	EFHL	EFHM	EFHS	EFMC	EFPS	EFS1	EFS2	FALSE		
C1	5	3	1	3	52	18	256	12	563	913	0.30%
C2	-	1	-	-	16	-	-	-	64	81	0.03%
MH	-	-	-	-	-	-	-	-	43	43	0.01%
PC1	-	-	-	-	6	-	-	-	1,104	1,110	0.36%
PC2	-	-	-	-	-	-	-	-	35	35	0.01%
RM	-	-	-	-	2	8	-	-	1,590	1,600	0.52%
S1	2	9	2	-	108	1	9	4	3,473	3,608	1.18%
S3	-	-	-	-	1	-	2	-	3,519	3,522	1.15%
URM	5	1	-	-	25	7	3	-	11,100	11,141	3.64%
W1	1	-	-	-	87	-	3	-	271,762	271,853	88.84%
W2	-	-	2	-	61	1	6	-	12,027	12,097	3.95%
Totals	13	14	5	3	358	35	279	16	305,280	306,003	100.00%
Percent	0.00%	0.00%	0.00%	0.00%	0.12%	0.01%	0.09%	0.01%	99.76%	100.00%	

EFFS	Fire Stations
EFHL	Low-Rise Healthcare Facilities
EFHM	Mid-Rise Healthcare Facilities
EFHS	High-Rise Healthcare Facilities

EFMC	Medical Clinics, Labs, Offices
EFPS	Police Stations
EFS1	Schools
EFS2	Colleges and Universities

Table A.18 -- Building counts by General occupancy and Number of Dwellings in structure

General Occupancy	Number of Dwelling Units in Building								Row Total	Row Percent
	0	1	2	3 - 4	5 - 9	10 - 19	20 - 49	Over 50		
Education	291	5	-	-	-	-	-	-	296	0.10%
Food and Entertainment	1,314	8	-	-	-	-	-	-	1,322	0.43%
Health Care	22	-	-	-	-	-	-	-	22	0.01%
Heavy Industrial	709	-	-	-	-	-	-	-	709	0.23%
Light Industrial	1,951	17	-	1	-	-	-	-	1,969	0.64%
Multi-family Residential	4	70	7,597	2,210	4,311	2,941	924	121	18,178	5.95%
Office Commercial	3,469	135	-	1	-	-	-	-	3,605	1.18%
Parking Garage	50	-	-	-	-	-	-	-	50	0.02%
Places of Worship	988	30	1	2	-	-	-	-	1,021	0.33%
Retail Trade	3,946	42	12	9	3	1	-	-	4,013	1.31%
Single-family Residential	222	269,229	-	-	1	-	-	-	269,452	88.19%
Wholesale Trade	4,866	23	-	-	-	-	-	-	4,889	1.60%
Totals	17,832	269,559	7,610	2,223	4,315	2,942	924	121	305,526	100.00%
Percent	5.84%	88.23%	2.49%	0.73%	1.41%	0.96%	0.30%	0.04%	100.00%	

n.b. Temporary Lodging, Institutional Dormitories and Nursing Homes have not been included in this tabulation

Table A.19 -- Building counts by Structure type and Number of Dwellings in structure

Structure Type	Number of Dwelling Units in Building								Row Total	Row Percent
	0	1	2	3 - 4	5 - 9	10 - 19	20 - 49	Over 50		
C1	876	2	-	4	18	4	3	2	909	0.30%
C2	47	-	-	-	-	-	-	25	72	0.02%
MH	-	15	2	1	-	-	3	22	43	0.01%
PC1	1,094	2	-	13	-	-	-	-	1,109	0.36%
PC2	35	-	-	-	-	-	-	-	35	0.01%
RM	1,590	8	-	1	-	-	-	-	1,599	0.52%
S1	3,076	10	5	35	162	84	34	34	3,440	1.13%
S3	3,483	5	-	-	21	5	-	-	3,514	1.15%
URM	3,642	6,412	461	334	164	70	16	5	11,104	3.63%
W1	2,430	259,727	7,140	1,525	925	16	-	-	271,763	88.95%
W2	1,559	3,378	2	310	3,025	2,763	868	33	11,938	3.91%
Totals	17,832	269,559	7,610	2,223	4,315	2,942	924	121	305,526	100.00%
Percent	5.84%	88.23%	2.49%	0.73%	1.41%	0.96%	0.30%	0.04%	100.00%	

n.b. Temporary Lodging, Institutional Dormitories and Nursing Homes have not been included in this tabulation

APPENDIX B . Influence of Input Variables on Structure Type Outcome Pairs in the Multinomial Logistic Regression Model

**Table B.1 -- Influence of Height (Number of stories) on Factor change in Structure
type Odds**

Input Variable	Odds Comparing Alternative 1 to 2	Raw Coefficient	Odds Factor Change	
			exp(b)	exp(b*SD(x))
STORIES (sd=2.1402625)	Concrete-Steel	-0.15843	0.8535	0.7124
	Concrete-URM	0.24792	1.2814	1.7000
	Concrete-Wood	0.25727	1.2934	1.7343
	Steel-Concrete	0.15843	1.1717	1.4036
	Steel-URM	0.40635	1.5013	2.3862
	Steel-Wood	0.41570	1.5154	2.4344
	URM-Concrete	-0.24792	0.7804	0.5882
	URM-Steel	-0.40635	0.6661	0.4191
	URM-Wood	0.00935	1.0094	1.0202
	Wood-Concrete	-0.25727	0.7732	0.5766
	Wood-Steel	-0.41570	0.6599	0.4108
	Wood-URM	-0.00935	0.9907	0.9802

	significant at 95% confidence
	significant at 99% confidence

Table B.2 -- Influence of Area (Square Feet) on Factor change in Structure type

Odds

Input Variable	Odds Comparing Alternative 1 to 2	Raw Coefficient	Odds Factor Change	
			exp(b)	exp(b*SD(x))
SQUARE FEET (sd=45584.275)	Concrete-Steel	0.00000	1.0000	0.9527
	Concrete-URM	0.00006	1.0001	14.3518
	Concrete-Wood	0.00003	1.0000	3.3665
	Steel-Concrete	0.00000	1.0000	1.0496
	Steel-URM	0.00006	1.0001	15.0636
	Steel-Wood	0.00003	1.0000	3.5335
	URM-Concrete	-0.00006	0.9999	0.0697
	URM-Steel	-0.00006	0.9999	0.0664
	URM-Wood	-0.00003	1.0000	0.2346
	Wood-Concrete	-0.00003	1.0000	0.2970
	Wood-Steel	-0.00003	1.0000	0.2830
	Wood-URM	0.00003	1.0000	4.2631

	significant at 95% confidence
	significant at 99% confidence

Table B.3 -- Influence of Year of Construction on Factor change in Structure type

Odds

Input Variable	Odds Comparing Alternative 1 to 2	Raw Coefficient	Odds Factor Change	
			exp(b)	exp(b*SD(x))
YEAR BUILT (sd=22.9592)	Concrete-Steel	-0.06116	0.9407	0.2456
	Concrete-URM	0.07542	1.0783	5.6493
	Concrete-Wood	-0.01107	0.9890	0.7755
	Steel-Concrete	0.06116	1.0631	4.0720
	Steel-URM	0.13658	1.1463	23.0041
	Steel-Wood	0.05008	1.0514	3.1578
	URM-Concrete	-0.07542	0.9274	0.1770
	URM-Steel	-0.13658	0.8723	0.0435
	URM-Wood	-0.08649	0.9171	0.1373
	Wood-Concrete	0.01107	1.0111	1.2895
	Wood-Steel	-0.05008	0.9512	0.3167
	Wood-URM	0.08649	1.0903	7.2848

significant at 95% confidence
 significant at 99% confidence

Table B.4 -- Influence of Wholesale Trade, Commercial Office and Bank occupancies on Factor change in Structure type Odds

Input Variable	Odds Comparing Alternative 1 to 2	Raw Coefficient	Odds Factor Change	
			exp(b)	exp(b*SD(x))
COM2 (Wholesale Trade)	Concrete-Steel	1.17328	3.2326	1.5675
	Concrete-URM	1.51681	4.5577	1.7880
	Concrete-Wood	1.41450	4.1144	1.7193
	Steel-Concrete	-1.17328	0.3093	0.6379
	Steel-URM	0.34353	1.4099	1.1407
	Steel-Wood	0.24122	1.2728	1.0968
	URM-Concrete	-1.51681	0.2194	0.5593
	URM-Steel	-0.34353	0.7093	0.8767
	URM-Wood	-0.10231	0.9028	0.9616
	Wood-Concrete	-1.41450	0.2430	0.5816
	Wood-Steel	-0.24122	0.7857	0.9117
	Wood-URM	0.10231	1.1077	1.0400
	COM4 (Commercial Office)	Concrete-Steel	3.80386	44.8742
Concrete-URM		3.41236	30.3368	3.1836
Concrete-Wood		2.57086	13.0770	2.3927
Steel-Concrete		-3.80386	0.0223	0.2750
Steel-URM		-0.39150	0.6760	0.8756
Steel-Wood		-1.23301	0.2914	0.6581
URM-Concrete		-3.41236	0.0330	0.3141
URM-Steel		0.39150	1.4792	1.1421
URM-Wood		-0.84151	0.4311	0.7516
Wood-Concrete		-2.57086	0.0765	0.4179
Wood-Steel		1.23301	3.4315	1.5196
Wood-URM		0.84151	2.3199	1.3305
COM5 (Banks)		Concrete-Steel	2.86616	17.5694
	Concrete-URM	2.64597	14.0971	1.4862
	Concrete-Wood	1.61937	5.0499	1.2744
	Steel-Concrete	-2.86616	0.0569	0.6510
	Steel-URM	-0.22019	0.8024	0.9676
	Steel-Wood	-1.24679	0.2874	0.8297
	URM-Concrete	-2.64597	0.0709	0.6729
	URM-Steel	0.22019	1.2463	1.0335
	URM-Wood	-1.02660	0.3582	0.8575
	Wood-Concrete	-1.61937	0.1980	0.7847
	Wood-Steel	1.24679	3.4791	1.2053
	Wood-URM	1.02660	2.7915	1.1662

n.b. reference level is COM1 (Retail Trade)

significant at 95% confidence
 significant at 99% confidence

Table B.5 -- Influence of Restaurant, Heavy Industrial and Multi-family residential occupancies on Factor change in Structure type Odds

Input Variable	Odds Comparing Alternative 1 to 2	Raw Coefficient	Odds Factor Change	
			exp(b)	exp(b*SD(x))
COM8 (Restaurants)	Concrete-Steel	1.38134	3.9802	1.3607
	Concrete-URM	1.67898	5.3601	1.4540
	Concrete-Wood	0.31642	1.3722	1.0731
	Steel-Concrete	-1.38134	0.2512	0.7349
	Steel-URM	0.29764	1.3467	1.0686
	Steel-Wood	-1.06492	0.3448	0.7887
	URM-Concrete	-1.67898	0.1866	0.6877
	URM-Steel	-0.29764	0.7426	0.9358
	URM-Wood	-1.36256	0.2560	0.7380
	Wood-Concrete	-0.31642	0.7288	0.9319
	Wood-Steel	1.06492	2.9006	1.2680
	Wood-URM	1.36256	3.9062	1.3550
IND1 (Heavy Industrial)	Concrete-Steel	0.67894	1.9718	1.1121
	Concrete-URM	1.17584	3.2409	1.2021
	Concrete-Wood	-1.23244	0.2916	0.8245
	Steel-Concrete	-0.67894	0.5072	0.8992
	Steel-URM	0.49689	1.6436	1.0809
	Steel-Wood	-1.91138	0.1479	0.7414
	URM-Concrete	-1.17584	0.3086	0.8319
	URM-Steel	-0.49689	0.6084	0.9252
	URM-Wood	-2.40828	0.0900	0.6859
	Wood-Concrete	1.23244	3.4296	1.2128
	Wood-Steel	1.91138	6.7624	1.3488
	Wood-URM	2.40828	11.1148	1.4579
RES3 (Multi-family Residential)	Concrete-Steel	1.65841	5.2510	2.1877
	Concrete-URM	1.12487	3.0798	1.7006
	Concrete-Wood	-2.84780	0.0580	0.2607
	Steel-Concrete	-1.65841	0.1904	0.4571
	Steel-URM	-0.53354	0.5865	0.7774
	Steel-Wood	-4.50621	0.0110	0.1192
	URM-Concrete	-1.12487	0.3247	0.5880
	URM-Steel	0.53354	1.7050	1.2864
	URM-Wood	-3.97266	0.0188	0.1533
	Wood-Concrete	2.84780	17.2497	3.8355
	Wood-Steel	4.50621	90.5774	8.3908
	Wood-URM	3.97266	53.1259	6.5226

n.b. reference level is COM1 (Retail Trade)

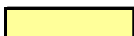

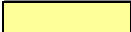
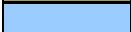
 significant at 95% confidence
 significant at 99% confidence

Table B.6 -- Influence of Fire Rating descriptor on Factor change in Structure type

Odds

Input Variable	Odds Comparing Alternative 1 to 2	Raw Coefficient	Odds Factor Change	
			exp(b)	exp(b*SD(x))
Fire Resistant Fire Rating	Concrete-Steel	-6.23672	0.0020	0.0466
	Concrete-URM	-6.18536	0.0021	0.0478
	Concrete-Wood	-0.72320	0.4852	0.7008
	Steel-Concrete	6.23672	511.1811	21.4543
	Steel-URM	0.05136	1.0527	1.0256
	Steel-Wood	5.51353	248.0240	15.0354
	URM-Concrete	6.18536	485.5891	20.9194
	URM-Steel	-0.05136	0.9499	0.9751
	URM-Wood	5.46216	235.6068	14.6605
	Wood-Concrete	0.72320	2.0610	1.4269
	Wood-Steel	-5.51353	0.0040	0.0665
	Wood-URM	-5.46216	0.0042	0.0682
Wood Joist Fire Rating	Concrete-Steel	-6.95986	0.0009	0.0308
	Concrete-URM	-8.61423	0.0002	0.0135
	Concrete-Wood	-7.49728	0.0006	0.0235
	Steel-Concrete	6.95986	53.4831	32.4699
	Steel-URM	-1.65437	0.1912	0.4372
	Steel-Wood	-0.53742	0.5843	0.7643
	URM-Concrete	8.61423	509.5132	74.2615
	URM-Steel	1.65437	5.2298	2.2871
	URM-Wood	1.11695	3.0555	1.7481
	Wood-Concrete	7.49728	803.1279	42.4809
	Wood-Steel	0.53742	1.7116	1.3083
	Wood-URM	-1.11695	0.3273	0.5720

n.b. reference level is Fire Proof Fire Rating

 significant at 95% confidence
 significant at 99% confidence

References

- Abrams, Daniel P., Amr S. Elnashai and J. E. Beavers. (2002). "A New Engineering Paradigm: Consequence-Based Engineering." Mid-America Earthquake Center. Retrieved October 22, 2007, from <http://mae.ce.uiuc.edu/documents/cbepaper.pdf>.
- Acharya, Tinku and Ajoy K. Ray (2005). *Image Processing: Principles and Applications*. Hoboken, N. J.: John Wiley & Sons, Inc.
- ACSE (2005). Minimum Design Loads for Buildings and Other Structures, SEI/ACSE 7-05. Reston, VA: American Society of Civil Engineers.
- Adamek, Tomasz and Noel O'Connor (2003). Efficient contour-based shape representation and matching. Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval, Berkeley, California.
- Akaike, Hirotugu (1974). "A new look at the statistical model identification." IEEE Transactions on Automatic Control **19**(6): 716-723.
- Aldrich, John H. and Forrest D. Nelson (1984). Linear Probability, Logit, and Probit Models. Series / Number 07-045 - Quantitative Applications in the Social Sciences. Newbury Park, CA: Sage publications.
- American Community Survey Office. (2006). "American FactFinder -- B25024. UNITS IN STRUCTURE - Universe: HOUSING UNITS." 2006 American Community Survey. US Census Bureau. Retrieved May 31, 2008, from http://factfinder.census.gov/servlet/DTTable?_bm=y&-geo_id=05000US47157&-ds_name=ACS_2006_EST_G00_&-mt_name=ACS_2006_EST_G2000_B25024.
- American Community Survey Office. (2007, August 24). "2003 ACS Narrative Profile for Memphis City, Shelby County." Population and Housing Profile. US Census Bureau. Retrieved May 31, 2008, from <http://www.census.gov/acs/www/Products/Profiles/Single/2003/ACS/Narrative/155/NP15500US4748000157.htm>.
- Anderson, Dave and George McNeil. (1992). "Artificial Neural Networks Technology." Data & Analysis Center for Software. Retrieved October 22, 2007, from https://www.dacs.dtic.mil/techs/neural/neural_ToC.php.
- Anderson, J. A. (1984a). "Regression and ordered categorical variables." Journal of the Royal Statistical Society, Series B **46**: 1-30.

- Anderson, James A. and Edward Rosenfeld (1990). *Neurocomputing: Foundations of Research*. Cambridge, MA: MIT Press.
- Anderson, Theodore. W. (1984b). *An Introduction to Multivariate Statistical Analysis*. Second Edition. New York: Wiley.
- Ansari, Nirwan and Edward J. Delp (1990). "Partial shape recognition: a landmark-based approach." IEEE Transactions on Pattern Analysis and Machine Intelligence **12**(5): 470-483.
- Antenucci, John C., Kay Brown, Peter L. Croswell et al. (1991). *Geographic Information Systems: A Guide to the Technology*. New York: Van Nostrand Reinhold.
- Arcelli, Carlo and Gabriella Sanniti di Baja (1985). "A width-independent fast thinning algorithm." IEEE Transactions on Pattern Analysis and Machine Intelligence **7**(4): 463--474.
- Arkin, Esther M., L. Paul Chew, Daniel P. Huttenlocher et al. (1991). "An efficiently computable metric for comparing polygonal shapes." IEEE Transactions on Pattern Analysis and Machine Intelligence **13**(3): 209 - 216.
- Arnold, Christopher and Robert Reitherman (1982). *Building Configuration and Seismic Design*. New York: John Wiley & Sons.
- Ashbrook, A. and N. A. Thacker. (1998, January 12). "Tutorial: Algorithms for 2-Dimensional object recognition." TINA Memos: Human and Machine Vision. Retrieved July 26, 2008, from http://www.tina-vision.net/docs/memos_vision.php.
- Association for the Advancement of Artificial Intelligence. (2007, March). "Genetic Algorithm & Genetic Programming: (A subtopic of Machine Learning)." Retrieved September 17, 2007, from <http://www.aaai.org/AITopics/html/genalg.html>.
- ATC-13 (1985). Earthquake Damage Evaluation Data for California - Report ATC-13. Redwood City, CA: Applied Technology Council.
- ATC-21 (1988). Rapid Visual Screening of Buildings for Potential Seismic Hazards - Report ATC-21: FEMA 154. Redwood City, CA: Applied Technology Council. .
- ATC-21 (1991). Seismic Vulnerability and Impact of Disruption on Lifelines in the Conterminous United States. - Report ATC-21. Redwood City, CA: Federal Emergency Management Association, Applied Technology Council. .
- ATC-69 (2008). Reducing the Risks of Nonstructural Earthquake Damage: State-of-the-Art and Practice Report - ATC-69. Redwood City, CA: Federal Emergency Management Association, Applied Technology Council. .

- Attneave, F. (1954). "Some informational aspects of visual perception." Psychological Review **61**(3): 183-193.
- August, Jonas, Steven W. Zucker and Allen Tannenbaum (1999). On the evolution of the skeleton. Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece.
- Baelia, B., M. Pla and Dan Lee (1995). Digital map generalization in practice. Proceedings of ICA Workshop on Progress in Automated Map Generalization, Barcelona, Spain.
- Barile, Margherita. (2008, February 10). "Taxicab Metric." A Wolfram Web Resource. E. W. Weisstein, Ed. MathWorld. Retrieved February 12, 2008, from <http://mathworld.wolfram.com/TaxicabMetric.html>.
- Basu, Mitra, Horst Bunke and Alberto Del Bimbo (2005). "Guest editors' introduction to the special section on syntactic and structural pattern recognition." IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(7): 1009 - 1012.
- Bea, Keith (1998). FEMA and Disaster Relief. Washington, D.C.: Congressional Research Service, The Library of Congress. **97-159 GOV**: 37.
- Beatley, Timothy and Philip R. Berke (1992). "Time to shake up earthquake planning." Issues in Science & Technology **9**(2): 82-89.
- Bellone, Tamara, Enrico Borgogno and Giuliano Comoglio (2004). Improving automated generalization for on demand web mapping by multi-scale databases. O. Altan, Ed. Proceedings of the XX ISPRS Congress, Commission III: Geo-Imagery Bridging Continents, Istanbul, Turkey: ISPRS.
- Belongie, Serge, Jitendra Malik and Jan Puzicha (2002). "Shape matching and object recognition using shape contexts." IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(4): 509-522.
- Bennett, J. R. and J. S. MacDonald (1975). "On the measurement of curvature in a quantized environment." IEEE Transactions on Computers **C-24**(8): 803 - 820.
- Berke, Philip R. (1995a). "Natural hazard reduction and sustainable development: A global assessment." Journal of Planning Literature **9**(4): 370-382.
- Berke, Philip R. (1995b). Reducing Natural Hazard Risks through Land Use Planning and Growth Management: Federal and State Policy Experience. College Station, TX: Hazard Reduction & Recovery Center, Texas A & M University.
- Berke, Philip R. (1998). "Reducing natural hazard risks through state growth management." Journal of the American Planning Association **64**(1): 76 - 87.

- Bicego, Manuele, Vittorio Murino, Marcello Pelillo et al. (2006). "Editorial: Similarity-based pattern recognition." Pattern Recognition **39**(10): 1813 – 1814.
- Bischof, H., W. Schneider and A. J. Pinz (1992). "Multispectral classification of Landsat-images using neural networks." IEEE Transactions on Geoscience and Remote Sensing **30**(3): 482-490.
- Bishop, Christopher M. (1995). *Neural Networks for Pattern Recognition*. New York: Oxford University Press, Inc.
- Black, Paul E. (2004a, December 17). "Euclidean distance." Dictionary of Algorithms and Data Structures [online]. P. E. Black, Ed. U.S. National Institute of Standards and Technology. Retrieved February 12, 2008, from <http://www.nist.gov/dads/HTML/euclidndstnc.html>.
- Black, Paul E. (2004b, December 17). "Levenshtein distance." Dictionary of Algorithms and Data Structures [online]. P. E. Black, Ed. U.S. National Institute of Standards and Technology. Retrieved February 12, 2008, from <http://www.nist.gov/dads/HTML/Levenshtein.html>.
- Black, Paul E. (2004c, December 17). "Lm distance." Dictionary of Algorithms and Data Structures [online]. P. E. Black, Ed. U.S. National Institute of Standards and Technology. Retrieved February 12, 2008, from <http://www.nist.gov/dads/HTML/lmdistance.html>.
- Black, Paul E. (2004d, December 17). "Manhattan distance." Dictionary of Algorithms and Data Structures [online]. P. E. Black, Ed. U.S. National Institute of Standards and Technology. Retrieved February 12, 2008, from <http://www.nist.gov/dads/HTML/manhattanDistance.html>.
- Black, Paul E. (2004e, December 17). "Rectilinear Distance." Dictionary of Algorithms and Data Structures [online]. P. E. Black, Ed. U.S. National Institute of Standards and Technology. Retrieved February 12, 2008, from <http://www.nist.gov/dads/HTML/rectilinear.html>.
- Blum, H. (1967). A transformation for extracting new descriptors of form. In *Models for the Perception of Speech and Visual Form*. W. Whalen-Dunn, Ed. Cambridge, MA: MIT Press: 362- 380.
- Blumenkrans, Alejandro (1991). "Two-dimensional object recognition using a two-dimensional polar transform." Pattern Recognition **24**(9): 879-890.
- Bookstein, Fred L. (1991). *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge: Cambridge University Press.

- Boxer, Laurence, Chun-Shi Chang, Russ Miller et al. (1993). "Polygonal approximation by boundary reduction." Pattern Recognition Letters **14**(2): 111-119.
- Bribiesca, Ernesto (1981). "Arithmetic operations among shapes using shape numbers." Pattern Recognition **12**(2): 123-137.
- Bribiesca, Ernesto and Adolfo Guzman (1980). "How to describe pure form and how to measure differences in shapes using shape numbers." Pattern Recognition **12**(2): 101-112.
- Briechle, Kendra J. (1999). Natural Hazard Mitigation and Local Government Decision Making. In *The Municipal Year Book 1999*. Washington, D. C. : International City/County Management Association: 3-9.
- British Broadcasting Corporation. (2008, May 27). "BBC News: Special Reports, 2008 - China Quake." British Broadcasting Corporation, International Edition. Retrieved May 28, 2008, from http://news.bbc.co.uk/2/hi/in_depth/asia_pacific/2008/china_quake/default.stm.
- Brody, Samuel D. (2003). "Are we learning to make better plans? A longitudinal analysis of plan quality associated with natural hazards." Journal of Planning Education and Research **23**(2): 191-201.
- Bunke, Horst and Alberto Sanfeliu, Eds. (1990). *Syntactic and Structural Pattern Recognition: Theory and Applications*. World Scientific Series in Computer Science. Singapore: World Scientific.
- Burby, Raymond J. (1994). "Floodplain planning and management: Research needed for the 21st Century." Journal of Contemporary Water Research and Education **97**: 44-47.
- Burby, Raymond J., Ed. (1998). *Cooperating with Nature: Confronting Natural Hazards with Land Use Planning for Sustainable Communities*. Washington, D. C.: Joseph Henry/The National Academies Press.
- Burby, Raymond J. (2005). "Have state comprehensive planning mandates reduced insured losses from natural disasters?" Natural Hazards Review **6**: 67-81.
- Burby, Raymond J. (2006). "Hurricane Katrina and the Paradoxes of Government Disaster Policy: Bringing about wise governmental decisions for hazardous areas." The ANNALS of the American Academy of Political and Social Science **604**(1): 171-191.
- Burby, Raymond J., Timothy Beatley, Phillip R. Berke et al. (1999). "Unleashing the power of planning to create disaster-resistant communities." Journal of the American Planning Association **65**(3): 247-258.

- Burby, Raymond J., Steven P. French and Arthur C. Nelson (1998). "Plans, code enforcement, and damage reduction: Evidence from the Northridge earthquake." Earthquake Spectra **14**(1): 59-74.
- Burby, Raymond J. and Peter J. May (1998). "Intergovernmental environmental planning: Addressing the commitment conundrum." Journal of Environmental Planning and Management **41**(1): 95 - 110.
- Cable News Network. (2005). "Hurricane Katrina - Special Reports from CNN.com." Cable News Network, International Edition. Retrieved May 28, 2008, from <http://www.cnn.com/SPECIALS/2005/katrina/>.
- California Scientific. (2007). "BrainMaker Neural Network Application Examples." Retrieved January 19, 2007, from <http://www.calsci.com/Applications.html>.
- Campbell, M. Karen and Allan Donner (1989). "Classification efficiency of multinomial logistic regression relative to ordinal logistic regression." Journal of the American Statistical Association **84**(6): 587-592.
- Carling, A. (1992). *Introducing Neural Networks*. Wilmslow, UK: Sigma Press.
- Cecconi, Alessandro, Robert Weibel and Mathieu Barrault (2002). Improving automated generalization for on demand web mapping by multi-scale databases. C. Armenakis and Y. C. Lee, Eds. Proceedings of the Joint ISPRS Commission IV symposium, Spatial Data Handling and 95th Annual CIG Geomatics Conference : Geospatial Theory, Processing and Applications, Ottawa, Canada: Canadian Institute of Geomatics.
- Cesar, Roberto Marcondes and Luciano Da Fontoura Costa (1995). "Piecewise linear segmentation of digital contours in $O(N \cdot \log(N))$ through a technique based on effective digital curvature estimation." Real-Time Imaging **1**: 409-417.
- Cesar, Roberto Marcondes and Luciano Da Fontoura Costa (1996). "Towards effective planar shape representation with multiscale digital curvature analysis based on signal processing techniques." Pattern Recognition **29**(9): 1559-1569.
- Chaikin, George M. (1974). "Short note: An algorithm for high-speed curve generation." Computer Graphics and Image Processing **3**: 346-349.
- Chang, C. C., S. M. Hwang and D. J. Buehrer (1991). "A shape recognition scheme based on relative distances of feature points from the centroid." Pattern Recognition Letters **24**(11): 1053-1063.
- Chang, S. E. (1998). Direct Economic Impact. In *Engineering and Socioeconomic Impacts of Earthquakes: An Analysis of Electricity Lifeline Disruptions in the New Madrid Area*. M. Shinozuka, A. Rose and R. Eguchi, Eds. Buffalo, NY: MCEER.

- Chang, S. E. (2001). "Structural change in urban economies: Recovery and long-term impacts in the 1995 Kobe earthquake." Journal of Economics and Business Administration **183**: 47-66.
- Chaudhuri, D. and A. Samal (2007). "A simple method for fitting of bounding rectangle to closed regions." Pattern Recognition **40**(7): 1981-1989.
- Chen, S. W., S. T. Tung, C. Y. Fang et al. (1998). "Extended attributed string matching for shape recognition." Computer Vision and Image Understanding **70**(1): 36-50.
- City-Data.com. (2008). "Shelby County, Tennessee detailed profile - houses, real estate, agriculture, wages, work, ancestries, and more." City-Data.com. Retrieved May31, 2008, from http://www.city-data.com/county/Shelby_County-TN.html.
- Comaniciu, Dorin and Peter Meer (2002). "Mean shift: a robust approach toward feature space analysis." IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(5): 603-619.
- Congressional Hazards Caucus. (2007a, July 23). "Congressional Hazards Caucus Fact Sheet: Floods." Congressional Hazards Caucus Fact Sheets. American Geographic Institute. Retrieved March 21, 2008, from <http://www.hazardscaucus.org/floods-factsheet0207.pdf>.
- Congressional Hazards Caucus. (2007b, July 23). "Congressional Hazards Caucus Fact Sheet: Hurricanes." Congressional Hazards Caucus Fact Sheets. American Geographic Institute. Retrieved March 21, 2008, from http://www.hazardscaucus.org/hurricanes_factsheet0306.pdf.
- Costa, Luciano da Fontoura and Roberto Marcondes Junior Cesar (2001a). *Shape Analysis and Classification: Theory and Practice*. New York: CRC Press.
- Costa, Luciano da Fontoura and Roberto Marcondes Junior Cesar. (2001b). "Shape Analysis by Costa & Cesar: Basic Mathematical Concepts." L. d. F. Costa and R. M. C. Junior, Eds. CRC Press. Retrieved March 11, 2008, from http://www.ime.usp.br/~cesar/shape_crc/chap1.html.
- Cover, T. M. (1965). "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition." IEEE Transactions on Electronic Computers **14**: 326-334.
- Cutter, Susan L., Ed. (2005). *American Hazardscapes: The Regionalization of Hazards and Disasters*. Washington, D. C.: Joseph Henry/The National Academies Press.
- De la Rosa, D., F. Mayol, E. Diaz-Pereira et al. (2004). "A land evaluation decision support system (MicroLEIS DSS) for agricultural soil protection: With special

reference to the Mediterranean region." Environmental Modelling & Software **19**(10): 929-942.

Demers, Michael N. (1999). *Fundamentals of Geographic Information Systems*. 3rd Edition. New York: John Wiley & Sons, Inc.

Devijver, Pierre A. and Josef Kittler, Eds. (1982). *Pattern Recognition: A Statistical Approach*. Englewood Cliffs, N.J.: Prentice-Hall.

Douglas, D. H. and T. K. Peucker (1973). "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature." Canadian Cartographer **10**(2): 112-122.

Dreiseitl, Stephan and Lucila Ohno-Machado (2002). "Logistic regression and artificial neural network classification models: a methodology review." Journal of Biomedical Informatics **35**: 352-359.

Dryden, Ian L. and Kanti V. Mardia (1993). "Multivariate Shape Analysis." Sankhya: The Indian Journal of Statistics **Special Volume 55**(Series A, Part 3): 460-480.

Dryden, Ian L. and Kanti V. Mardia (1998). *Statistical Shape Analysis*. New York: John Wiley & Sons.

Duda, Richard O. and Peter E. Hart (1973). *Pattern Classification and Scene Analysis*. New York: Wiley.

Duda, Richard O., Peter E. Hart and David G. Stork (2001). *Pattern Classification*. Second. New York: John Wiley & Sons, Inc.

Dudani, S., K. Breeding and R. McGhee (1977). "Aircraft identification by moment invariants." IEEE Transactions on Computers **26**(1): 39-46.

Dwinnell, Will. (2006, November 17). "Mahalanobis Distance." Data Mining in MATLAB. W. Dwinnell, Ed. Retrieved February 12, 2008, from <http://matlabdatamining.blogspot.com/2006/11/mahalanobis-distance.html>.

Ellingwood, Bruce R. (2007). Quantifying and communicating uncertainty in seismic risk assessment. Proceedings of the Special Workshop on Risk Acceptance and Risk Communication, Stanford University, Palo Alto.

Environmental Systems Research Institute, Inc. (2007, March 15). "ArcGIS Desktop Help 9.2." ESRI: Redlands, CA. Retrieved February 13, 2008, from <http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=welcome>.

- ESRI (1996). Automation of Map Generalization: the Cutting-Edge Technology. ESRI White Paper Series. Redlands, CA: Environmental Systems Research Institute, Inc.
- Fabel, George. (1997, October). "Machine vision systems looking better all the time." Quality Digest Online Magazine. Retrieved March 13, 2008, from <http://www.qualitydigest.com/oct97/html/machvis.html>.
- Fahlman, Scott E. (1989). Fast learning variations of back-propagation: An empirical study. In *Proceedings of the 1988 Connectionist Models Summer School*. D. S. Touretzky, G. Hinton and T. Sejnowski, Eds. San Mateo: Morgan Kaufmann.
- FEMA - DHS (2007). HAZUS MH-MR3 Technical Manuals. Multi-hazard Loss Estimation Methodology. Emergency Preparedness and Response Directorate -- Department of Homeland Security. Washington, D. C. : Federal Emergency Management Association
- FEMA (1992a). A Benefit-Cost Model for the Seismic Rehabilitation of Buildings - Volume 1 (A User's Manual). FEMA 227. Washington, D. C.: Federal Emergency Management Association
- FEMA (1992b). A Benefit-Cost Model for the Seismic Rehabilitation of Buildings - Volume 2 (Supporting Documentation). FEMA 228. Washington, D. C.: Federal Emergency Management Association
- FEMA (1994). Typical Costs for Seismic Rehabilitation of Buildings - Volume 1 (Summary). FEMA 156. Washington, D. C.: Federal Emergency Management Association
- FEMA (1995). Typical Costs for Seismic Rehabilitation of Buildings - Volume 2 (Supporting Documentation). FEMA 157. Washington, D. C.: Federal Emergency Management Association
- FEMA (2000). Disaster Mitigation Act of 2000. Public Law 106-390. Washington, D.C.: Federal Emergency Management Association: 26.
- FEMA (2001). Understanding your Risks -- Identifying hazards and estimating losses. State and Local Mitigation Planning How-to Guide. FEMA 386-2. Washington, D. C.: Federal Emergency Management Association: 168.
- FEMA (2002a). Getting started -- Building support for mitigation planning. State and Local Mitigation Planning How-to Guide. FEMA 386-1. FEMA. Washington, D. C.: Federal Emergency Management Association
- FEMA (2002b). Handbook for the Seismic Evaluation of Buildings - A Prestandard, prepared by the American Society of Civil Engineers (ASCE). FEMA Hazard

Mitigation Handbooks. FEMA 310. Washington, D.C.: Federal Emergency Management Association

FEMA (2004). Using HAZUS-MH for Risk Assessment. HAZUS-MH Risk Assessment and User Group Series. FEMA 433. Federal Emergency Management Association. Washington, D. C. : Federal Emergency Management Association

FEMA (2007). Robert T. Stafford Disaster Relief and Emergency Assistance Act, as amended, and Related Authorities. FEMA 592: Public Law 93-288. Washington, D.C.: Federal Emergency Management Association: 125.

Fischler, M. A. and H. C. Wolf (1994). "Locating perceptually salient points on planar curves." IEEE Transactions on Pattern Analysis and Machine Intelligence **16**(2): 113-129.

Fletcher, Roger (1987). *Practical Methods of Optimization*. New York: Wiley.

Fowler, Robert (2000). Topographic LIDAR. In *Digital Elevation Model Technologies and Applications: The DEM Users Manual*. D. F. Maune, Ed. Bethesda, MD: The American Society for Photogrammetry and Remote Sensing: 207-236.

Freiss, T-T. and R. Harrison (1998). Support Vector neural networks: The Kernel Adatron with bias and soft margin - Technical Report ACSE-TR-752. Sheffield, UK: Department of ACSE, University of Sheffield.

French, Steven P., John C. Castanon and Alan Henson (1992). A Knowledge-Base Approach to Using Existing Data for Seismic Risk Assessment. NSF Report BCS-8822125. San Luis Obispo, CA: Department of City and Regional Planning, California Polytechnic State University.

French, Steven P. and Mark S. Isaacson (1984). "Applying earthquake risk analysis techniques to land use planning." Journal of the American Planning Association **50**(4): 509-522.

French, Steven P. and Subrahmanyam Muthukumar (2006). "Advanced Technologies for Earthquake Risk Inventories " Journal of Earthquake Engineering **10**(2): 207-236.

Fu, King Sun (1982). *Syntactic Pattern Recognition and Applications*. Englewood Cliffs, NJ: Prentice-Hall.

Fu, King Sun (1986). "A step towards unification of syntactic and statistical pattern recognition." IEEE Transactions on Pattern Analysis and Machine Intelligence **8**(3): 398-404.

- Fukunaga, Keinosuke (1990). *Introduction to Statistical Pattern Recognition*. Second Edition. Boston: Academic Press.
- Gabriel, A. K. and R. M. Goldstein (1988). "Repeat Pass Interferometry." International Journal of Remote Sensing **9**: 857-872.
- Gdalyahu, Yoram and Daphna Weinshall (1999). "Flexible syntactic matching of curves and its application to automatic hierarchical classification of silhouettes." IEEE Transactions on Pattern Analysis and Machine Intelligence **21**(12): 1312-1328.
- Geman, S. (1992). "Neural networks and the bias/variance dilemma." Neural computation **4**: 1-58.
- Gil-Jimenez, P., S. Lafuente-Arroyo, H. Gomez-Moreno et al. (2005). Traffic sign shape classification evaluation. Part II. FFT applied to the signature of blobs. Proceedings of the Intelligent Vehicles Symposium, 2005, Traverse City, MI: IEEE.
- Godschalk, David R. and Stephen Baxter (2002). Urban hazard mitigation: Creating resilient cities. Urban Hazards Forum. New York, N. Y.: CUNY.
- Godschalk, David R., Timothy Beatley, Philip Berke et al. (1999). *Natural Hazard Mitigation: Recasting Disaster Policy and Planning*. Washington, D. C. : Island Press.
- Godschalk, David R., Samuel Brody and Raymond Burby (2003). "Public participation in natural hazard mitigation policy formation: Challenges for comprehensive planning." Journal of Environmental Planning and Management **46**(5): 733-754.
- Godschalk, David R., Edward J. Kaiser and Philip R. Berke (1998). Integrating Hazard Mitigation and Local Land Use Planning. In *Cooperating with Nature: Confronting Natural Hazards with Land Use Planning for Sustainable Communities*. R. J. Burby, Ed. Washington, D. C.: Joseph Henry/The National Academies Press: 85-118.
- Goldberger, Arthur Stanley (1991). *A Course in Econometrics*. Cambridge, MA: Harvard University Press.
- Golland, Polina, W. Eric L. Grimson, Martha E. Shenton et al. (2005). "Detection and analysis of statistical differences in anatomical shape." Medical Image Analysis **9**: 69-86.
- Gonzalez, Rafael C. and Michael C. Thomason (1978). *Syntactic Pattern Recognition: An Introduction*. Reading, MA: Addison-Wesley

- Greene, William H. (2008). *Econometric Analysis*. 6th. Upper Saddle River, NJ: Prentice-Hall.
- Grenander, Ulf (1996). *Elements of Pattern Theory*. Baltimore: Johns Hopkins University Press.
- Gribov, Alexander and Eugene Bodansky (2004). A new method of polyline approximation. A. Fred, T. Caelli, R. P. W. Duin, A. Campilho and D. d. Ridder, Eds. Proceedings of the Structural, Syntactic, and Statistical Pattern Recognition Joint IAPR International Workshops, SSPR 2004 (Structural and Syntactic Pattern Recognition) and SPR 2004 (Statistical Techniques in Pattern Recognition). Lisbon, Portugal: Springer Berlin/Heidelberg.
- Grigorishin, T., G. Abdel-Hamid and Y. -H. Yang (1998). "Skeletonisation: An electrostatic field-based approach." Pattern Analysis & Applications **1**(3): 163-177.
- Gupta, L. and Mandyam D. Srinath (1987). "Contour sequence moments for the classification of closed planar shapes." Pattern Recognition **20**(3): 267-272.
- Gupta, L. and Mandyam D. Srinath (1988). "Invariant planar shape recognition using dynamic alignment." Pattern Recognition **21**(3): 235-239.
- Haddow, George D., Jane A. Bullock and Damon P. Coppola (2008). *Introduction to Emergency Management*. Third Edition. Burlington, MA: Elsevier, Inc.
- Hagan, Martin T., Howard B. Demuth and Mark Beale (1996). *Neural Network Design*. Boulder: University of Colorado Press.
- Hanson, S. J. (1990). Meiosis networks In *Advances in Neural Information Processing Systems 2*. D. S. Touretzky, Ed. San Francisco: Morgan Kaufmann: 533-541.
- Haykin, Simon (1994). *Neural Networks: A Comprehensive Foundation*. Second Edition. New York: Macmillan Publishing.
- Hebb, Donald O. (1949). *The Organization of Behavior*. New York: Wiley.
- Hensher, David A., John M. Rose and William H. Greene (2005). *Applied Choice Analysis: A Primer*. Cambridge, UK: Cambridge University Press.
- Hinton, Geoffrey E. and T. J. Sejnowski (1986). Learning and relearning in Boltzmann machines. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. D. E. Rumelhart and J. L. McClelland, Eds. Cambridge, MA: MIT Press: 282-317.

- Hopfield, J. J. (1982). "Neural networks and physical systems with emergent collective computational abilities." Proceedings of the National Academy of Sciences **79**: 2554-2558.
- Hu, Ming-Kuei (1962). "Visual pattern recognition by moment invariants." IEEE Transactions on Information Theory **8**(2): 179-187.
- ICC (2000). International Building Code 2000. International Conference of building Officials, Whittier, CA: International Code Council: 756.
- Institute of Electrical and Electronics Engineers, Inc., IEEE. (2008). "IEEE Journals and Magazines." Transactions on Geoscience and Remote Sensing. J. A. Benediktsson, Ed. IEEE Geoscience and Remote Sensing Society. Retrieved March 11, 2008, from <http://ieeexplore.ieee.org/servlet/opac?punumber=36>.
- Jain, Anil K. and Richard C. Dubes (1988). *Algorithms for Clustering Data*. Englewood Cliffs, N. J.: Prentice Hall.
- Jain, Anil K., Robert P.W. Duin and Jianchang Mao (2000). "Statistical pattern recognition: A review." IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(1): 4-37.
- Jiang, X. Y. and H. Bunke (1991). "Simple and fast computation of moments." Pattern Recognition **24**(8): 801-806.
- Jin, X. and C. H. Davis (2005). "Automated building extraction from high-resolution satellite imagery in urban areas using Structural, Contextual, and Spectral information." EURASIP Journal on Applied Signal Processing **14**: 2196–2206.
- Joao, E. M. (1998). *Causes and consequences of map generalization*. London: Taylor & Francis.
- Johnston, Mark R., Christine D. Scott and Robert G. Gibb (1999). Problems arising from a simple GIS Generalisation Algorithm. The 11th Annual Colloquium of the Spatial Information Research Centre, SIRC 99. Dunedin, New Zealand.
- Jones, Barclay G. (1978). "The Eclecticism of Regional Science – expanding the choices of scientific method: With an application to estimating building stocks." Northeast Regional Science Review **8**: 1-19.
- Jones, Barclay G. and Stephanie E. Chang (1994). A comparison of indirect and direct estimates of the built physical environment in the Memphis region. Proceedings of the Fifth U.S. National Conference on Earthquake Engineering: Earthquake Awareness and Mitigation Across the Nation, Chicago, IL: Chicago, IL: Earthquake Engineering Research Institute.

- Jones, Barclay G. and Ajay Madan Malik (1997). Building Inventory. In *Loss Estimation of Memphis Buildings*. D. P. Abrams and M. Shinozuka, Eds. Buffalo: National Center for Earthquake Engineering Research, State University of New York at Buffalo: 11-20.
- Jones, Barclay G., Donald M. Manson, Charles M. Hotchkiss et al. (1987). *Estimating Building Stocks and their Characteristics*. Ithaca, NY: Cornell Institute for Social and Economic Research.
- Kaiser, Edward J., David R. Godschalk and Stuart F. Chapin (1995). *Urban Land Use Planning*. Fourth Edition. Urbana: University of Illinois Press.
- Kashyap, Rangasami L. and Ramalingam Chellappa (1981). "Stochastic models for closed boundary analysis: Representation and reconstruction." IEEE Transactions on Information Theory **27**(5): 627 - 637.
- Katz, Robert A. and Stephen M. Pizer (2003). "Untangling the Blum Medial Axis Transform." International Journal of Computer Vision **55**(2-3): 139–153.
- Kauppinen, Hannu, Tapio Seppanen and Matti Pietikainen (1995). "An experimental comparison of autoregressive and Fourier-based descriptors in 2D shape classification." IEEE Transactions on Pattern Analysis and Machine Intelligence **17**(2): 201-207.
- Kavzoglu, Taskin, Jasmee Jaafar and Paul M. Mather (2000). Extraction of field boundary information from classified satellite images. Proceedings of GIS Research, York, UK.
- Kavzoglu, Taskin and Paul M. Mather (2000). The use of feature selection techniques in the context of artificial neural networks. Proceedings of the 26th Annual Conference of the Remote Sensing Society, Leicester, UK.
- Kaygin, Serkan and M. Mete Bulut (2002). "Shape recognition using attributed string matching with polygon vertices as the primitives." Pattern Recognition Letters **23**(1-3): 287-294.
- Kazemi, Sharon (2003). A generalization framework to derive multi-scale GEODATA. Proceedings of the Spatial Sciences Conference, Canberra, Australia.
- Kazemi, Sharon, Samsung Lim and C. Rizos (2001). A review of map and spatial database generalization for developing a generalization framework. Proceedings of the 4th Workshop ACI on Progress in Automated Map Generalisation, Beijing, China.
- Kendall, David G. (1984). "Shape manifolds, Procrustean metrics and complex projective spaces." Bulletin of the London Mathematical Society **16**: 81-121.

- Kircher, C. A. (2003). It makes dollars and sense to improve nonstructural system performance. Proceedings of Seminar on Seismic Design, Performance, and Retrofit of Nonstructural Components in Critical Facilities, ATC 29-2, Newport Beach, CA: Applied Technology Council.
- Kiryati, N. and D. Maydan (1989). "Calculating geometric properties from Fourier representation." Pattern Recognition **22**(5): 469-475.
- Lakhan, Chris V. (1996). *Introductory Geographical Information Systems*. Toronto, Canada: Summit Press.
- Lam, Louisa, Seong-Whan Lee and Ching Y. Suen (1992). "Thinning methodologies - a comprehensive survey." IEEE Transactions on Pattern Analysis and Machine Intelligence **14**(9): 869-885.
- Langley, Pat (1996). *Elements of Machine Learning*. San Francisco: Morgan Kaufmann.
- Latecki, Longin Jan and Rolf Lakamper (1999). Polygon evolution by vertex deletion. M. Nielsen, P. Johansen, O. F. Olsen and J. Weickert, Eds. Proceedings of the Second International Conference on Scale-Space Theories in Computer Vision, Corfu, Greece: Springer-Verlag.
- Latecki, Longin Jan and Rolf Lakamper (2000). "Shape similarity measure based on correspondence of visual parts." IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(10): 1185- 1190.
- Lee, Dan Scott (2003). Generalization within a geoprocessing framework. International Workshop on Semantic Processing of Spatial Data (GEOPRO 2003), Mexico City, Mexico.
- Lee, Dan Scott, Jie Shan and James S. Bethel (2003). Class-guided building extraction from Ikonos imagery. Photogrammetric Engineering and Remote Sensing, Journal of the American Society for Photogrammetry and Remote Sensing. **69**(2): 143–150.
- Leu, Jia-Guu and Limini Chen (1988). "Polygonal approximation of 2-D shapes through boundary merging." Pattern Recognition Letters **7**(4): 231 238.
- Leventon, Michael E., W. Eric L. Grimson and Olivier Faugeras (2000). Statistical shape influence in geodesic active contours. 2000 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '00), Hilton Head Island, SC: IEEE Computer Society.
- Li, Xia and Anthony Gar-on Yeh (2002). "Neural-network-based cellular automata for simulating multiple land use changes using GIS." International Journal of Geographical Information Science **16**(4): 323-343.

- Li, Yajun (1992). "Reforming the theory of invariant moments for pattern recognition." Pattern Recognition **25**(7): 723-730.
- Limeng, L. and W. Lixin (2001). Map generalization from scale of 1:500,000 to 1:2,500,000. Proceedings of the 20th International Cartography Conference, Beijing, China.
- Liu, Hong-Chih and Mandyam D. Srinath (1990). "Partial shape classification using contour matching in distance transformation." IEEE Transactions on Pattern Analysis and Machine Intelligence **12**(11).
- Loncaric, Sven (1998). "A survey of shape analysis techniques." Pattern Recognition **31**(8): 983-1001.
- Long, J. Scott (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: SAGE Publications.
- Long, J. Scott and Jeremy Freese (2006). *Regression Models for Categorical Dependent Variables Using Stata*. College Station, TX: Stata Press.
- Lopez, Oscar A. and Elizabeth Raven (1999). "An overall evaluation of irregular-floor-plan-shaped buildings located in seismic areas." Earthquake Spectra **15**(1): 105-120.
- Luenberger, David G. (1984). *Linear and Nonlinear Programming*. Reading, MA: Addison-Wesley.
- Luger, George F. (2002). *Artificial Intelligence, Structures and Strategies for Complex Problem Solving*. Fourth Edition. Harlow, England: Addison-Wesley.
- Makhfi, Pejman. (2007, January 12). "Introduction to Knowledge Modeling and Neural Networks." Retrieved September 27, 2007, from <http://www.makhfi.com/index.htm>.
- Malik, Ajay Madan (1995). Building Stocks Cross-classified by Use and Structural Type: Memphis-Shelby County, Tennessee 1992 and Wichita-Sedgwick County, Kansas 1982. Working Paper in Estimating Building Stocks for Earthquake Mitigation and Recovery Planning. Ithaca, NY: Cornell Institute for Social and Economic Research Program in Urban and Regional Studies.
- May, Peter J. (1985). *Recovering from Catastrophes: Federal Disaster Relief Policy and Politics*. Westport, CT: Greenwood Press.
- McCulloch, Warren and Walter Pitts (1943). "A logical calculus of the ideas immanent in nervous activity." Bulletin of Mathematical Biophysics **5**: 115-133.

- McFadden, Daniel (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In *Frontiers in Econometrics*. P. Zarembka, Ed. New York: Academic Press: 105-142.
- McKeown, David, Jeff McMahill and Douglas Caldwell (1999). The use of spatial context in linear feature simplification. The IV International Conference on GeoComputation, Fredericksburg, VA.
- McLachlan, Geoffrey J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley.
- McMaster, Robert B. and K. Stuart Shea (1988). Cartographic generalization in a digital environment: A framework for implementation in a Geographic Information System. Proceedings of the Third Annual International Conference in GIS/LIS: Accessing the World, San Antonio, TX: ACPRS-ACSM-URISA.
- McMaster, Robert B. and K. Stuart Shea (1989). Cartographic generalization in a digital environment: When and how to generalize. E. Anderson, Ed. Proceedings of the Ninth International Symposium on Computer-Assisted Cartography (Auto-Carto 9), Baltimore, MD: ACPRS-ACSM.
- McMaster, Robert B. and K. Stuart Shea (1992). *Generalization in Digital Cartography*. Washington, DC: Association of American Geographers.
- Michie, D., D. J. Spiegelhalter and C. C. Taylor. (1994, April 16, 1999). "Machine Learning, Neural and Statistical Classification." University of Leeds. Retrieved October 20, 2007, from <http://www.amsta.leeds.ac.uk/~charles/statlog/>.
- Mid-America Earthquake Center. (2006). "Mid-America Earthquake Center -- Research." Retrieved October 22, 2007, from <http://mae.ce.uiuc.edu/research/index.html>.
- Minsky, Marvin and Seymour Papert (1969). *Perceptrons*. Cambridge, MA: MIT Press.
- Mollander, Craig W. (2000). Photogrammetry. In *Digital Elevation Model Technologies and Applications: The DEM Users Manual*. D. F. Maune, Ed. Bethesda, MD: The American Society for Photogrammetry and Remote Sensing: 121-142.
- Morse, Bryan. (2007, August 1). "Points of extreme curvature: Sec. 7.6.7. Available: http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/MORSE/boundary_rep_desc.pdf." In *CVonline: On-Line Compendium of Computer Vision [Online]*. R. B. Fisher, Ed. Retrieved February 14, 2008, from <http://homepages.inf.ed.ac.uk/rbf/CVonline/>.
- Muller, Jean-Claude, Jean-Philippe Lagrange and Robert Weibel, Eds. (1995a). *GIS and Generalization: Methodology and Practice*. Bristol, England: Taylor & Francis.

- Muller, Jean-Claude, Robert Weibel, Jean-Phillippe Lagrange et al. (1995b). Generalization: state of the art and issues. In *GIS and Generalization: Methodology and Practice*. J.-C. Muller, J.-P. Lagrange and R. Weibel, Eds. Bristol, England: Taylor & Francis: 3-18.
- Multihazard Mitigation Council (2005a). Natural Hazard Mitigation Saves: An Independent Study to Assess the Future Savings from Mitigation Activities. Volume 1 -- Findings, Conclusions and Recommendations. National Institute of Building Sciences. Washington, D. C. : NIBS
- Multihazard Mitigation Council (2005b). Natural Hazard Mitigation Saves: An Independent Study to Assess the Future Savings from Mitigation Activities. Volume 2 -- Study Documentation. National Institute of Building Sciences. Washington, D. C. : NIBS
- Murty, C. V. R. (2002 a, September). "How Architectural Features Affect Buildings During Earthquakes?". National Information Center of Earthquake Engineering. Retrieved February 4, 2008, from <http://www.iitk.ac.in/nicee/EQTips/EQTip06.pdf>.
- Murty, C. V. R. (2002 b, October). "How Buildings Twist During Earthquakes?". National Information Center of Earthquake Engineering. Retrieved February 4, 2008, from <http://www.iitk.ac.in/nicee/EQTips/EQTip07.pdf>.
- National Research Council (1999). *The Impacts of Natural Disasters: A Framework for Loss Estimation*. Washington, D. C. : The National Academies Press.
- National Research Council (2006). *Improved Seismic Monitoring - Improved Decision-Making: Assessing the Value of Reduced Uncertainty*. Washington, D. C. : The National Academies Press.
- National Research Council and the Division on Earth and Life Studies (2006). *Facing Hazards and Disasters: Understanding Human Dimensions*. Washington, D. C.: The National Academies Press.
- National Science and Technology Council. (2003, July). "Reducing Disaster Vulnerability through Science and Technology: An Interim Report of the Subcommittee on Disaster Reduction." National Oceanic and Atmospheric Administration, National Environmental, Satellite, Data and Information Service. Retrieved May 10, 2008, from http://www.sdr.gov/SDR_Report_ReducingDisasterVulnerability2003.pdf.
- National Science and Technology Council. (2005, June). "Grand Challenges for Disaster Reduction: Earthquake." National Oceanic and Atmospheric Administration, National Environmental, Satellite, Data and Information Service. Retrieved May 27, 2008, from http://www.sdr.gov/185820_Earthquake_FINAL.pdf.

- Neaupane, K. M. and S. H. Achet (2004). "Use of back propagation neural network for landslide monitoring: A case study in the higher Himalaya." Engineering geology **74**(3-4): 213-226.
- Nelson, Arthur C. and Steven P. French (2002). "Plan quality and mitigating damage from natural disasters." Journal of the American Planning Association **68**(2): 194-207.
- Nilsson, Nils. (1996, August 22, 2005). "Introduction to Machine Learning -- An Early Draft of a Proposed Textbook." Retrieved October 14, 2007, from <http://robotics.stanford.edu/people/nilsson/MLDraftBook/MLBOOK.pdf>.
- Niu, Xutong, Rongxing Li and Morton O'Kelly (2002). Truck Detection from Aerial Photographs. Proceedings of the ISPRS Technical Commission II Symposium: Integrated Systems for Spatial Data Production, Custodian and Decision Support, Xi'an, China: ISPRS.
- Ogniewicz, R. L. (1993). *Discrete Voronoi Skeletons*. Konstanz, Germany: Hartung-Gorre Verlag.
- Olshansky, Robert B. (2001). "Land use planning for seismic safety -- The Los Angeles experience, 1971-1994." Journal of the American Planning Association **67**(2): 173-185.
- Oosterom, P. van (1995). The GAP-tree, and approach to 'on-the-fly' map generalization of an area partitioning. In *GIS and Generalization: Methodology and Practice*. J.-C. Muller, J.-P. Lagrange and R. Weibel, Eds. Bristol, England: Taylor & Francis: 120-132.
- Patterson, Dan W. (1996). *Artificial Neural Networks*. New York: Prentice Hall.
- Pavlidis, Theodosios (1972). Structural pattern recognition: primitives and juxtaposition relations. In *Frontiers of Pattern Recognition*. S. Watanabe, Ed. New York: Academic Press.
- Pavlidis, Theodosios (1978). "A review of algorithms for shape analysis." Computer Graphics and Image Processing **7**: 243-258.
- Pavlidis, Theodosios (1979). "Hierarchies in structural pattern recognition." Proceedings of the IEEE **67**(5): 737- 744.
- Pavlidis, Theodosios (2003). "36 years on the pattern recognition front: Lecture given at ICPR'2000 in Barcelona, Spain, on the occasion of receiving the K.S. Fu prize." Pattern Recognition Letters **24**(1-3): 1-7.

- Pearce, Laurie (2003). "Disaster management and community planning, and public participation: How to achieve sustainable hazard mitigation." Natural Hazards **28**(2): 211-228.
- Pentland, A. (1987). Recognition by parts. Proceedings of the First International Conference on Computer Vision, London, England: Computer Society of the IEEE -- IEEE Computer Society Press: Los Alamitos, CA.
- Persoon, E. and K. S. Fu (1977). "Shape discrimination using Fourier descriptors." IEEE Transactions on Systems, Man and Cybernetics **7**: 534-541.
- Porter, Keith A. (2002). Learning from earthquakes: a survey of surveys. EERI Invitational Workshop: An Action Plan to Develop Earthquake Damage and Loss Data Protocols, Pasadena, CA: Earthquake Engineering Research Institute.
- Porter, Keith A. (2005). A Taxonomy of Building Components for Performance-Based Earthquake Engineering - PEER Report 2005/03. Berkeley, CA: University of California, Berkeley.
- Porter, Keith A., Anne S. Kiremidjian and Jeremiah S. LeGrue (2001). "Assembly-based vulnerability of buildings and its use in performance evaluation." Earthquake Spectra **17**(2): 291-312.
- Prager, R. W. and F. Fallside (1989). "The modified Kanerva model for automatic speech recognition." Computer Speech and Language **3**: 61-82.
- Principe, Jose C., Neil R. Euliano and W. Curt Lefebvre (2000). *Neural and Adaptive Systems: Fundamentals through Simulation*. New York: John Wiley & Sons, Inc.
- R. S. Means (2008). *Square Foot Costs*. Kingston, MA: Reed Construction Data, Inc.
- Ramachandran, V. S. and Sandra Blakeslee (1998). *Phantoms in the Brain: Probing the Mysteries of the Human Mind*. New York: Quill William Morrow.
- Realpe, Alvaro and Carlos Velázquez (2006). "Pattern recognition for characterization of pharmaceutical powders." Powder Technology **169**(2): 108-113.
- Reddy, Swaroop D. (2000). "Examining hazard mitigation within the context of public goods." Environmental Management **25**(2): 129-141.
- Reitherman, Robert (1998). The need for improvement in post-earthquake investigations of the performance of nonstructural components. Proceedings of Seminar on Seismic Design, Retrofit and Performance of Nonstructural Components, ATC 29-1, San Francisco, CA: Applied Technology Council.

- Richard, C. W. and H. Hemami (1974). "Identification of three dimensional objects using Fourier descriptors of the boundary curve." IEEE Transactions on Systems, Man and Cybernetics **4**(4): 371-378.
- Richter, Charles F. (1957). *Elementary Seismology*. San Francisco, CA: W. H. Freeman Co.
- Ripley, Brian D. (1996). *Pattern Recognition and Neural Networks*. Cambridge, UK: Cambridge University Press.
- RMS & CUREe (1993). Assessment of the State of the Art Earthquake Loss Estimation Methodologies -- Task 1 Report prepared by: Risk Management Software, Inc. and California Universities for Research in Earthquake Engineering. Washington, D. C.: National Institute of Building Sciences and the Federal Emergency Management Association. **ENW-92-IL-3973**: 376.
- Rodrigues, E. and J. M. Martin (1992). "Theory and Design of Interferometric Synthetic Aperture Radar." IEEE Proceedings **139**(2): 147-159.
- Rojas, Raul (1995). *Neural Networks: A Systematic Introduction*. New York: Springer-Verlag.
- Rose, Adam Zachary (2004). Economic Principles, Issues, and Research Priorities in Hazard Loss Estimation. In *Modeling Spatial And Economic Impacts Of Disasters*. Y. Okuyama, S. E.-L. Chang and S. E. Chang, Eds. New York: Springer-Verlag: pp. 13-36.
- Rose, Adam Zachary, J. Benavides, S.E. Chang et al. (1997). "The regional economic impact of an earthquake: direct and indirect effects of electricity lifeline disruptions." Journal of Regional Science, **37**: 437-458.
- Rose, Adam Zachary and Howard Kunreuther, Eds. (2004). *The Economics of Natural Hazards*. Northampton, MA: Edward Elgar Publishing.
- Rose, Adam Zachary and Dongsoo Lim (2002). "Business interruption losses from natural hazards: conceptual and methodological issues in the case of the Northridge earthquake." Global Environmental Change Part B: Environmental Hazards **4**(1): 1-14.
- Rosenblatt, Frank (1958). "The perceptron: A probabilistic model for information storage and organization in the brain." Psychological Review **65**: 386-408.
- Rothe, Irene, Herbert Susse and Klaus Voss (1996). "The method of normalization to determine invariants." IEEE Transactions on Pattern Analysis and Machine Intelligence **18**(4): 366-376.

- Ruas, A. and C. Plazanet (1996). Strategies for automated generalization. In *Advances in GIS Research II: Proceedings of the 7th International Symposium on Spatial Data Handling*. M.-J. Kraak, M. Molenaar and E. M. Fendel, Eds. London: Taylor & Francis: 319-335.
- Rumelhart, David E., Geoffrey E. Hinton and R. J. Williams (1986). Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. D. E. Rumelhart and J. L. McClelland, Eds. Cambridge, MA: MIT Press: 318-362.
- Rumelhart, David E. and James L. McClelland, Eds. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press.
- Russell, Stuart and Peter Norvig (1995). *Artificial Intelligence: A Modern Approach*. Second Edition. Englewood Cliffs, NJ: Prentice-Hall.
- Saeki, T., H. Tsubokawa and S. Midorikawa (2000). Seismic damage evaluation of household property by using geographic information systems (GIS) - Paper 1968. Proceedings of the 12th World Conference on Earthquake Engineering, Auckland, New Zealand: International Association for Earthquake Engineering.
- Sahar, Liora and Amnon Krupnik (1999). "Semiautomatic Extraction of Building Outlines from Large-Scale Aerial Images." Photogrammetric Engineering and Remote Sensing **65**(4): 459-465.
- Salih, Nghan D., David Chek Ling Ngo and Hakim Mellah (2006). "2D object description with discrete segments." Journal of Computer Science **2**(7): 572-576.
- Savonis, Michael John (1985). The Residential Building Stock: Characteristics and Trends in Wichita-Sedgwick County, Kansas. Working Paper in Estimating Building Stocks for Earthquake Mitigation and Recovery Planning. Ithaca, NY: Cornell Institute for Social and Economic Research Program in Urban and Regional Studies.
- Schalkoff, Robert J. (1992). *Pattern Recognition - Statistical, Structural and Neural Approaches*. New York: John Wiley & sons.
- Sebastian, Thomas B., Philip N. Klein and Benjamin B. Kimia (2001). Recognition of shapes by editing shock graphs. Proceedings of the Eighth IEEE International Conference on Computer Vision, Vancouver, BC, Canada.
- Shahrokni, Ali, Tom Drummond and Pascal Fua. (2004, June 21). "Minkowski-form: Non-parametric measures. Available: http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/SHAHROKNI1/nod_e8.html." In CVonline: On-Line Compendium of Computer Vision [Online]. R. B.

- Fisher, Ed. Retrieved February 14, 2008, from <http://homepages.inf.ed.ac.uk/rbf/CVonline/>.
- Sherman, Deborah. (2007, 3/30/2007). "Undercover agents slip bombs past DIA screeners." 9news.com. Retrieved August 27, 2007, from <http://www.9news.com/news/article.aspx?storyid=67166>.
- Shinozuka, M., Adam Rose and Ron Eguchi, Eds. (1998). *Engineering and Socioeconomic Impacts of Earthquakes: An Analysis of Electricity Lifeline Disruptions in the New Madrid Area*. Buffalo, NY: MCEER.
- Silva, Fernando M. and Luis B. Almeida (1990). Acceleration techniques for the back-propagation algorithm. In *Neural Networks*. L. B. Almeida and C. J. Wellekens, Eds. New York: Springer: 110-119.
- Skiena, Steven S. (1997, June 2). "Medial-Axis Transformation." The Algorithm Design Manual. Springer-Verlag. Retrieved March 18, 2008, from <http://www2.toki.or.id/book/AlgDesignManual/BOOK/BOOK5/NODE193.HTM>.
- Skopeliti, Andriani and Lysandros Tsoulos (2001). A knowledge based approach for the generalization of linear features. Proceedings of the 20th International Cartography Conference, Beijing, China.
- Sohn, Hong-Gyoo, Choung-Hwan Park, Ho-Sung Kim et al. (2005). 3-D building extraction using IKONOS multispectral images. Proceedings of the 2005 Geoscience and Remote Sensing Symposium (IGARSS 2005), Seoul, South Korea.
- Srinivasan, Ashwin (2001). "Extracting context-sensitive models in Inductive Logic Programming." Machine Learning **44**(3): 301-324.
- StatSoft, Inc. (2003). "Neural Networks." Retrieved 2007, March 14, from <http://www.statsoft.com/textbook/stneunet.html>.
- Super, Boaz J. (2004). "Fast correspondence-based system for shape-retrieval." Pattern Recognition Letters **25**(2): 217-225.
- Taghavi, S. and E. Miranda (2003). Response Assessment of Nonstructural Building Elements - PEER Report 2003/05. Richmond, CA: Pacific Earthquake Engineering Research Center.
- Teague, M. R. (1980). "Image analysis via the general theory of moments." Journal of Optical Society of America **70**: 920-930.
- Torsello, Andrea and Edwin R. Hancock (2004). "A skeletal measure of 2D shape similarity." Computer Vision and Image Understanding **95**: 1-29.

- Tralli, David M. (2000). Assessment of Advanced Technologies for Loss Estimation: Multidisciplinary Center for Earthquake Engineering Research. University at Buffalo, State University of New York.
- Trier, Oivind Due, Anil K. Jain and Torfinn Taxt (1996). "Feature extraction methods for character recognition - A survey." Pattern Recognition **29**(4): 641-662.
- Tun, Aung Hla. (2008, May 6). "Myanmar cyclone toll climbs to nearly 22,500." Thomson Reuters, International Edition. Retrieved May 25, 2008, from <http://www.reuters.com/article/topNews/idUSBKK1919620080506?feedType=RSS&feedName=topNews>.
- United Nations (2001). Disaster Reduction and Sustainable Development: Understanding the links between development, environment and natural disasters. World Summit on Sustainable Development. Geneva, Switzerland: United Nations International Strategy for Disaster Reduction.
- United Nations (2003). Disaster Reduction and Sustainable Development: Understanding the links between vulnerability and risk to disasters related to development and environment. World Summit on Sustainable Development. Geneva, Switzerland: United Nations International Strategy for Disaster Reduction.
- United Nations (2005). Building the Resilience of Nations and Communities to Disasters: Hyogo Framework for Action 2005-2015 United Nations World Conference on Disaster Reduction, Kobe, Hyogo, Japan: Inter-Agency Secretariat of the International Strategy for Disaster Reduction.
- US Census Bureau. (2008, January 2). "Shelby County QuickFacts from the US Census Bureau." State and County QuickFacts. US Census Bureau. Retrieved March 27, 2008, from <http://quickfacts.census.gov/qfd/states/47/47157.html>.
- US General Accounting Office (2003). Disaster Assistance: Information on FEMA's Post 9/11 Public Assistance to the New York City Area. Report to the Committee on Environment and Public Works, US Senate. Washington, D.C.: US General Accounting Office. **GAO-03-926**: 48.
- Vapnik, Vladimir Naumovich (1999). *The Nature of Statistical Learning Theory*. Second Edition. New York: Springer-Verlag.
- Veltkamp, Remco C. (2001). Shape matching: Similarity measures and algorithms. Utrecht: Technical Report UU-CS-2001-03, Utrecht University.
- Veltkamp, Remco C. and Michiel Hagedoorn (1999). State-of-the-art in shape matching. Utrecht: Technical Report UU-CS-1999-27, Utrecht University.

- Visvalingam, Mahes and J. D. Whyatt (1990). "The Douglas-Peucker algorithm for line simplification: Re-evaluation through visualization." Computer Graphics Forum **9**: 213-228.
- Wang, Shuenn-Shyang, Po-Cheng Chen and Wen-Gou Lin (1994). "Invariant pattern recognition by moment fourier descriptor." Pattern Recognition **27**(12): 1735-1742.
- Waugh, William L. Jr. (2000). *Living with Hazards; Dealing with Disasters: An Introduction to Emergency Management*. Armonk, N. Y.: M. E. Sharpe.
- Webb, Andrew R. (2002). *Statistical Pattern Recognition*. Second. New York: John Wiley & Sons, Inc.
- Wei, Yanfeng, Zhongming Zhao and Jianghong Song (2004). Urban building extraction from high-resolution satellite panchromatic image using clustering and edge detection. Proceedings from Geoscience and Remote Sensing Symposium, IEEE International (IGARSS 2004), Anchorage, AL.
- Weibel, Robert and G. Dutton (1999). Generalizing spatial data and dealing with multiple representations. In *Geographical Information Systems: Volume 1. Principles and Technical Issues*. 2nd Edition. P. A. Longley, M. F. Goodchild, D. J. Maguire and D. W. Rhind, Eds. New York: John Wiley & Sons: 125-155.
- Weibel, Robert and Christopher B. Jones (1998). "Computational perspectives on map generalization." Geoinformatica **2**(4): 307-314.
- Werbos, Paul John (1974). *Beyond Regression: New Tools for prediction and analysis in the behavioural sciences*. Cambridge, MA: Harvard University.
- Whittaker, A. S. and T. T. Soong (2003). An overview of nonstructural components research at three US earthquake engineering research centers. Proceedings of Seminar on Seismic Design, Performance, and Retrofit of Nonstructural Components in Critical Facilities, ATC 29-2, Newport Beach, CA: Applied Technology Council.
- Widrow, Bernard and M.E. Hoff (1960). Adaptive switching circuits. IRE WESCON Convention Record. New York: IRE Part 4: 96-104.
- Widrow, Bernard and Samuel Sterns (1985). *Adaptive Signal Processing*. Upper Saddle River, NJ: Prentice Hall.
- Wu, W. Y. and M. J. J. Wang (1999). "Two-dimensional object recognition through twostage string matching." IEEE Transactions on Image Processing **8**(7): 978-981.

- Wynne-Jones, M. (1993). "Node splitting: A constructive algorithm for feed-forward neural networks." Neural Computing and Applications **1**(1): 17-22.
- Yang, W. and C. M. Gold (1997). A system approach to automated map generalization. Y. C. Lee and Z. L. Li, Eds. Proceedings of International Workshop on Dynamic and Multi-Dimensional GIS, Hong Kong, China.
- Yang, Y. P. and Theodosios Pavlidis (1990). "Optimal correspondence of string subsequences." IEEE Transactions on Pattern Analysis and Machine Intelligence **12**(11): 1080-1087.
- Young, I., J. Walker and J. Bowie (1974). "An analysis technique for biological shape." Computer Graphics and Image Processing **25**: 357-370.
- Zahn, Charles and Ralph Roskies (1972). "Fourier descriptors for plane closed curves." Computer Graphics and Image Processing **21**: 269-281.
- Zhang, J., X. Zhang, H. Krimb et al. (2003). "Object representation and recognition in shape spaces." Pattern Recognition Letters **36**(5): 1143 – 1154.
- Zurada, Jacek M., Aleksander Malinowski and Ian Cloete (1994). Sensitivity analysis for minimization of input data dimension for feedforward neural network. ISCAS 1994, IEEE International Symposium on Circuits and Systems, London, UK.