Modeling Plant Succession with Markov Matrices

Concluding Paper

Undergraduate Biology and Math Training Program

New Jersey Institute of Technology

Catherine Morrison

November 3, 2008

Partner: David Hamoui

Advisors: Prof. Gareth Russell

Prof. Amitaba Bose

Abstract:

Plant succession is the process by which plants die and their place is taken by other plants that grow in the same location. This process repeats over time causing, for example, the growth of forests from abandoned fields. The replacement of plants follows a mostly predictable sequence. This paper is about recent research into the possibility of modeling plant succession by the Markov process. It is part of a larger project the goal of which is exploring the possibility of manipulation in the successional process. In other words, whether by small changes the result of succession can be altered to achieve a more desirable state.

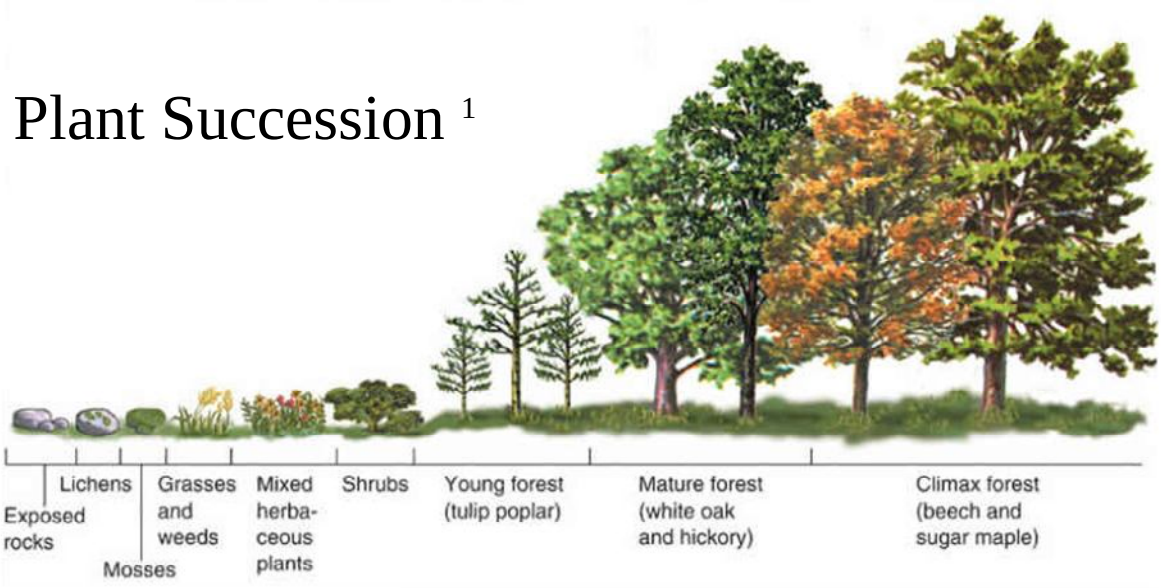
Our research was focused on building the tools necessary for the larger experiment. We first explored the sources of stochasticity inherent in the use of the Markov process for modeling succession. Stochasticity comes from individual species when they are modeled instead of proportions of species. Another kind of stochasticity comes from the use of a set matrix to better approximate the uncertainty inherent in the Markov model. From this analysis we created a method for the determination of the appropriate number of plots to be used based on their size. Secondly, we explored methods to determine how many previous time steps should be used to determine the appropriate Markov matrix for predicting a certain time step and noted that this number will vary based on the homogeneity of the Markov process. Throughout, we use real world examples to demonstrate our theories and the usefulness of our tools.

Table of Contents

List of Tables

List of Figures

Introduction



Plant succession is the replacement of plants over time steps. This process, for example, allows an abandoned field to become a forest. However, though this diagram shows the general plant groups that replace their predecessors as time progresses, the specific plant sequence or sequence options with their probabilities are highly variable from plot to plot depending on many external as well as innate properties.

Figure 1.1

When warehouses are leveled, landfills are capped or fields are abandoned, a similar process occurs. This process is called plant succession. Originally small weeds and grasses grow. Over time larger, stronger and longer lasting plants tend to take their place (Figure 1.1). This process can be studied over time and with enough information the transitions might be able to be predicted with a degree of certainty. This could be a very important step towards making these environments friendlier to native species or at least less hostile. The goal of our research was to investigate the possibility of using mathematical systems and computer modeling to predict the transitions and to suggest possible small changes that could be made in the plot in order to reach a more desirable

outcome. From prior research, we hypothesize that the outcome of plant succession can be predicted with sufficient precision to enable identification of small interventions that will push the community towards a more desirable end state. A more desirable end state would be one in which the habitat would support more native species, improve the appearance of the area or become more easily manageable.

Studying plant succession is not a new field. Historically, however, scientists have focused more on learning the general transitions that are likely to be made and tracking these transitions through laborious field work. In this experiment, no field work was engaged in although some of the data the methods were tested on was taken from actual field work. Other data was obtained from studies unrelated to ecology that exemplified transitions over time. Much of the data was taken from the PhD dissertation of an ecology student, Corey Samuels, from the University of Tennessee.[2] Her data was from grasses in Kansas. They were grouped into annuals and perennials. In her paper, Samuels suggests that Markov set chains could be a good model for the process of plant succession. Her idea was the basis of this research.

Markov matrices are a method of obtaining data about a time step from its predecessor. Given a matrix of probabilities (the Markov matrix) and a vector of numbers representing the population (either by percents or actual numbers), the vector of the population in the next time step can be found by the multiplication of the matrix and the vector. In each case, only the population information from the last time step is

needed to generate the next step along with the transition matrix. Figure 1.2 shows an example of a vector of population percentiles being transformed.

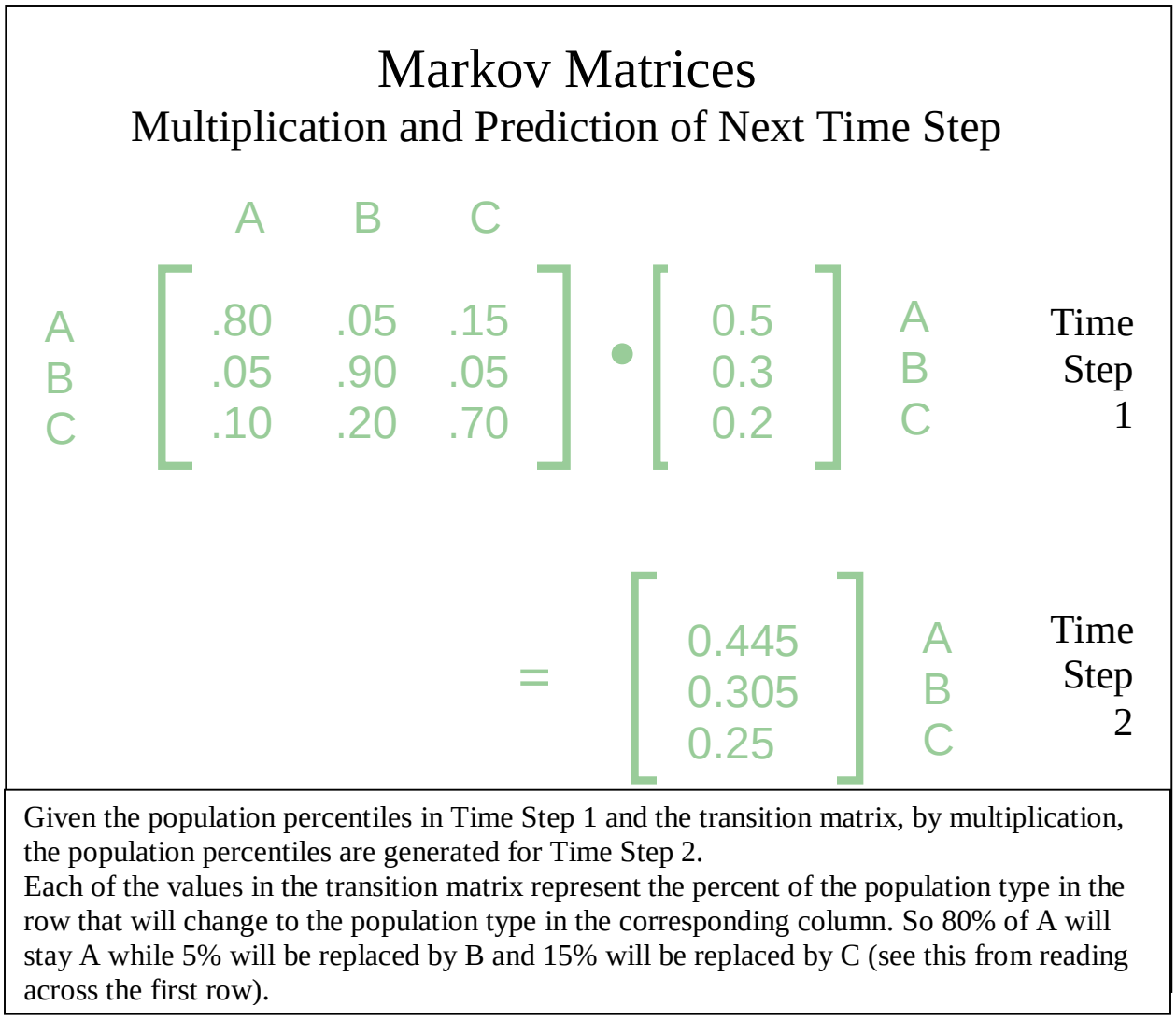
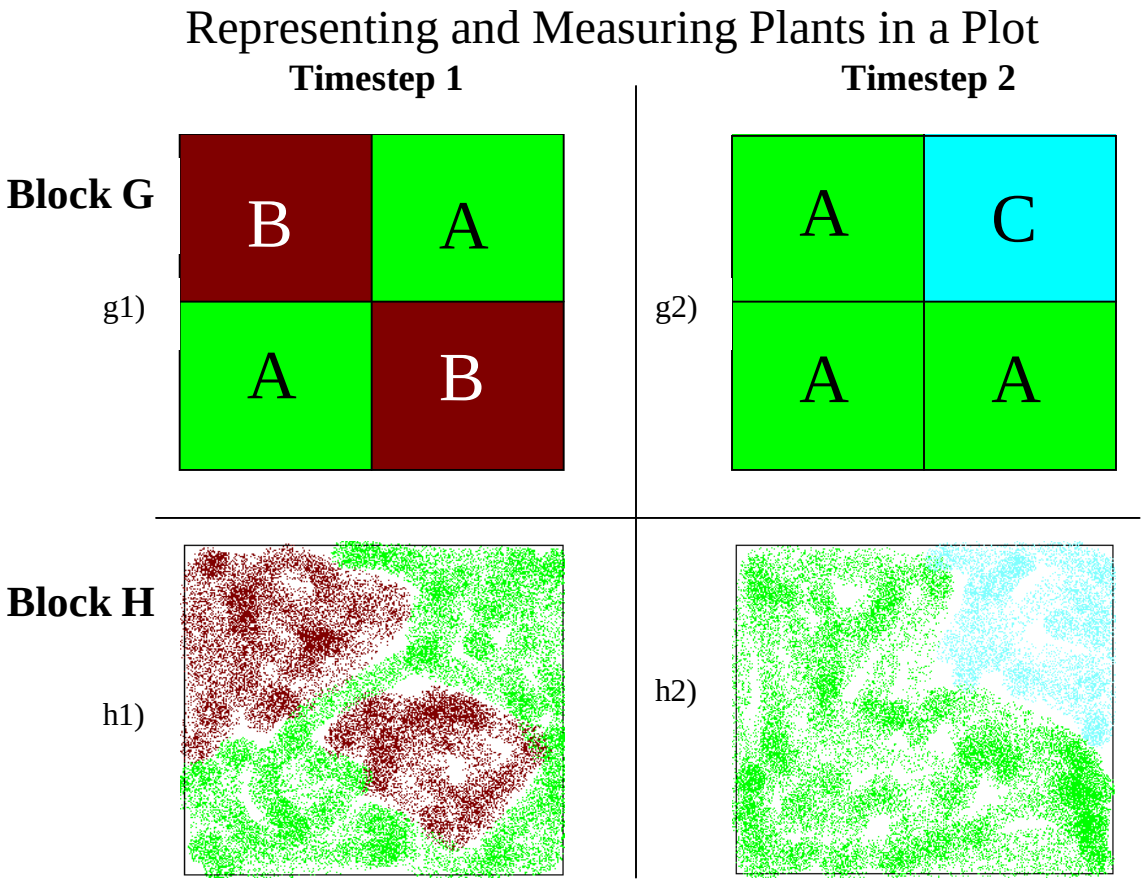


Figure 1.2

If the matrix of transitions is viewed as a grid, for each of the squares, the value inside represents the percentage of the row population type that will be replaced by the column population type. When this is applied to plants, the same principles hold. Only A,B and C represent types of plants. The process of replacement is a discrete process.



Block G could be interpreted as a plot with four plants in it (A,B and C). Block H is a plot with a large number of plants in it. Since it is nearly impossible to keep track of every plant or make sure every plant is in one piece of the grid, it is more likely that Block G is a representation of Block H. Boxes h1 and h2 represent what a data collector would observe in the field. He would estimate what percentage of plants are A, B or C and may represent this information by the grids in Block G.

Both blocks show plant succession. Block G makes interpretation easier. From Block G we see that all plants of type B were succeeded by plants of type A while fifty percent of plants of type A were succeeded by plants of type C.

Figure 1.3

The plot has some set of species. During one of the time steps, one or more of these species will be replaced by another. The state of the system at any point in time is a function of the state at some previous time step †. This process is encoded in the matrix of transition probabilities. Figure 1.3 shows what the process looks like ecologically. Block G could be interpreted as having two plants of type B and two plants of type A in box g1, and three plants of type A and one plant of type C in box g2. These would be the numbers placed in the population vector. However Block G could be a generalization of Block H which would mean that the proper interpretation would be that box g1 was 50% B and 50% A while box g2 was 75% A and 25% C. The vector would now be made up of proportions and the data would be continuous.

Given only one time step to gather the data from, the transition values appear to be:

A→ A: 50%	A→ B: 0%	A→ C: 50%	(Table 1.1)
B→ A: 100%	B→ B: 0%	B→ C: 0%	

There is not enough information to determine the transitions from C since no transition from C is shown in the data. Further data might reveal slightly different patterns of transition but if all the values were based on the first transition (between Time step 1 and Time step 2) the values would be those in Table 1.1.

---

† Most often the time step used is the previous time step. However, as explained on page 21, more than one time step or a time step that not immediately preceding the one to be predicted can be used under certain circumstances.

Foundational Functions

The first step in using Markov Matrices to model plant succession is developing functions that calculate these matrices. First a method of least squares fitting was applied to the data. The input for least squares is a set of two matrices. With the data in table 2.2 (one of Samuels data sets) the matrices in Figure 2.1a are formed.

{1994, 0.692, 0.308},
{1995, 0.315, 0.685},
{1996, 0.226, 0.774},
{1997, 0.075, 0.925},
{1998, 0.020, 0.980},
{1999, 0.070, 0.930}

(Table 2.1)

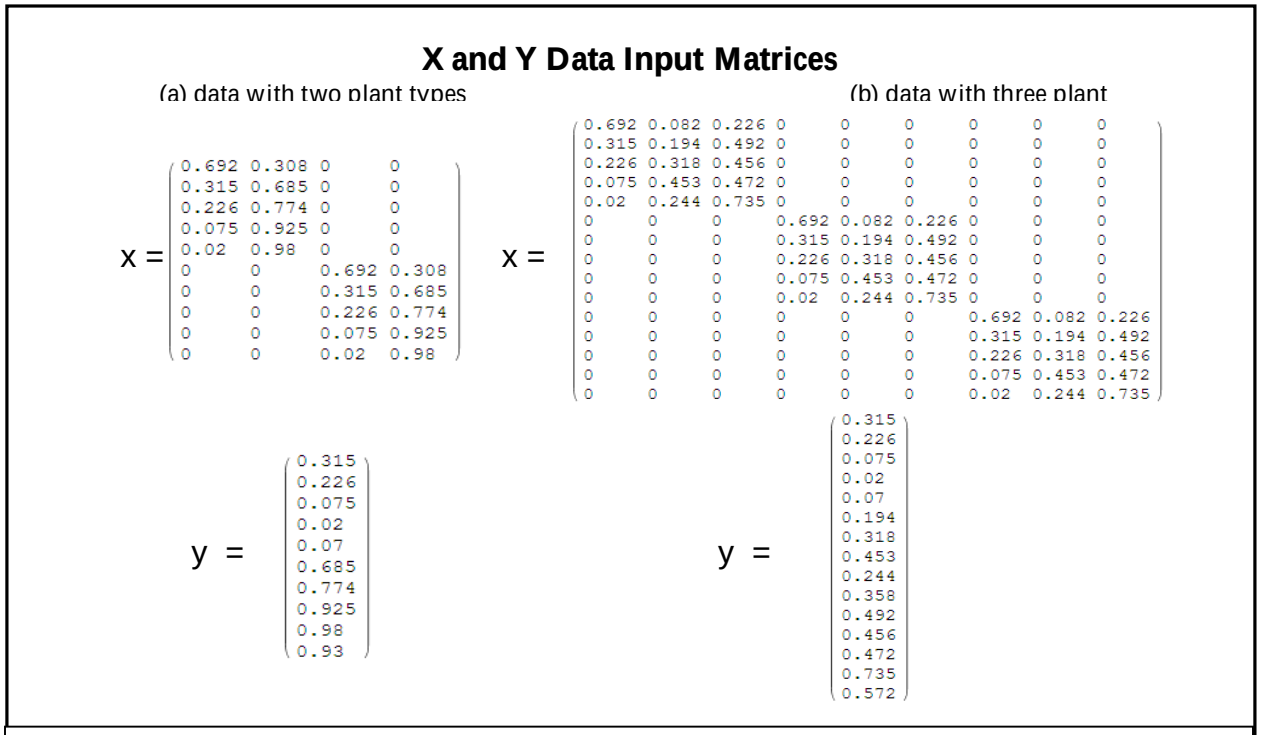
These data are in the form {date, type1, type2}. Some data are also in the form

{1994, 0.692, 0.082, 0.226},
{1995, 0.315, 0.194, 0.492},
{1996, 0.226, 0.318, 0.456},
{1997, 0.075, 0.453, 0.472},
{1998, 0.020, 0.244, 0.735},
{1999, 0.070, 0.358, 0.572}

(Table 2.2)

where the third column in the data of table 2.2 are broken into two groups. Essentially, the plot was differentiated into two types of plants in Table 2.2 and the second of these types was further differentiated in table 2.3 to form three types. This becomes clear when the last two rows in table 2.3 are summed because they equal the last row in table 2.2. Ecologically, the plot is made up, for example, of shrubs and grasses. This characterization is useful on one level. However, on another level, the grass can be split

up into two types such as Kentucky Bluegrass and Annual Bluegrass. The groups that are placed in the matrix can have many different degrees of specificity which could be useful if there are different degrees of certainty associated with the degrees of specificity.



The two matrices in (a) for two plant types and(b) for three plant types are examples of the matrices necessary for input for both the least squares and quadratic programming methods.

Figure 2.1

Using the least squares method , it was possible to generate negative values for the probabilities which are nonsensical. Instead, a method of quadratic programming, which is built into Mathematica, as it is in many other programs, was employed. This method takes the same input and does not generate any negative values (constraint: $0 \leq n \leq 1$). It is a mathematical optimization technique for several variables subject to linear

constraints. This code now became the foundation of this research because it generates the Markov matrix.

A second useful tool is a simulator which is a piece of code that produces matrices based on inputs and in some cases random generators. These matrices can then be fed to a Markov matrix simulator which generates a sequence of proportions (the Markov matrix). This tool is very useful when data are needed to test a new program but is not available. It is also helpful in testing different kinds of variability, starting vectors and starting matrices. In this simulator there are four different options for stochasticity that can be used (see Figure 2.3). One of the options is no stochasticity (deterministic); the other three options involve different forms of stochasticity.

In the deterministic option, a vector of proportions and a matrix are given. The simulator runs the Markov process and since the proportions and matrix are fixed, the process will converge to a vector of proportions. Using the Eigen vector these final proportions can be determined.

The first option for stochasticity comes in the form of a set matrix and a fixed vector of proportions. In a set matrix there are not single values but sets of values. These are an attempt to better represent the uncertainty inherent in the system and are generated by the High Low Method, a method developed in Samuels' dissertation which solves the problem generated by a set matrix- the rows do not sum to one. The simulator randomly picks a value in the range of each of the sets. Figure 2.2 contains further explanation and examples of the set matrix.

The second option for stochasticity uses a fixed matrix as in the deterministic option but does not use a vector of proportions but a vector of individuals. A vector of individuals gives stochasticity because though a vector of proportions such as

{0.5, 0.5} (Table 2.3)

and {1,1} (Table 2.4)

may seem identical, their result can be very different. If the vector for the first time step

was {a, b} = {1,1} (Table 2.5)

the next time step vector could be {1,1}, {0,2} or {2,0} (Table 2.6)

with the relative probabilities {0.5,0.25,0.25}. (Table 2.7)

This means that fifty percent of the time, one of the species would die out.

These are much bigger changes than if given a vector

{a, b}={100,100} (Table 2.8)

where the result is far more likely to be very close to itself- such as

{94, 106} (Table 2.9)

rather than {0,200} (Table 2.10)

Essentially, as the number of individuals increases, the vector predicted for a certain time step approaches the vector prediction for proportions. If the number of individuals is small, there is a much greater degree of variance. Using individuals generates a second type of stochasticity.

The third option combines the first two with a set matrix and a vector of individuals, increasing the stochasticity.

Simple Matrices vs. Set Matrices

$$\begin{bmatrix} 0.25 & 0.75 \\ 0.5 & 0.5 \end{bmatrix}$$

$$\begin{bmatrix} \{0.1\ ,\ 0.35\} & \{0.65\ ,\ 0.9\} \\ \{0.35\ ,\ 0.55\} & \{0.45\ ,\ 0.65\} \end{bmatrix}$$

Since it is nearly impossible that a transition from one plant type to another would have a set probability with no variance, it is much more realistic to produce a Markov where each transition is represented by a set of numbers which are the minimum and maximum percentage of plants that underwent that transition. For example, instead of giving constant percentages such as in Table 2.4a it is much more likely that the percentages will lie in ranges as in Table 2.4b

Figure 2.2

Simulator

Options, Components and Results

Matrix Type	Input Type	Proportions	Individuals
		$\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$	$\begin{bmatrix} 15 \\ 15 \end{bmatrix}$
Markov Matrices $\begin{bmatrix} 0.25 & 0.75 \\ 0.5 & 0.5 \end{bmatrix}$		Deterministic Eigen Vector	Stochastic Type 2 Fluctuating
Markov Set Matrices $\begin{bmatrix} \{0.1 , 0.35\} & \{0.65 , 0.9\} \\ \{0.35 , 0.55\} & \{0.45 , 0.65\} \end{bmatrix}$		Stochastic Type 1 Fluctuating	Stochastic Type 1 and 2 Fluctuating

The simple matrix has no inherent stochasticity and neither does the vector of proportions. Both the vector of individuals and the set matrix have inherent stochasticity. With these four options four types of stochasticity can be obtained. These may converge (deterministic) or fluctuate (stochastic). The fluctuation will be asymptotic but not deterministic.

Figure 2.3

Number of Plots and Plot Size

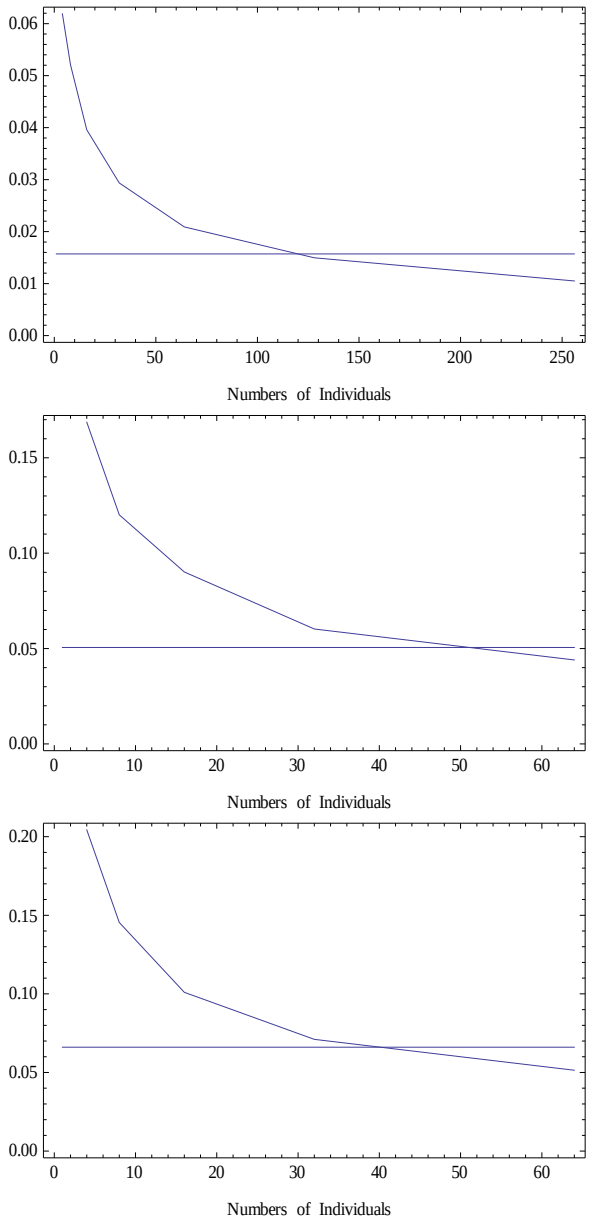
One of the first questions to be answered is how many plots need to be tracked and how big these plots should be in order to get accurate data by balancing the stochasticity from these two sources. A measurement called Bray Curtis Distance[4] was employed to measure the effect of noise in the process on the final community.

Bray Curtis Distance is a measure of the difference between two communities. The limits of this measurement are between 0 and 1. Given vectors {a,b,c} and {x,y,z}, the equation is:

$$\text{Bray Curtis Distance} = \frac{\text{Abs}[a-x] + \text{Abs}[b-y] + \text{Abs}[c-z]}{\text{Abs}[a+x] + \text{Abs}[b+y] + \text{Abs}[c+z]} \qquad \text{(Table 2.11)}$$

A zero would result when the proportions (vectors) are identical and one would be obtained when the for example, a equaled 1 and x equaled 0. This was used to determine the noise generated by plots of different sizes where the size of a plot was "measured" by the number of individuals it contained. The results of the Bray Curtis Distances were graphed when the number of individuals was varied and the matrix was kept as a constant. As expected, the more individuals are introduced, the less effect the noise has on the results. The other line on the graph was generated with four plots by assuming an infinite population of the individuals with a set matrix where the values of the matrix were picked randomly every time. Where these two plots intersect is the place where the noise generated by these two sources of stochasticity (number of individuals and randomly picked values from a set to fill the matrix) is equal. This point gives us the optimal number of individuals per plot when there are four plots.

Graphs of Noise Based on Population Variation vs. Noise Based on Matrix Variation



The intersection point of these graphs represents the number of individuals per plot where the level of noise generated by the number of plots is equal to the level of noise generated by the number of individuals. This is the best number of individuals given the number of plots. [Note: the y value on these graphs in the Bray Curtis Distance]

Figure 3.1

Optimal Number of Previous Time Steps

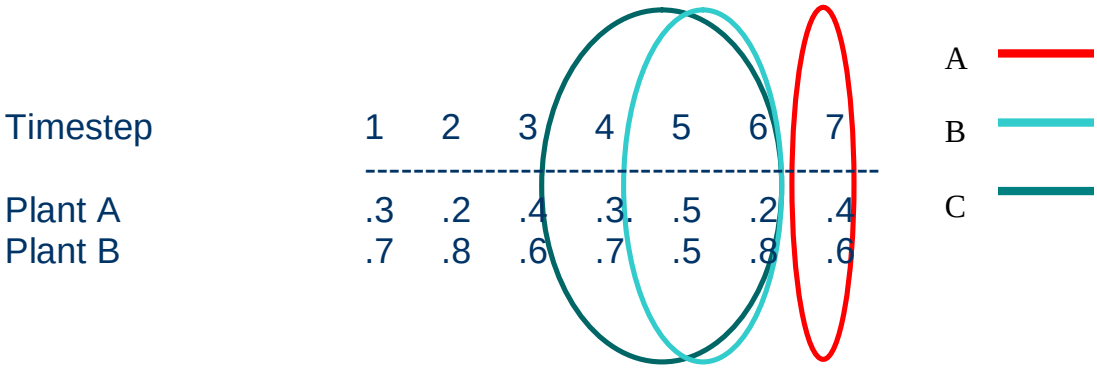
Now that the method of generating matrices has been developed, it is important to design the experiment. Since every matrix needs at least one previous data point to estimate the current one, the next question is how many data points are best? Is there a predictable result achieved by using more previous time steps? These questions are important because in order to have the least error in predictions, since these matrices depend on the previous time steps it is important to know that, for example, with 100 previous time steps available the last 50 will generate the answer with the smallest error (Bray Curtis Distance) (Figure 4.1).

It may seem that the more data points the better the result will always be, however this will not always be the case. If the data is homogenous it should level out, making for a better approximation the more time steps are used. Also an increase in time steps will decrease the negative affect that fluctuations due to bad readings, unexpected changes, etc. will have on the result. But, if the data is non-homogenous, there will be a point in time where the previous data is irrelevant which will make the results worse. In the long run, every data set will be significantly different so the probability that through some program the number of previous time steps to be utilized could be magically estimated is very small.

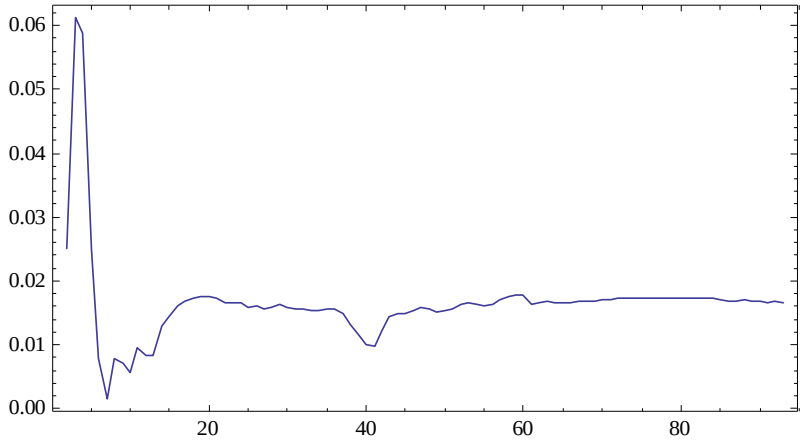
Figure 5.2 shows another way to visualize these types of results and some conclusions that can be drawn using the other method of visualization.

Predictions:
Using the Optimal Number of Previous Time Steps

(a) Previous Time Steps:
Which is better at predicting A? B or C?



(b) Data Showing Accuracy of Prediction Based on Previous Time Steps



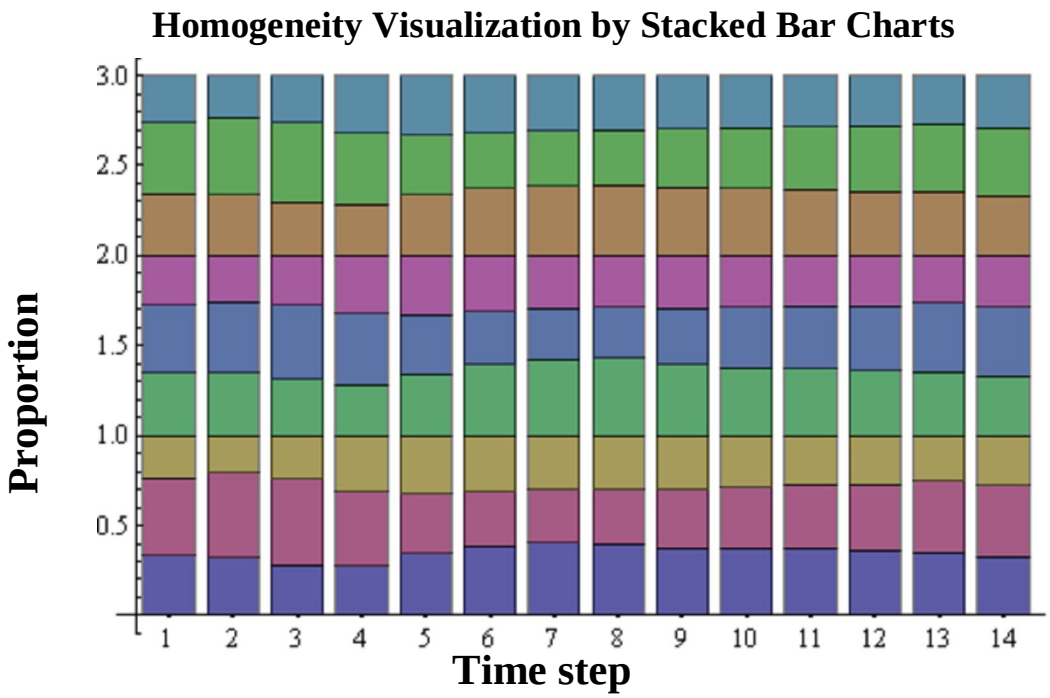
x axis= number of previous time steps used
y axis= Bray Curtis Distance (0 is most accurate, 1 is most inaccurate)

In (a) the theory is illustrated with random numbers. If A is to be predicted, is it better to use the data enclosed in B or C (or any other set of data for that matter)?
In (b) one of the graphs generated in Mathematica by graphing the Bray Curtis Distance and the number of previous time steps used is shown. The data was closest around 6 or 7 previous time steps. In the area before that, the noise seems to be too great to obtain reasonable results. In the area around 20 and after, the noise seems to settle to a basically constant level.

Figure 4.1

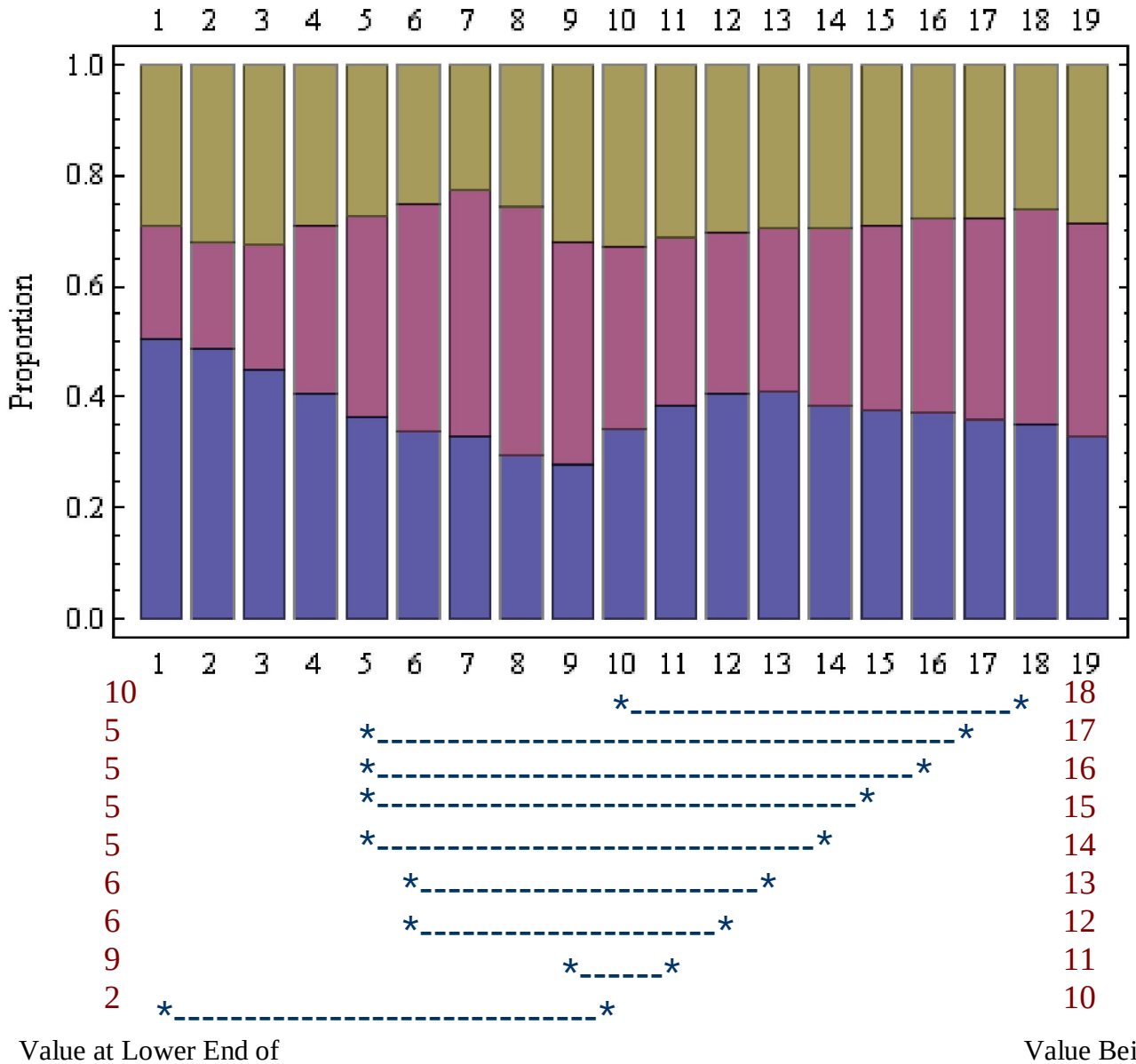
Visualizations of Homogeneity of Data

One way that homogeneity can be visualized is the amount of difference over time between data points. One easy way to visualize this is stacked bar charts. In Figure 5.1 one of the bar charts used in this experiment is shown. The data appears homogenous with a small amount of noise. These graphs can be a useful visualization. Figure 5.2 shows a way this type of graph can be useful to visualize optimal previous time steps.



This graph attempts to visualize the relative "sameness" of points with respect to one another by graphing the values of each of the types over time (represented by the bars). The sum of all the data values at all times is equal to 3 and the width of each color in a bar represents the amount of data made up by that type of plant. Through this, the relative fluctuations in the amount of a certain type of plant can be visualized by the width of its color band.

Figure 5.1



This figure is an attempt to show any possible correlation between the optimal number of previous time steps and the percentage of the population made up by each of the three types (gold, red, blue). The left star shows the value at the lower end of the optimal number of previous time steps, while the star on the right shows the time step being predicted. The figure shows that predicting time steps from 12 to 17 is best accomplished by using previous time steps with a lower value in the range of 5 or 6. Other time steps are less uniform. Looking at the chart above the lines, the values of the red, blue and gold are mostly uniform from 12 to 17 which agrees with the uniformity of their best predicting values.

Figure 5.2

Conclusion

The goals of this research were to explore the stochasticity inherent in the modeling system in order to determine the size and number of plots and to develop a method to find the optimum number of previous time steps. Before we could do this we needed to replicate Samuel's methods in Mathematica and test them on her data to determine if our replications were accurate. The results of these tests proved they were.

Moving on to our goals we found that Stochasticity was inherent in the use of individuals instead of proportions (stochasticity type 1) and the different Markov matrices estimated from different plots which were represented by set matrices (stochasticity type 2). This knowledge of the sources of stochasticity resulted in a method for determining the best number of individuals in a plot and number of plots.

For the second part of our research, we explored the homogeneity of Markov matrices in real data sets and experimented with data sets to see if we could develop a method of determining how many time steps should be included in the calculations of the Markov matrix calculated for the prediction of a certain time step. The method we developed enabled us to determine the best place to begin using previous time steps from in order to have a Markov matrix that gave the best prediction. Since in homogenous data, this time step should be the same in all cases, when the calculations revealed different starting points for a different time step to be predicted, it became a strong reason to believe that the data was non homogenous.

Further Research

Now that the methods are generated, the next step is working with some new larger sets of ecological data in order to test and fine tune the model as well as discover new pieces of the design. In the next steps of the larger experiment, field testing will begin to gather more information about successional steps and how small changes in composition, environment, etc. affect those steps.

We also hope to explore our results using standard deviation and standard error. We would also like to change the number of plots tested in the intersection of stochasticities graph to see how the change affects the intersection point.

References

[1] Cuny Geography Program. Last access date: 11/01/08 http://www.geography.hunter.cuny.edu/~tbw/ncc/chap4.wc/vegetation/plant.succession.jpg

[2] Samuels, Corey. "Markov Set-chains as Models of Plant Succession." May 2001.

[3] Lee, T.C., Judge, G.G. & Zellner, A. (1977). *Estimating the parameters of the Markov probability model from aggregate time series data*. New York: North-Holland Publishing Company.

[4] Wolfram Mathematica Reference: Bray Curtis Distance Article. Last Access Date 12/10/08. http://reference.wolfram.com/mathematica/ref/BrayCurtisDistance.html