D

DATA ANALYSIS WITH NEURAL NETWORK MODELING

D.1 PURPOSE OF THE MODELING

The purpose of data modeling in this study was to quantify the causal relationship between the severity of moisture problems and the hypothesized major contributors to those problems. The modeling developed mathematical relationships to help explain behavior in the sample population of homes and to provide a measure to allow prediction and prevention of moisture problems in the general population of manufactured homes. Because much of the data was based on qualitative assessments in the field, the mathematical models were considered to be of less significance than the correlations and general trends. The correlations and general trends provide an understanding of the strength and general nature of relationships represented by the mathematical model.

D.2 SELECTION OF THE DATA MODELING METHOD

Mathematical modeling is the process of generating an equation or series of equations based on the set of dependent and independent variables that comprise a given data set. The resulting model allows one to predict the behavior of a system by varying the independent variables. The dependent variables or model output, describe the resulting behavior of a system. In the case of this study, the system consists of the dynamic relationships among contributors to moisture problems in manufactured homes; the output is the severity of moisture problems expected. In general, the sample from which the data was drawn consisted of homes experiencing significant moisture problems in hot, humid climates.

Mathematical modeling was undertaken in order to generalize from the individual homes sampled to the larger universe of manufactured homes in hot, humid climates. In theory, the resulting model from homes with significant moisture problems, would allow one to predict the relative significance of any one of the input factors to indicate risk of moisture problems in the home.

A best fit "response function" that is described by the data was created. For this analysis, the response function is an equation with characteristics of the study homes as inputs and degree of moisture problem as output. The equation provided a framework to associate how characteristics of the home affect the degree of observed moisture problems and quantified the relative importance of these characteristics. A correlation coefficient was calculated that describes how well a series of data points "fits" the equation – or how well the model output is explained by the data input.

There are several widely used statistical techniques to model systems; one of the most common is regression analysis. However, when systems become very complex, as in the case of moisture problems in homes, more sophisticated techniques such as neural network analysis are becoming more common.

D.2.1 Regression Analysis

One of the simplest types of modeling methods is regression analysis. Two data points plotted on a set of axes define a straight line. The equation of the line is the best fit "response function" specified

by the two points. If the data is indeed linear, then the straight-line equation provides a predictive model; for each input (data) there is an output (response) in a repeatable pattern. In such a linear regression model, the two-point data series will achieve the maximum correlation coefficient of 1.0, or perfect correlation with the straight-line model. If additional points are added to the data series that do not fall in a straight line, a "best fit" line can be calculated with a correlation coefficient somewhere between 0 and 1. In general, the closer the correlation coefficient is to 1.0, the better the "fit"; or the better the model variance is explained by the data. For simple systems this analysis is straightforward; however, if the output is dependent upon two inputs, the equation is not a line, but a two-dimensional surface. When there are many inputs, the equation is too complex to be illustrated graphically. However, the input and outputs can still be evaluated mathematically through multiple variable regression analysis yielding a mathematical model and a correlation coefficient.

A major drawback in classic regression analysis is that the form of the equation must be assumed in advance. Review of the project data set and the anticipated complexity of moisture problems belied any assumption as to the form of the response function and drove the analysis toward a more sophisticated modeling method known as neural network analysis.

D.2.2 Neural Network Analysis

Neural network analysis, a modeling technique for highly complex systems, has been in use for over twenty years. Today, neural network analysis uses sophisticated algorithms that are appropriate for general applications and problems of considerable complexity. In general, neural network analyses are good pattern recognition and classification tools, with the ability to effectively process imprecise input data. Neural networks offer unique solutions to a variety of classification problems such as speech recognition, highly complex system modeling in which physical processes are not fully understood.

Neural network methods have a strong similarity to models of the biological brain and therefore a great deal of the terminology is borrowed from neuroscience. Like the neurons of the brain, factors that make up a "system" often have multi-faceted, complex relationships with each other, and function together to produce a result. Many such complex systems are difficult to understand using traditional methods.

Like multiple regression analysis, the object of neural network analysis is to find an equation that provides a "best fit" representation of the data. The data set is analyzed in an iterative manner similar to multiple regression by progressively deleting lower significance factors. However, unlike regression analysis, neural network analysis does not require the form of the equation to be known in advance - a tremendous advantage over multiple regression analysis. Neural network analysis uses trial and error to shape an equation to fit data. Once the type of equation is determined, further analysis develops the equation that models the data. The amount of variation in the data explained by the equation is generally higher in neural network modeling than in multiple regression analysis.

D.3 PREPARING THE DATA FOR ANALYSIS

The success of neural network analysis, like other forms of modeling, depends greatly on the sample size and data quality. In general, the larger and more complete the sample, the higher is the confidence in the results. Without a sample of sufficient size or quality, the neural network analysis method will not produce a reliable model. Likewise, samples with incomplete data will also cause problems. Outliers in data sets also make it difficult for the analysis to converge to a model and tend to disproportionately influence the shape of the equation. If included, outliers will skew the model and result in a poor fit of the data. Data from a few samples had obvious outliers; for example: two sample homes had measured duct leakage in excess of 60% - over three times that seen in any other sample home. In such cases, the duct leakage value used for the analysis was capped at the maximum value from the remainder of the set. This maintains the influence of this characteristic in the sample, but prevents it from dominating the model. If outliers and missing data could not be reasonably

amended or approximated, then either the sample home or the subject contributor was removed from the analysis.

The remaining data set was then divided into three separate series depending upon the location of the moisture problems; the wall, ceiling or floor. Preliminary modeling efforts determined that the data series with moisture problems associated only with the floor were not of sufficient size to model. These samples were not included in the analytical portion of the study.

D.3.1 Quantifying the moisture problems

To prepare the data set analysis it was necessary to quantify the extent of the moisture problems in each home. Since one premise of the analysis is that some hypothesized factors are stronger indicators of potential moisture problems than others, the level of input into the model should be related to the severity of the moisture problem. A scoring system was devised to describe the severity of moisture problem. Simple problems such as "odor" were assigned a moisture damage score of one, while more severe problems, such as buckling over 100ft2, were assigned greater values. The highest moisture problem score attained in the sample homes was twelve. These scores were assigned to each problem by the building scientist collecting the data. Note that the scoring methodology is not independent and affects the specific structure of the model developed. However, the scores are proportional to the overall degree of moisture problems in each sample home and are thus reasonable values to use.

Table D-1. Moisture problem scoring system

Moisture Problem	Score	Number of homes surveyed in category
Rust	1	
Odor	1	
Staining <10 square feet	1	
Staining 10 to 100 square feet	2	
Staining >100 square feet	3	
Structural softening <10 square feet	1	
Structural softening 10 to 100 square feet	2	
Structural softening >100 square feet	3	
Bowing and buckling <10 square feet	1	
Bowing and buckling 10 to 100 square feet	2	
Bowing and buckling >100 square feet	3	

Eleven characteristics of the home were selected to help quantify as many as possible of the 12 hypothesized contributors to moisture problems discussed in Section IV (the first column of Table D-2 contains nine of the 11 - wind zone and thermal zone are the other two). Some of the hypothesized contributors were dropped from the statistical analysis because data could not be adequately quantified. For example, although the lack of an exterior air barrier was hypothesized to contribute to moisture problems, it was not possible to non-invasively determine the degree to which the homes possessed this attribute. As such, no determination could be made about the influence of external air barriers on moisture problems. In the case of imbalanced distribution of conditioned air, three metrics were captured.

The impact of certain hypothesized contributors could not be included in the model and so the corresponding factor analyzed is listed in the table as "not included". For example, an interior vapor retarder (in the form of vinyl-covered wallboard) and attic ventilation were present in all homes in the sample since they are stipulated by the HUD-code, and so their impact on the moisture problems could not be modeled.

Table D-2. Hypothesized contributing factors with corresponding metric included in the neural network analysis, if any

Hypothesized Contributor	Corresponding metric analyzed	Explanation
Imbalanced distribution of conditioned air, creating negative pressures relative to the outside	House pressure from closing bedroom door	Pressure differential increase between living area and outside induced by closing master bedroom door when the air handler is on, in UNITS
	Master bedroom pressure	Pressure differential between the master bedroom and the remainder of the home when the air handler is on and the bedroom door is closed, in UNITS
	House pressure	Pressure differential between the house and outside when the air handler is on and all interior doors are open, in UNITS
High rate of shell leakage	Shell leakage	Shell leakage measured by a blower door at 50 Pa, converted to air changes per hour
Ventilated attic space	Not included	Required of all homes surveyed. Not possible to measure integrity of this element.
No ground vapor barrier under house	Ground vapor retarder coverage	Percentage of the home's foot print covered by the ground vapor retarder, visually approximated
Damage to the bottom board	Bottom board integrity	Area of holes in the bottom board, visually approximated in square feet
Duct leakage to the outside	Duct leakage	Duct leakage to the outside, in cubic ft per minute per sf of interior floor area measured at 25 Pa with a duct blaster.
Oversized A/C equipment	Air conditioner capacity	Area of homes served per ton of installed cooling, in sf per ton of cooling capacity
Low indoor thermostat setting	Interior temperature	The lower of the: measured interior temperature, thermostat set point at the time of visit, and set-point reported by the resident, in degrees F
Introduction of unconditioned outside air	Not included	Although data was gathered on the status of the ventilation systems, ventilation function could not be measured in a one-day testing protocol.
Other wall/ceiling penetrations	Not included	Recessed ceiling light fixtures, through-the-wall fans, etc.

Hypothesized Contributor	Corresponding metric analyzed	Explanation
Interior vapor retarder	Not included	Present in all homes surveyed. Not possible to measure integrity of this element.
Lack of exterior air barrier on walls	Not included	Not possible to measure integrity of this element.
Night flushing	Not included	Residents opening windows at night to let in cool air, thereby allowing large amounts of moisture inside.
Localized cold spots	Not included	For example, a cold air supply directed against a nearby section of wall or floor surface and thereby cooling it well below thermostat set point
n/a	Wind zone	Value recorded from data plate
n/a	Thermal zone	Value recorded from data plate

D.4 DEVELOPING DATA RELATIONSHIPS

Models were developed through the neural network analysis using both the data series for moisture problems in the walls and moisture problems in the ceiling. However, the best correlation was found for the combined data series composed of either wall or ceiling moisture problems. Combining these two data sets also increased the sample size for analysis. In this data series the higher of the moisture damage scores for either the wall or the ceiling was selected as the overall moisture problem score for the house.

The model produced by the neural network analysis of data had a correlation coefficient of 0.79. The variation explained by the model, which is expressed as the square of the correlation coefficient is 62.8%. This means that the amount of moisture damage in walls or ceilings that remains to be explained by factors not included in the study was 37.2% (or 100% less 62.8%). Given the inherent variability of some of the data collected, as explained above, a model explaining nearly two-thirds of the variation in the data is considered accurate.

The shell leakage and house pressure factors were found to be of lower importance and were not included in further analysis.

The neural network model developed a response function for the data from which was developed predictive equations that predict the behavior of the system (in this case the impact on the moisture problem score) based on the data input (in this case the hypothesized contributing factors).

The output of the equations is a number (the “moisture impact rating”) whose magnitude indicates the degree to which each factor is associated with moisture problems in homes built to the respective wind and thermal zone standards. The analysis alone does not establish a direct cause and effect relationship between the hypothesized contributing factor and the moisture problem, however, observation and professional experience indicate there is a strong likelihood of such a relationship existing.

The neural network analysis developed predictive equations for each combination of three wind zones and two thermal zones, a total of 6 categories (the sample homes were all located in HUD Thermal Zone I, however, 30% of the homes were constructed to Zone II standards). Due to the differing

construction characteristics1 of homes built for each thermal and wind zone2 category, it was not possible to combine them into a single predictive equation.

In dividing the sample into these 6 categories, the sample sizes became quite small in some cases. The relative ranking of each hypothesized factor was generally consistent, however, and since the intent of this analysis is to understand, in general terms, the relative importance of the various factors, a combined ranking was created and is shown in Table D-3.

Table D-3. Moisture impact rating averaged for each hypothesized contributing factor

Factor	Setting that will result in lowest probability of high moisture problem score	Ranking	Range of factors
Master bedroom pressure	lower	1	0 – 26 pascals
Air conditioner sizing	higher	2	3.8 – 1.6 tons per 1,000 sf floor area
Interior temperature	higher	3	65-80 degrees F
Ground liner coverage	higher	4	0 – 100% coverage
House pressure from closing bedroom door	higher	5	-8.6 - +1 pascals
Duct leakage	lower	6	1 – 25 cfm / sf floor area
Bottom board holes	lower	7	0 – 200 sf
Shell leakage	Little effect	n/a	n/a
House pressure	Little effect	n/a	n/a

The “Range of factors is that for the whole data set. The actual range used in the analysis should be found in the set of homes analyzed

D.5 INTERPRETING THE RESULTS AND LIMITS OF THE ANALYSIS

The validity of the analysis is dependent on the quality and quantity of the data. The quantity was limited by the difficulty in finding affected homes in the summer of 2001. Even with the sample size limited, the amount of moisture damage in walls or ceilings that can be explained by the model is a respectable 62%.

The quality of the data was affected by the difficulty in getting accurate, consistent, and objective measurements. Neural network analysis handles imprecise data quite well and thus, the relative magnitude of the average moisture contributing factors is considered reasonably accurate. In other words, for the hypothesized contributing factors that could be modeled, the prioritization in Table 4 is thought to generally represent the degree of association with moisture problems in walls and ceilings in hot, humid climates. A/C sizing, master bedroom pressure and interior temperature are thought to

[1] Thermal zone II homes will often have more insulation. Homes designed for higher wind zones may have added sheathing on exterior walls.
[2] 28% of the homes were built to wind zone 1 standards, 58% wind zone 2, and 6% wind zone 3.

be highly important; ground liner coverage and house pressure from closing the bedroom door are thought to be of medium significance; duct leakage and bottom board holes are thought to be of lower importance; and shell leakage and house pressure are thought to be of little significance. The analysis makes no claims to the impact, or lack thereof, of other potential factors, such as interior vapor barrier and attic venting, which were not included in the neural network analysis.