CAVES Generalization Framework

March 5, 2008

1 The Generalization Framework

Diversity of issues and modeling paradigms do not allow for a direct generalization of the models built for CAVES case studies. Instead, we propose to investigate abstract representations (concepts) of certain aspects of CAVES case studies models. Then we can introduce measures related to these representations and compare the same set measures over all case studies models.

We assume that agents' behavior is constructed based on her/his perception of variety of influences, which can be classified as individual, interpersonal and environmental factors (see Figure 1).

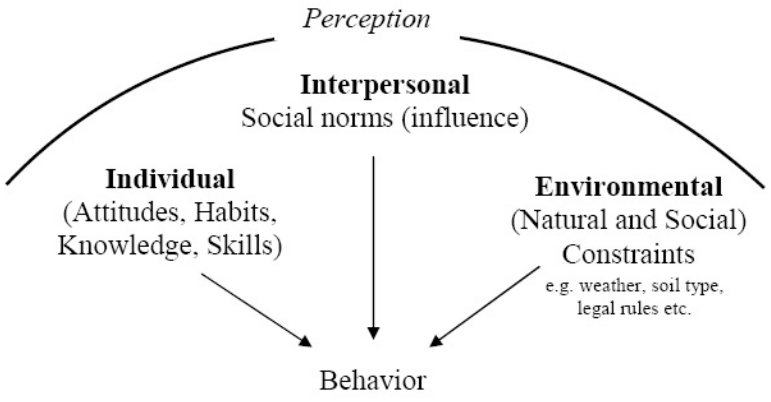

Figure 1: Modeling agents' behavior

Based on the influence categories introduced in Figure 2 we can conceptualize a set of agents linked in social networks interacting with other agents and their environment. Certain features of the detailed case studies models can be reflected in the following abstract representations (see Figure 2):

- agents
  Agents constitute nodes in social networks. Their states can be represented through the set of variables and parameters.

- agents' choices (decisions, strategies)
  In this framework we focus only on choice decisions which often can be a choice of the specific strategy. The choice of an agent can be influenced by his own state, the state of other agents linked with him in a social network and environment. Agents decisions can lead to modifications of his social network and/or environment.

- links between agents
  Agents can be linked with other agents with one or more connections. Different types of connections can be named and conceptualized as network layers. Different layers can be used in different decisions.

- groups - agents affiliations
  Agents can belong to one or more groups which can influence the shape of social network(s).

- environment
  Environment can be represented through a set of environmental variables and parameters, some of them can be spatially explicit.

- system's interactions
  Two-way interactions between agents, network and environment define the system dynamics. However in some models only some interaction types can be present. For example in some models a network can be static so it is not influenced by agents' choices. The interactions listed in Figure 2 provide a full set of possibilities in the generalization framework.

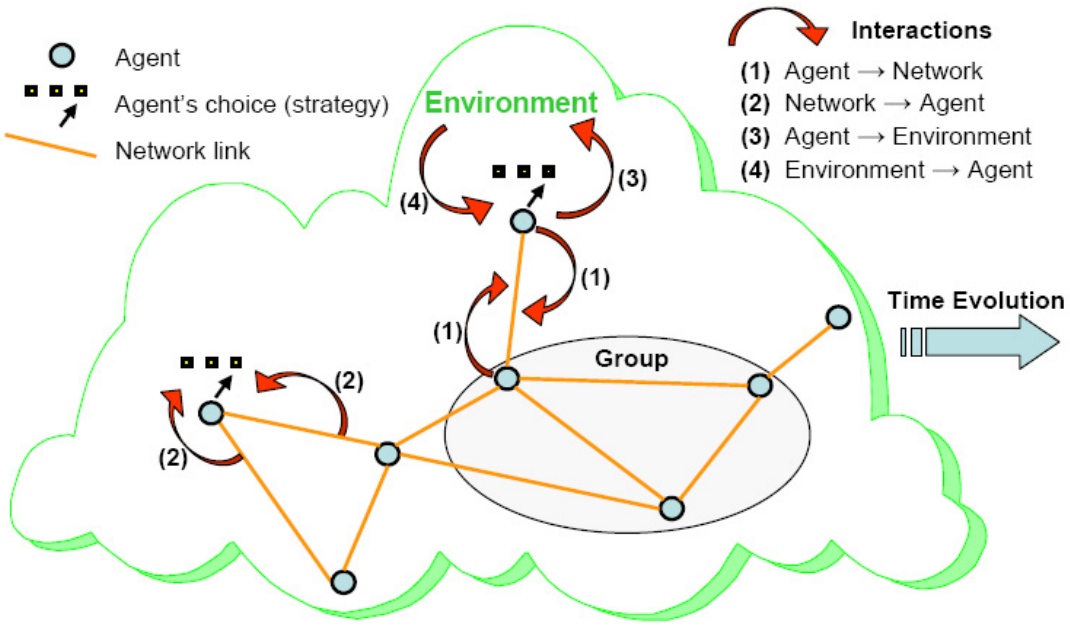

Figure 2: CAVES generalization framework

2 Complexity Measures - Introduction

The basic distinction in complexity measures as applied to the above framework, lies in the focus of analysis - it can be network based or agent-based. As a modeled system evolves we can trace how the measures change in time.

Network analysis has a long history and many of the developed network metrics such as average path length or degree centrality can be used successfully in the CAVES project. Recent advances in complex networks increase the repertoire of measures - we recommend to include clustering coefficient and degree correlation.

Agents-focused complexity measures use agents' states as substrate for calculations. Example measures: dynamism, polarization or clustering as applied to agents' choices.

Both network-based and agent-based measures can be analyzed as time series - this type of analysis can focus on features such as volatility clustering or other examples of heteroscedasticity.

Looking into a stability (or instability) of time behavior of selected variables or measures we can apply resilience measures. This kind of analysis focuses on phase space looking for alternative stability domains and how they change in time. It should be mentioned that certain level of volatility can make this type of analysis inapplicable.

3 Network Measures

3.1 Introduction – notation and supplementary algorithms

Notation

- V – set of nodes
- E – set of edges, e_{jk} – an edge linking nodes j and k for undirected graph or an edge pointing from j to k for a directed graph
- W – adjacency matrix; $N \times N$ binary matrix such that $W(i,j) = 1$ if there is a link between nodes i and j and $W(i,j) = 0$ if there is no link (in particular $W(i,i) = 0$)
- N – total number of nodes
- M – total number of links
- $d(i)$ – degree of the node i, for directed graph we distinguish $d^{in}(i)$ – number of incoming edges and $d^{out}(i)$ – number of outcoming edges
- $\overline{d}$ – average node degree
- $g(i,j)$ – distance (geodesic distance) between nodes i and j; length of the shortest path between nodes i and j
- connected graph – each node is reachable from any other node
- geodesic – shortest path between two nodes

3.1.1 Geodesic distance (shortest path length)

The path length algorithm (Dijkstra algorithm) returns the N - elements vector `distance` of geodesic (shortest) distances between the node s and all the other vertices. An input to Dijkstra algorithm is the matrix $A(i,j)$ such that

- $A(i,i) = 0$,
- $A(i,j) = 1$ if i and j are linked (for directed graph: if there is a link pointing from i to j, i.e. $e_{ij} \in E$)
- $A(i,j) =$ inf if i and j are not linked (for directed graph: if $e_{ij} \notin E$)

and a node s from which we compute the geodesic distances. The algorithm can be generalized to when the direct connections between the nodes are expressed as connection "distances" (inversely proportional to "strength" of the connection), rather than *there is/there is not* a link.
Path length $g(i,j)$ (where $g(i,\ldots) =$`distance`(i)) is further used to determine the network measures such as *diameter* and *average path length*.

```
%create an $N$-element zero vector
visited(1:N) = 0;
%assign all vertices,except s, the initial distance from s equal to infinity,
%set distance(s,s)=0
distance(1:N) = inf;
distance(s) = 0;
% The main loop does not repeat the contents exactly.
% In each run the vector no_visited is updated.
for i = 1:(N-1)

    %read no visited node
    %no_visited = [];
    for j = 1:N
         if visited(j)==0
             no_visited(j) = distance(j);
         else
             no_visited(j) = inf;
         end
     end;

     %selection of min
     %x - minimal value of no_visited, j_min - the position of x;
     %if there are more than one minimal element,
     %the index of the first one is returned.
     [x,j_min] = min(no_visited)

     %mark visited node
```

```
    visited(j_min) = 1;

    if(x==inf)
        break
    end;

    %expand and update last visited node
    for v = 1:n,
        if ( ( A(j_min, v) + distance(j_min)) < distance(v) )
            distance(v) = distance(j_min) + A(j_min, v)
        end;
    end;
end;

return
```

3.2 Transitivity (clustering) [3]

A deviation from the behavior of random graph can be seen in the property of network transitivity (clustering). Some networks exhibit significantly higher density of triangles in the network: if vertex A is connected to vertex B and vertex B is connected to vertex C, then there is some higher probability that A is also connected to C ("the friend of your friend is likely also to be your friend").
Clustering coefficient is defined as:

$$C^{(1)} = \frac{3 \times \text{ number of triangles in the network}}{\text{number of connected triples of vertices}} \tag{1}$$

where "'connected triple"' is a single vertex with edges running to an unordered pair of other vertices. An alternative definition of the clustering coefficient (Watts and Strogatz) is given as

$$C^{(2)} = \frac{1}{N}\sum_i C_i \tag{2}$$

where

$$C_i = \begin{cases} \dfrac{\text{number of triangles connected to vertex } i}{\text{number of triples centered in vertex } i}, & d(i) > 1 \\ 0, & d(i) \leq 1 \end{cases} \tag{3}$$

is referred to as *local clustering.*

- undirected graph

$$C_i = \frac{\#\{e_{jk} \in E : j < k, j, k \in N(i)\}}{\frac{1}{2}\sum_{i=1}^{N} d(i)(d(i)-1)}$$

- directed graph

$$C_i = \frac{\#\{e_{jk} \in E : j, k \in N(i)\}}{\sum_{i=1}^{N} d(i)(d(i)-1)}$$

here $d(i)$ is the total (in + out) degree of the vertex $d(i) = d^{in}(i) + d^{out}(i)$ and $N(i) = \{j : e_{ij} \vee e_{ji} \in E\}$

The two definitions (1) and (2) differ in the order of the operations of taking the ratio of triangles/triples and averaging over all vertices. $C^{(2)}$ tends to assign higher weights to the low-degree vertices than $C^{(1)}$. As far as calculation, $C^{(2)}$ is easily calculated on computer whereas $C^{(1)}$ is more convenient for analytical calculations.

3.3 Assortativity

The network is called assortative if it exhibits positive degree-degree correlation. To quantify this effect degree correlation coefficient r is introduced as a normalized connected degree-degree correlation function $\langle jk \rangle - \langle j \rangle \langle k \rangle$, where $\langle \ldots \rangle$ expresses an average over the edges [4]. The normalization assures that r takes values within the range$[-1, +1]$. For a given network degree correlation coefficient r is calculated as follows [4]

$$r = \frac{M^{-1} \sum_i j_i k_i - [M^{-1} \sum_i (j_i + k_i)]^2}{M^{-1} \sum_i \frac{1}{2}(j_i^2 + k_i^2) - \left[M^{-1} \sum_i \frac{1}{2}(j_i + k_i)\right]^2}, \tag{4}$$

where j_i and k_i are the degrees of nodes at the end of ith edge, the sums are taken over all M edges.

Assortativity coefficient r is computable **both** for directed and undirected graphs. For directed graphs M is the total number of edges (in and out) and so are the degrees k_i, j_i – both *in* and *out* summed up.

3.4 Average path length

Average path length is well defined for connected graphs

$$l_G = \frac{1}{N(N-1)} \sum_{i,j \in V} g(i,j) \tag{5}$$

where $g(i,j)$ is the geodesic distance from i to j. In order to generalize l_G to the case of not connected graph one may assign some high (but finite) value to the distance of nodes that are not reachable from each other $\tilde{g}(i,j) = D$, if $g(i,j) = \infty$. (D is a parameter[1]). If $D = 0$ than l_G is the average path length over all connected subgraphs.

Average path length may be computed **both** for directed and undirected graph. For an undirected graph (5) may be rewritten as $l_G = \frac{2}{N(N-1)} \sum_{i,j \in V; i<j} g(i,j)$.

[1]Note that the geodesic-path the algorithm should be run as it is. First we obtain geodesic distances and then swap all infinities for D. (This order is important, as otherwise the rules of adding to infinity (last loop) are altered).

3.5 Resilience-related measures on network, [5]

3.5.1 Level of connectivity

- density of the links within the network (density of the graph) – the ratio of existing links M to all possible links in the network.
  - undirected graph
$$\Delta = \frac{M}{\binom{N}{2}} = \frac{2M}{N(N-1)} = \frac{\overline{d}}{N-1}. \tag{6}$$
  - directed graph
$$\Delta = \frac{\sum_{i=1}^{N} d(i)}{N(N-1)}, \tag{7}$$
  where $d(i) = d^{in}(i) + d^{out}(i)$.

  We can also define the density of a subgraph, Δ_S
$$\Delta_S = \frac{2M_S}{N_S(N_S-1)}$$
  Δ_S measures the density of links between the nodes in a given subset and is used to evaluate the cohesiveness of subgroups. (For directed graphs Δ_S may be computed analogously to (7)).

- reachability – Janssen et al. [5] suggest using network diameter as a measure of reachability. The diameter of a connected graph is defined as the length of the largest geodesic connecting any pair of nodes in the network, [6]
$$\text{diam} = \max_{i,j \in V} g(i,j) \in \{1, 2, \ldots, N-1\}. \tag{8}$$
  If a graph is not connected, diam $= \infty$ (or is undefined).

  Reachability applies to **both** directed and undirected graphs.

3.5.2 Level of centrality

The measures of centrality are defined for undirected graphs.
Degree based measures

- actor level: node degree $C_D(i) = d(i)$, or standardized measure $C'_D(i) = \frac{d(i)}{N-1}$
- group level:
  - general centralization index
$$C_D = \frac{\sum_{i=1}^{N} C_D^* - C_D(i)}{(N-1)(N+1)}, \tag{9}$$

where C_D^* is the largest observed value, $C_D^* = \max_i C_D(i)$. The maximum value (1) and minimum value (0) are attained for *star* and *regular graph*, respectively.

- variance of the degrees

$$S_D^2 = \frac{(\sum_{i=1}^{N} C_D(i) - \overline{C}_D)^2}{N}, \tag{10}$$

where $\overline{C}_D = \frac{1}{N}\sum_{i=1}^{N} C_D(i)$. S_D^2 is zero for a regular graph.

Closeness centrality

- actor level:

$$C_c(i) = \left[\sum_{j=1}^{N} g(i,j)\right]^{-1} \tag{11}$$

$C_c(i)$ reflects how close vertex i is to other vertices. It depends not only on direct ties but the vertices that are not adjacent to i are also taken into account. The standardized measure is given as $C_c'(i) = (N-1)C_c(i)$.

- group level:
  - Index of group closeness

$$C_c = \frac{\sum_{j=1}^{N} C_c'^* - C_c'(i)}{[(N-2)(N-1)]/(2N-3)} \tag{12}$$

where $C_c'^*$ is a maximum obtained value. This measure attains minimum when all geodesics are equal and maximum for a *star graph*.

  - the variance of standardized actor closeness indices

$$S_c^2 = \frac{(\sum_{i=1}^{N} C_c'(i) - \overline{C'}_c)^2}{N}. \tag{13}$$

4 Agents-focused measures

4.1 Measures of decision clustering. Spatial autocorrelation

The concept of clustering is to find the degree to which neighbors in a space share common attitudes. It relates the probability of sharing the same attitude and the distance between two sites. In a random configuration the probability of any two individuals sharing a common view is independent of the distance and depends only on the proportions of of people holding each attitude.

The term 'spatial autocorrelation' refers to the correlation of locational and attribute similarities among spatial objects and has its roots in the area data analysis. It resembles the Pearson's R^2 correlation coefficient. I varies (though not strictly) between the values '-1' and '$+1$.' Positive spatial autocorrelation indicates that like values tend to cluster in space whereas the negative SA suggests that neighbors are dissimilar. The random patterns exhibit zero spatial autocorrelation, see fig. 3.

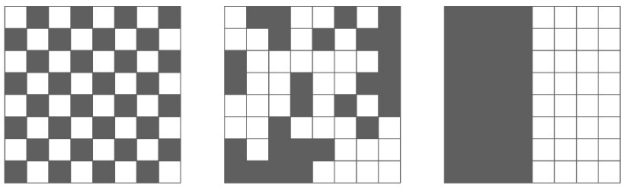

Figure 3: From the left: negative, zero and positive spatial autocorrelation pattern.

There is a number of indices of spatial autocorrelation, e.g. Moran (1950) (mainly applied to the continuous type of data but using it in a discrete case does not pose any difficulties).

$$I = \frac{N \sum_i \sum_j W_{ij} \left(X_i - \overline{X}\right) \left(X_j - \overline{X}\right)}{(\sum_i \sum_j W_{ij}) \sum_i \left(X_i - \overline{X}\right)^2} \tag{14}$$

where X_i, $i = 1, \ldots, N$ is the variable value (attribute) at i location, $\overline{X}$ is the average value taken over all locations and W_{ij} is a weight applied to the comparison between locations i and j. Weights can be based on adjacency matrix ($W_{ij} \in \{0, 1\}$) or distance (e.g. inverse distance).

We compare the observed value of spatial autocorrelation with the value that we would expect in the random case. The first way to assess the significance of the Moran's I statistics is restricted to the data assumed to follow approximately a normal distribution. In this case

$$\frac{I - \mathrm{E}(I)}{S_{\mathrm{E}(I)}} \overset{d}{\sim} N(0, 1),$$

where

$$\mathrm{E}(I) = -\frac{1}{N-1}$$

and

$$S_{\mathrm{E}(I)} = \sqrt{\frac{N^2 \sum_{ij} W_{ij}^2 + 3(\sum_{ij} W_{ij})^2 - N \sum_i (\sum_j W_{ij})^2}{(N^2 - 1)(\sum_{ij} W_{ij})^2}},$$

particularly, for the adjacency matrix

$$S_{\mathrm{E}(I)} = \sqrt{\frac{N^2 S + 3S^2 - N \sum_i (k_i)^2}{(N^2 - 1) S^2}}.$$

The p-value corresponding to the null hypothesis of no spatial autocorrelation is of the form

$$p = 2\left(1 - F\left(\frac{I + (N-1)^{-1}}{S_{\mathrm{E}(I)}}\right)\right),$$

where F is the standard Gaussian cumulative distribution function.

Significant spatial autocorrelation is indicated by p-value p lower than the significance level (usually 0.05).

The other way bases on the idea of sampling. We may realize spatial randomness both by free sampling or nonfree sampling (randomization/permutation). In free sampling the nodes are assigned attributes independently with the probabilities corresponding to their frequencies. In non-free sampling (exact test, permutation test) the nodes are assigned attributes with the constraint that the total number of each attribute is maintained (the data is permuted). In both cases of sampling the value of the test statistic (here, I) is compared to reference distribution being the distribution of the test statistic assuming the random sampling (null hypothesis is true). The p-value is the proportion of the distribution that is at least as extreme than the observed statistic, i.e.

$$p = \frac{\#\{I_i^{\pi} \geq I\}}{\text{no of permutations}},$$

where I_i^{π} stands for the autocorrelation of ith realization of sampling.

5 Time series based measures

5.1 Kurtosis

Kurtosis is a measure of the "peakedness" of the probability distribution. Higher kurtosis means that more of the variance is due to infrequent extreme deviations, as opposed to frequent modestly-sized deviations. Most commonly kurtosis is defined as the fourth cumulant divided by the square of the variance of the probability distribution, namely

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3, \tag{15}$$

which is also known as excess kurtosis. For a sample $x_1, x_2, \ldots, x_n$ the sample kurtosis is

$$g_2 = \frac{n\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} - 3 \tag{16}$$

where $\overline{x}$ is the sample mean. The distributions are classified with respect to the value of kurtosis as follows

- mesokurtic (mesokurtotic) – distributions with zero kurtosis (example: normal distribution family)
- leptokurtic (leptokurtotic, "super Gaussian") – distribution with positive kurtosis; a leptokurtic distribution has a more acute "peak" around the mean and "fat (heavy) tails" (examples: Laplace distribution, logistic distribution, t-student distribution)

- platykurtic (platykurtotic, "sub Gaussian") – distribution with negative kurtosis; a platykurtic distribution has a smaller "peak" around the mean and "thin tails" (examples: continuous or discrete uniform distributions, raised cosine distribution)

5.2 Estimation of the tail exponent

This section refers to data that exhibits power law behavior for right tails (power laws are easily detectable with log-log plots, i.e. power-law data with tail exponent α is seen as a straight line with slope α)

A simple method follows from the fact that for large x the logarithm of the distribution tail (1- cdf) is linear, i.e.,

$$\log P(X > x) \approx const - \alpha \log x. \tag{17}$$

For a sample $x_1, x_2, \ldots, x_T$ we build the order statistics as follows

$$x_{(1)} \geq x_{(2)} \geq \ldots \geq x_{(T)}.$$

Only m largest order statistics (m largest elements of the series x_i) are used to build the estimator (Hill,1975)

$$\hat{a} = \left(\frac{1}{m-1} \sum_{i=1}^{m} \log x_{(i)} - \log x_{(m)} \right)^{-1} = \left(\frac{1}{m-1} \sum_{i=1}^{m} \log \frac{x_{(i)}}{x_{(m)}} \right)^{-1}.$$

A simple rule of thumb says that m should be chosen so that the ratio of m to all elements (m/T) is around 0.5%–1%.
Note that Hill estimation is sensitive to the data set (choice of m).

5.3 Detecting heteroscedasticity

Useful notions:

- A sequence of random variables is **homoscedastic** if all random variables in the sequence have the same finite variance. This is also known as homogeneity of variance. See Fig. 4
- The complement of homoscedasticity is called **heteroscedasticity**. Heteroscedasticity can arise in a variety of ways. Typically tests for heteroscedasticity are designed to test the null hypothesis of homoscedasticity (equal error variance) against some specific alternative heteroscedasticity specification. See Fig. 4

Preliminary analysis. First step is to identify and remove the trend and the seasonality pattern from the data. The standard step is to plot the series and its autocorrelation function, *ACF*. The sample autocorrelation function (set of observations $\{x_1, x_2, \ldots, x_T\}$) is given as

$$ACF(k) = \frac{\bar{\gamma}(k)}{\bar{\gamma}(0)}$$

where $\bar{\gamma}(k)$ is sample autocovariance function is

$$\bar{\gamma}(k) = \frac{1}{T}\sum_{t=1}^{T-k}(x_t - \bar{x}_T)(x_{t+k} - \bar{x}_T), \quad k = 0, 1, ..., T-1$$

and $\bar{x}_T = T^{-1}\sum_{t=1}^{T} x_t$.
If the series is nonstationary, the ACF decays slowly and the usual solution is to analyze the differenced series (once or more, at lag one or more). Normally, the correct amount of differencing is the lowest order of differencing such that a time series fluctuates around a well-defined mean value and ACF plot decays fairly rapidly to zero (either from above or below). One has to be careful to avoid over-differencing, usually $d \leq 2$. In case of seasonal nonstationarity the ACF is zero except at lags $S, 2S, 3S, \ldots$ and decays very slowly. This can be made stationary by seasonal differencing $(1-B^S)x_t$, $(1-B^S)^2 x_t, \ldots, (1-B^S)^D x_t$. Usually $D = 1$ [7].
Detecting heteroscedasticity. *ACF* plot provides a tool to determine the conditional heteroscedasticity. Although the *ACF* of the observed data, i.e. $\{x_1^2, x_2^2, \ldots, x_T^2\}$ exhibits little correlation, the *ACF* of the squared data may still indicate significant correlation and persistence in the second-order moments [8], see Fig. 5

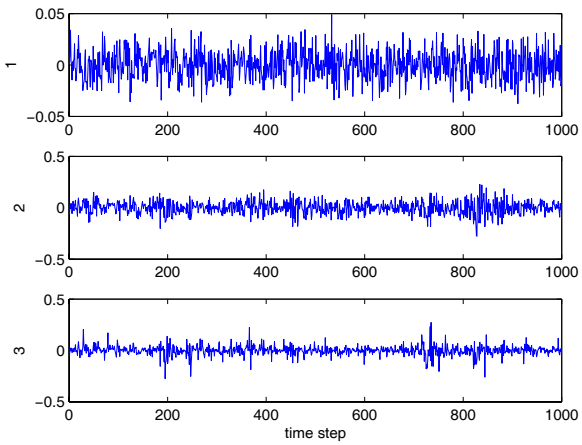

Figure 4: Examples of: homoscedastic data (1), heteroscedastic data: (2,3)

References

[1] Ph. Blanchard *et al.*, *Cluster percolation in O(n) spin models* J. Phys. A: Math. Gen. 33 (2000) 8603-8613.

[2] K.Christensen, N.Moloney, *Complexity and Criticality*, Imperial college Press, London, 2005.

[3] M.E.J. Newman, The Structure and Function of Complex Networks, SIAM 45 (2) 167-256 (2003).

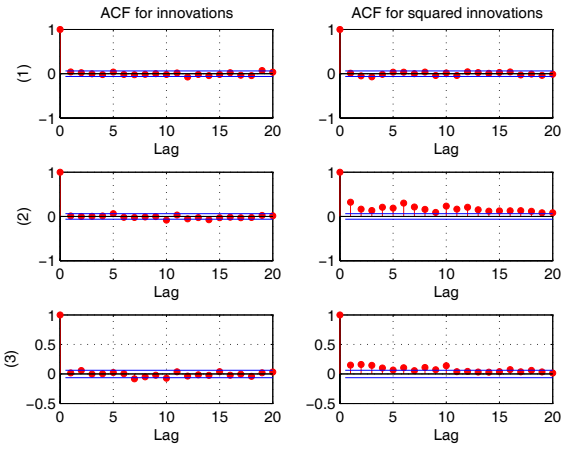

Figure 5: ACF analysis for data presented in Fig. 4. The observations themselves are largely uncorrelated but the variance process exhibits some correlation (data 2,3).

[4] M.E.J. Newman, Assortative Mixing in Networks, *Phys. Rev. Lett.* **89**, 208701 (2002).

[5] Janssen, M. A., Ö. Bodin, J. M. Anderies, T. Elmqvist, H. Ernstson, R. R. J. McAllister, P. Olsson, and P. Ryan. 2006. *A network perspective on the resilience of social-ecological systems*, Ecology and Society 11(1): 15. http://www.ecologyandsociety.org/vol11/iss1/art15/.

[6] Wasserman, S., Faust, K., Social Network Analysis: Methods and Applications Cambridge University Press 1994

[7] R.H. Shumway, Applied Statistical Time Series Analysis, Prentice Hall, New Jersey 1988.

[8] http://www.mathworks.com