Listen to Your Data: Econometric Model Specification through Sonification

Christopher S. McIntosh, Professor, Agricultural Economics and Rural Sociology, University of Idaho, Moscow, ID.

Ron C. Mittelhammer, Regents Professor, School of Economics Sciences, Washington State University, Pullman, WA.

Jonathan N. Middleton, Professor and Chair, Music Department, Eastern Washington University, Cheney, WA

Selected Paper prepared for presentation at the Agricultural & Applied Economics Association's 2013 AAEA &CAES Joint Annual Meeting, Washington, DC, August 4-6, 2013

ABSTRACT

Ever since 1927, when Al Jolson spoke in the first "talkie" film *The Jazz Singer*, there had been little doubt that sound added a valuable perceptual dimension to visual media. However, despite the advances of over 80 years, and the complete integration of sound and vision that has occurred in entertainment applications, the use of sound to channel data occurring in everyday life has remained rather primitive, limited to such things as computer beeps and jingles for certain mouse and key actions, low battery alarms on a mobile devices, and other sounds that simply indicate when some trigger state has been reached – the information content of such sounds is not high.

Non-binary, but still technically rather simple data applications include the familiar rattling sound of a Geiger counter, talking clocks and thermometers, or the sound output of a hospital EKG machine. What if deletion of larger and/or more recently accessed computer files resulted in a more complex sound than for deleting smaller or rarely accessed files, increasing the user's awareness of the loss of larger or more recent work efforts? All of these are examples of data sonification.

While sonification seems to be pursued mostly by those wishing to generate tuneful results, many are undertaking the process to simply provide another method of presenting data. Many examples are available at https://soundcloud.com/tags/sonification including some very tuneful arrangements of the Higgs Boson. Indeed, with complex data series one can often hear patterns or persistent pitches that would be difficult to show visually. Musical pitches are periodic components of sound and repetition over time can be readily discerned by the listener. Sonification techniques have been applied to a variety of topics (Pauletto and Hunt, 2009; Scaletti and Craig 1991; Sturm, 2005; Dunn and Clark, 1999). To the authors' knowledge, Sonification has yet to be applied in any substantive way to economic data.

Our goal is not to produce tuneful results. Rather, the purpose of this paper is to explore the potential application of Sonification techniques for informing and assessing the specification of econometric models for representing economic data outcomes. The purpose of this seminal and exploratory analysis is to investigate whether there appears to be significant promise in adding the data sonification approach to the empirical economists' toolkit for interpreting economic data and specifying econometric models. In particular, is there an advantage to using *both* the empirical analyst's eyes *and* ears when investigating empirical economic problems?

JEL classifications: C01, C18, C52

1.0 The What and Why of Data Sonification

People process auditory information differently than visual information. Much has been written in the education literature about learning styles based in part on auditory versus visual delivery of information. Students of all ages and abilities have preferences for the ways in which they receive information. The education literature provides a number of sources examining the various learning styles of students. These are typically classified as visual (V), auditory (A), reading/writing (R) and kinesthetic (K) and are referred to in the rest of the paper by the abbreviation VARK. Typical findings show that while students may prefer a specific style of learning, many of them benefit from being presented with multiple modes (see for example Lujan and Dicarlo, 2005; Felder and Silverman, 1988; and Rumburuth and McCormick, 2001)

As teachers of econometrics we have long suggested that our students look at their data by means of scatterplots - each variable, dependent and independent being examined as a function of other variables and by observation. Peter Kennedy (2008), in his popular econometrics textbook, advises "researchers should supplement their summary statistics with simple graphs: histograms, residual plots, scatterplots of residualized data and graphs against time." The time students invest in examining the data prior to running simple regression models will typically inform their analysis by visualizing things like correlations, outliers and structural shifts. Kennedy also states "the advantage of graphing is that a picture can force us to notice what we never expected to see." Instructors of econometrics have long suggested that students construct such graphs for just these reasons.

If the goal then is to specify and assess econometric models using visual depiction of various aspects of the model, then it seems logical to ask whether the analyst could reasonably engage other senses during this process. With the advent of massively increased computing power, larger and larger data sets, increasing complexity of model and system specifications, and with the attendant high dimensional multivariate nature of these models, understanding and interpreting model specifications, and their adequacy, has become increasingly more difficult. Indeed, the visual senses of most individuals become quickly overwhelmed once one leaves the familiar visualizable three-dimensional confines of the physical world. However, the auditory senses are comfortable and experienced with much higher dimensional simultaneous processing of inputs (sound signals and music). If sonification can be shown to provide potentially useful additional perspectives concerning multivariate relationships existent in economic data, the methodology could present a welcome addition to the methodology of empirical economics.

1.1 The Meaning of Sonification

What is the specific meaning of the term "*sonification*"? Thomas Hermann, whose influential PhD dissertation and continuing research has created significant increased interest in the possibility of the use of sonification for improving the effectiveness of data analysis (Hermann, 2002), offered both a laymen's definition, and a more formal definition of the term. Colloquially, and paraphrasing Hermann, sonification is the use of sound for representing or displaying data, and similar to scientific visualization, aims to enable human listeners to make use of their highly-developed perceptual skills (in this case listening skills) for making sense of data (Herrmann, 2012). Somewhat more formally, sonification is the transformation of data relations into perceived relations in an acoustic signal for the purposes of facilitating communication or

interpretation. Hermann went on to provide an axiomatic definition of sonification, which he expected would help position the methodology for use in scientific contexts, using the following characterization (with some minor editing to enhance clarity (Hermann, 2008)):

Definition: **Sonification**. Any technique that uses data as input, and generates sound signals, may be called sonification *iff*

A. *The sound reflects objective properties or relations in the input data.*
B. *The transformation is systematic, meaning that there is a* ***precise*** *definition of how data causes the sound to change.*
C. *The sonification is reproducible: given the same data, the resulting sound must be structurally identical.*
D. *The system can intentionally be used with different data, and also be used in repetition with the same data.*

1.2 Sonification Software: Musicalgorithms

Following Hermann's definition we have maintained a precise and reproducible methodology for converting our econometric data into sound. All sonifications presented in this paper were accomplished using the "Musicalgorithms" software developed by Middleton (2005). The software was accessed through the Musicalgorithms website (http://musicalgorithms.org). Musicalgorithms converts data into sound by transforming data observations through both the timing and pitch of sounds, the latter being a function of the magnitude of data observations, where a higher-valued data point is transformed into a higher pitched sound. Sonification in this program environment requires that the data be converted from their original values and range to

discrete numbers within the integer-valued range of 1 to 88. This is exactly the standard range of a modern piano keyboard, and it is to this instrument that we map our data.

In our applications of sonification all of the data refer to OLS error vectors, although the residuals from any econometric model could be used following precisely the same approach. Apart from an initial artificial example that we present for illustrative purposes, we mapped error vectors to the range of the piano keyboard using the Musicalgorithms' "division" option, which maps the numeric data proportionally throughout the 88 key range (Middleton, 2008). Using this approach, the data point with the smallest value will be mapped to the lowest key on the keyboard and the data point with the highest value will be mapped to the highest key on the keyboard. All other data points are mapped proportionally between these extreme values. This conversion process meets all of the requirements set forth in Hermann's definition of sonification One also has the option of restricting the data transformation process further by constraining the sound output to a subset of the full piano keyboard, but in our substantive applications, for all error vectors displaying violations of the Gauss-Markov assumptions, the full range of 88 semitones were used.

2.0 Conversions of OLS Residuals to Sound

In this section we present examples of sonifications of a number of error processes that represent various types of violations of general linear model assumptions. We begin with a simple but artificial econometric model that is designed to illustrate in a clear and straightforward manner how patterns in residual vectors, and concomitant violations of standard general linear model assumptions, can be recognized through an econometrician's auditory perceptions. We then present a series of examples in a more typical and familiar model setting, where we alter the

structure of the underlying data generating process to generate new data sets that exhibit error processes violating standard general linear model assumptions in various ways.

2.1 Omitted Variables: Mary Had a Little Lamb

As the title of this subsection suggests, our first example of sonification is admittedly a somewhat tongue-in-cheek application, but it produces a vivid illustration of how sound can be used to detect error assumption violations, in this case, the problem of "omitted variables" from the specification of the conditional regression function. Consider a linear model of the form

$$Y_t = \beta_0 + X_{1t}\beta_1 + X_{2t}\beta_2 + \varepsilon_t \qquad (1)$$

The error process is independent and identically distributed Bernoulli, with $p = .5$, translated to a zero mean, i.e., $\varepsilon_i = z_i - .5,\ z_i \sim iid\ \ Bernoulli(.5),\ i = 1,...,n$. The data consists of $n = 26$ observations and can be assumed to satisfy all of the standard general linear model assumptions that lead to BLUE estimates of the β parameter vector.

The midi file *Bernoulli.mid* (http://webpages.uidaho.edu/mcintosh/Brenoulli.mid) contains the outcomes of the error terms. The sonification of the error process consists of two different notes, played in sequential random order. It is immediately apparent from the dichotomy of the sounds that the support of the error process is a dichotomy, suggesting that a scaled and translated Bernoulli-type process underlies the generation of the errors. The apparent lack of "runs" or groupings in the sounds played suggests that the errors are likely generated at random (i.e., independently).

Now consider an omitted variables version of the model, whereby

$$Y_t = \beta_0 + X_{1t}\beta_1 + v_t,\ where\ v_t = X_{2t}\beta_2 + \varepsilon_t \qquad (2)$$

and the omitted component $X_{2t}\beta_2$ that now appears in the error term v_t is such that the pattern of the observations in X_{2t}, scaled by the value of β_2, sonifies to the tune of "Mary Had a Little Lamb". The sonification of the 26 outcomes in the error vector v_t is provided in the midi file *OmittedMary.mid* (http://webpages.uidaho.edu/mcintosh/OmittedMary.mid). In this sonification, one hears simultaneously the dichotomous sounds of the original error process in (1) together with the sounds of the omitted variable effect, suggesting there exists a systematic component to the error process, and signaling a misspecification of the conditional regression function.

While admittedly artificial, where the omitted variable component of the sonification turns out to be a highly familiar and recognizable tune, it is this type of auditory processing that lies at the root of the intent of econometric data sonification. While still in its infancy, one can conceive of a longer run context in which the song library of econometric misspecifications has been built up, and the associated tunes learned such that the use of auditory perception becomes a member of the econometrician's toolkit for exploring the specification of econometric models.

2.2 A Prototypical Econometric Model Setting

In this subsection, we examine an econometric model in which the original data are based on an acreage response function for wheat using the price of wheat, price of barley and price of potatoes along with time as explanatory variables. The data generating process was based on the equation:

$$Wheat\ Acres\ Planted_t = 1050 + 35Pwheat_t - 35Pbarley_t - 5Ppotatoes_t + 10Time_t + \varepsilon_t \qquad (3)$$

Data for the commodity prices used in this project are found in appendix B, Table B1. The time variable takes on integer values from one through fifty, corresponding to each observation. The error vector $\boldsymbol{\varepsilon} = [\varepsilon_1, ..., \varepsilon_n]'$, with $n = 50$, was modified to create models exhibiting specific Gauss-Markov violations and generate a dependent variable vector that would cause an estimated OLS model to exhibit the specific Gauss-Markov violations. In particular, these violations consisted of first order autocorrelation, two types of heteroscedasticity, and omitted relevant variables. The autocorrelation process reflected a strong positively autocorrelated error evolution, with $\rho = .9$. Regarding the two heteroscedastic error processes, the first represented a structural change in error variance occurring half-way through data series, and the second was an error variance structure that increased as the observation number increased. The omitted variables simulation involved a relevant time trend variable that was omitted from the structural part of the model specification. All simulations are based on $n = 50$ observations and assume normally distributed errors. The data are generated by first drawing a random sample from the normal error process, and then applying the regression model in (3) to produce the dependent variable values. Then the simulated data was used to fit the parameters of the following general linear model specification:

$$Wheat\ Acres\ Planted = \beta_0 + \beta_1 Pwheat + \beta_2 Pbarley + \beta_3 Ppotatoes + \beta_4 Time + \varepsilon \quad (4)$$

The OLS residuals from these regressions were then subjected to the sonification process. The analyzed scenarios are described in detail in the discussion that follows.

The baseline error structure was initially defined as *iid* normal with mean zero and variance equal to 15. A data series was generated, an Ordinary Least Squares regression model

was then estimated, and the OLS residuals were sonified. The result of this baseline simulation is shown in Figure 1 and the sonification of the estimated residuals can be heard in audio file 1 located at http://webpages.uidaho.edu/mcintosh/audio1normal.mid. This sequence is also shown in music notation for the musically adept econometrician in Appendix A as Figure A1.

2.2.1 First Order Autocorrelation

Redefining the error structure so that $\varepsilon_t = \rho\varepsilon_{t-1} + v_t$ where $v_t \sim N(0,15)$ and ρ is set equal to 0.9 generates a data set whose OLS residuals exhibit a statistically significant amount of autocorrelation. These residuals were sonified by order of observations as well as sorted by the magnitude of ε_{t-1}. Sorting by the magnitude of the lagged error vector is a commonly suggested method for determining first order autocorrelation. When a strong positive trend is exhibited, this would be indicative of positive first order autocorrelation (see for example Gujarati, 2011). These series are shown in Figures A2 and A3 and can be heard in audio files 2 and 3 located at http://webpages.uidaho.edu/mcintosh/audio2AR1obs.mid and http://webpages.uidaho.edu/mcintosh/audio3AR1et-1.mid

The audio patterns from these two series are what one would expect to hear from a strongly positively autocorrelated data series. In particular, when ordered by observation number, the audio reflects the repeated peaks and valleys that one would expect from residuals generated from an AR1 process. When the data are ordered by magnitude of ε_{t-1} and sonified the pitches begin in the lower register and steadily move higher, as expected.

2.2.2 Heteroscedasticity

As noted in the introduction to section 2, two types of heteroscedasticity were generated. The first is the result of a "structural change" in the error process with the first 25 observations

generated from a homoscedastic $N(0,15)$ population distribution and the last 25 observations being generated from a homoscedastic $N(0,75)$ population distribution. The residuals from this regression can be heard in audio file 4 located at http://webpages.uidaho.edu/mcintosh/audio4hetero.mid and shown in musical notation in Appendix Figure A4.

The change in the spread of the pitch between notes played in the first 25 observations and notes played in the last 25 observations is evident. This variation in pitch suggests that the spread of observations, i.e. the variance, changes midway through the set of observations, and thus is indicative of both heteroscedasticity, and of the type of heteroscedasticity, i.e., a structural shift in the variance of the process.

The second set of heteroscedastic data were generated from an error vector distributed as $N(0, 3 \times time)$. Thus as the observation number (time) increases, so does the variance. An OLS regression was estimated from this data and the resulting error vector can be heard in audio file 5 located at http://webpages.uidaho.edu/mcintosh/audio5hetero.mid. This sonified error vector is shown in musical notation in Appendix Figure A5.

In this sonification, a gradually growing spread in the pitch between notes is heard as the data observations progress. This variation in pitch suggests that the spread of observations, i.e., the variance, is widening over the observations in the data set, indicative of both heteroscedasticity, and the type of heteroscedasticity, i.e., a variance that has an upward drift as the observations progress over time.

2.2.3 Omitted Variable

The case of an omitted explanatory variable was examined by estimating a regression based on the data with error terms defined as $iid\ \varepsilon_t \sim N(0,15)$. A regression model omitting the time trend was then estimated from this data (the full correctly specified model results are found in audio file 1 and Figure A1). The residuals from the restricted model were then sonified. These can be found in audio file 6 located at http://webpages.uidaho.edu/mcintosh/audio6omitted.mid and are shown in musical notation in Figure A6.

The pitch of the sound of the error process gradually moves upward as the data series progresses, suggestive of an upward drift in the mean of the error process. The sonification suggests that there is an omitted variable error in the structural specification of the conditional mean function, and also suggests that the omitted influences represent an upward drift in the conditional mean of the regression function.

3.0 Summary, Conclusions, and Future Potential

To our knowledge this paper represents the first application of economic data sonification presented at a major economics conference. Our efforts here illustrate that sonification can indeed be used to identify patterns in error vectors. With respect to autocorrelation and heteroscedasticity we are hearing the patterns that we expected to hear in data series that violate the assumption of constant finite variances. Regarding omitted variables, we were able to hear an upward structural drift in the error observations that resulted from an omitted variable that exhibited such an upward drift.

As econometricians, we often make use of visualization without considering the other potential options for exploring our data. This paper serves to illustrate that the sonification approach can indeed generate output that is useful to the econometrician. Sonification holds promise as an intuitive way of displaying patterns in the data particularly when graphical means cannot do justice to the data because of the level of dimensionality or the complexity of the underlying patterns. Econometricians are comfortable with graphical displays of data; it is something the discipline has been doing for a century or more. Sonification, on the other hand, is in its infancy. There is clearly a lot to learn about both data generating processes and the "sounds" associated with process outcomes.

While this may be the first application of sonification to economic data it will certainly not be the last. Mathematica now includes functions useful for sound synthesis and sonification. The R programming language has a sonification package known as playitbyr. As it currently stands, the process of sonifying data is less than straightforward. However, software packages such as these will undoubtedly aid in the expansion of sonification as a data analysis tool.

In conclusion, it is worth reiterating that the purpose of this paper was to explore the potential application of Sonification techniques for informing and assessing the specification of econometric models for representing economic data outcomes. Having completed this exercise it is our opinion that there does indeed appear to be significant promise in adding the data sonification approach to the empirical economists' toolkit for interpreting economic data and specifying econometric models. It is likely that a great deal of interesting information will be revealed about data generating processes encountered in practice if we will only take the time and effort to listen to our data, and begin accumulating a historical record of what we heard.

References

Dunn, J. and M.A. Clark. 2008. "Atmospherics/Weather Works: A Spatialized Meteorological Data Sonification Project" *Leonardo* 38(1): 31-36.

Felder, R.M. and L.K. Silverman, 1988. "Learning and Teaching Styles in Engineering Education" *Engineering Education*, 78(7): 674-681.

Gujarati, D. 2011. "Econometrics by Example" Palgrave Macmillan, New York, NY.

Hermann, T. 2002 "Sonification for Exploratory Data Analysis". Ph.D. Thesis, University of Bielefeld, Bielefeld, Germany.

Herrmann, T. 2008. "Taxonomy and Definitions for Sonification and Auditory Display", Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008.

Hermann, T. 2013. "SONIFICATION.DE", Thomas Hermann's research website on Sonification, Data Mining and Ambient Intelligence, http://sonification.de/son, accessed on 1/10/2013.

Kennedy, Peter, 2008. "A Guide to Econometrics" Sixth Edition. Blackwell Publishing, Malden, MA.

Lujan, H.L. and S.E. DiCarlo, 2005 "First-Year Medical Students Prefer Multiple Learning Styles" *Advances in Physiology Education*, March, 30(1): 13-16.

Mathematica Sound and Sonification, 2013. http://reference.wolfram.com/mathematica/guide/SoundAndSonification.html accessed June 2, 2013

Middleton, J.N., 2005. "Musicalgorithms" http://musicalgorithms.ewu.edu accessed May 30, 2013.

Middleton, J.N. and D. Dowd. 2008. "Web-Based Algorithmic Composition from Extramusical Resources," *Leonardo*, Vol. 41, No. 2: 128-135.

Pauletto, S. and A. Hunt. 2009. "Interactive Sonification of Complex Data" *International Journal of Human-Computer Studies*. 67(11): 923-33.

Playitbyr. 2013. http://playitbyr.org/ accessed June 2, 2013.

Ramburuth, P and J. McCormick, 2001. “Learning Diversity in Higher Education: A Comparative Study of Asian International and Australian Students” *Higher Education* 42:333-350.

Scaletti, C. and A.B. Craig. 1991. “Using Sound to Extract Meaning from Complex Data” *Electronic Imaging*, San Jose, CA. 0001: 207-219

Sturm, B.L. 2005. “Pulse of an Ocean: Sonification of Ocean Buoy Data”, *Leonardo* 38(2): 143-49.

Appendix A: Musical Notation

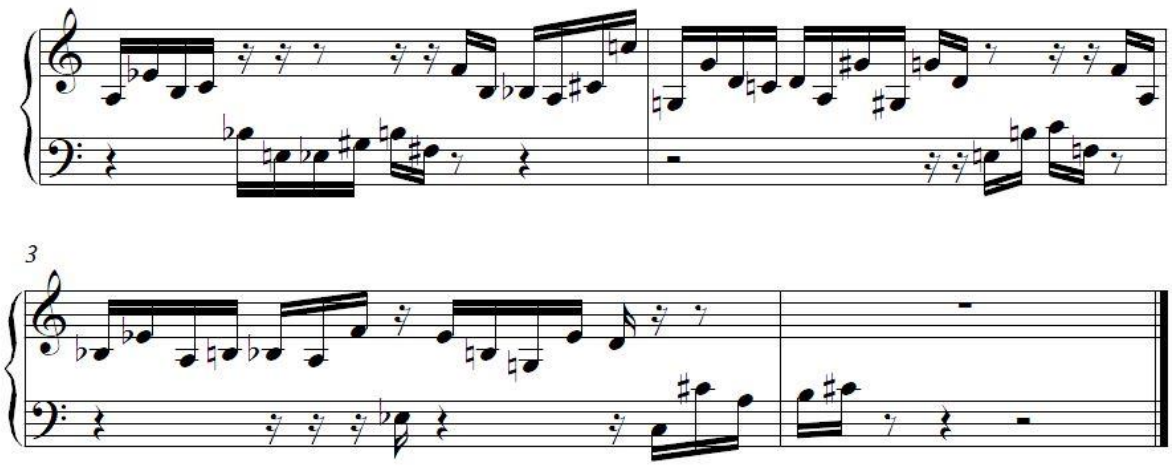

Figure A1. OLS residuals from an appropriately specified regression based on simulated data with $\varepsilon \sim N(0,15)$.

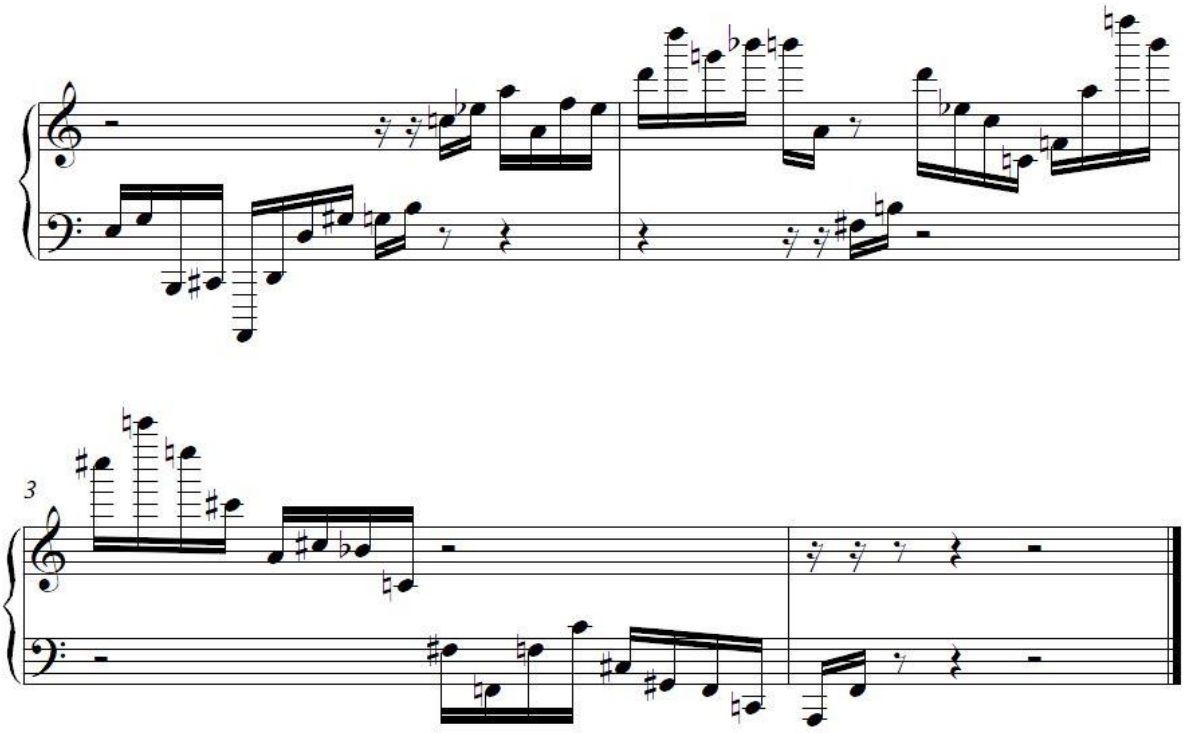

Figure A2. OLS residuals in observation order from a regression based on simulated data with $\varepsilon_t = \rho\varepsilon_{t-1} + v_t$ where $v_t \sim N(0,15)$ and ρ = 0.9.

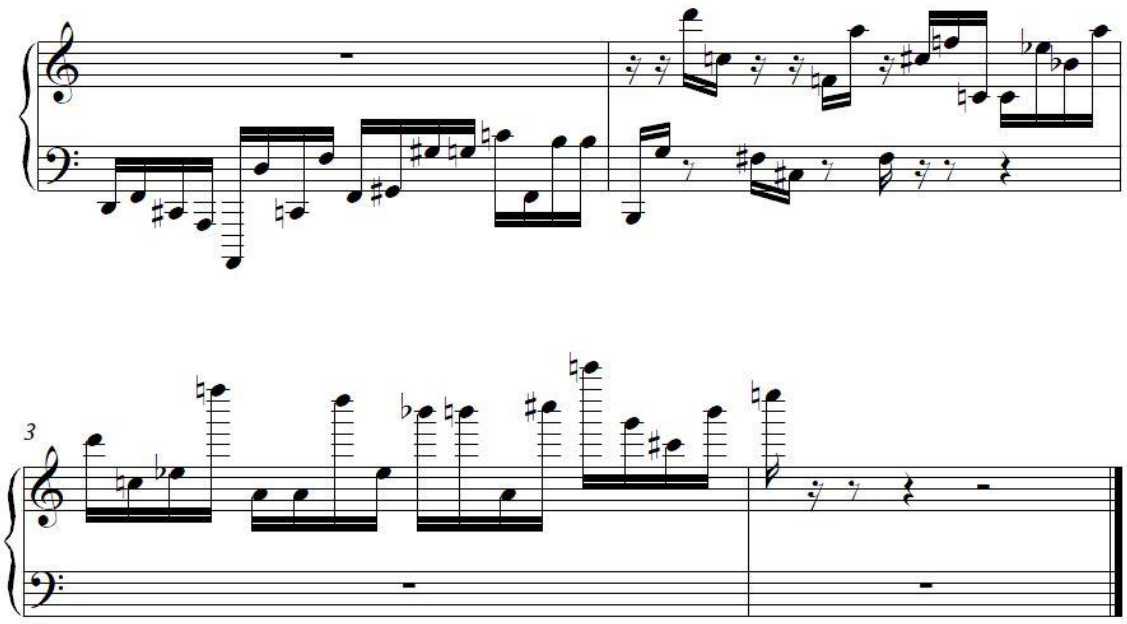

Figure A3. Figure A2. OLS residuals ordered by e_{t-1} based on simulated data with $\varepsilon_t = \rho\varepsilon_{t-1} + v_t$ where $v_t \sim N(0,15)$ and ρ = 0.9.

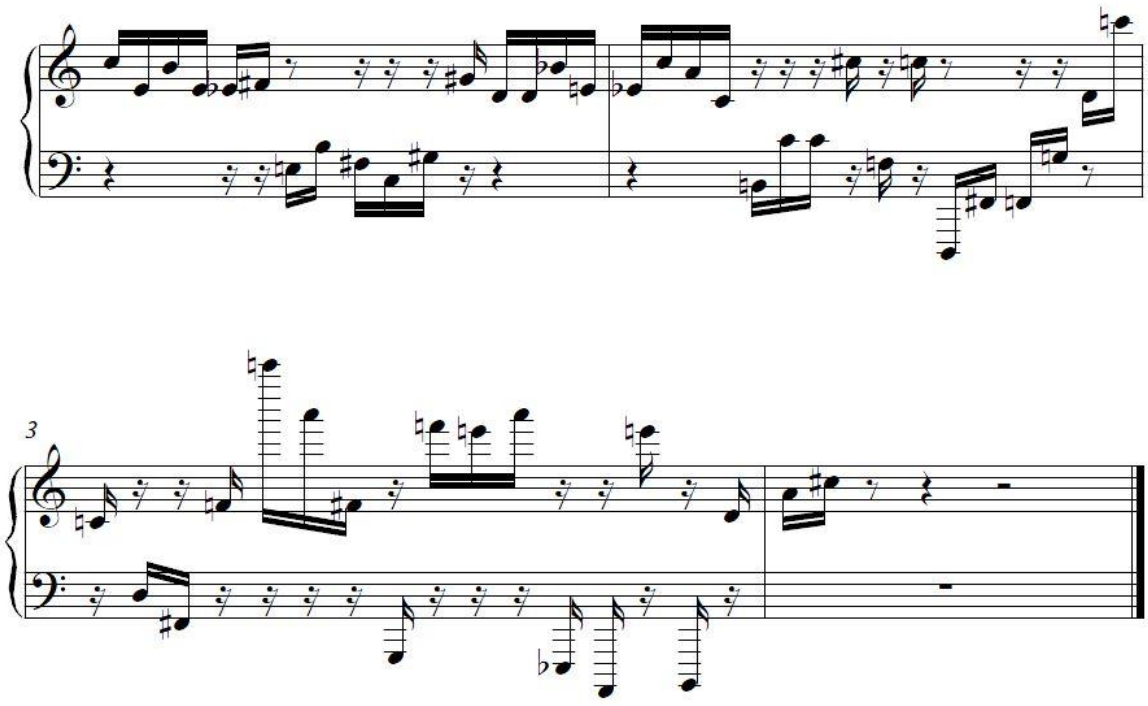

Figure A4. OLS residuals from a regression based on simulated data generated with the first 25 observations $\varepsilon \sim N(0,15)$ and the last 25 observations $\varepsilon \sim N(0,75)$

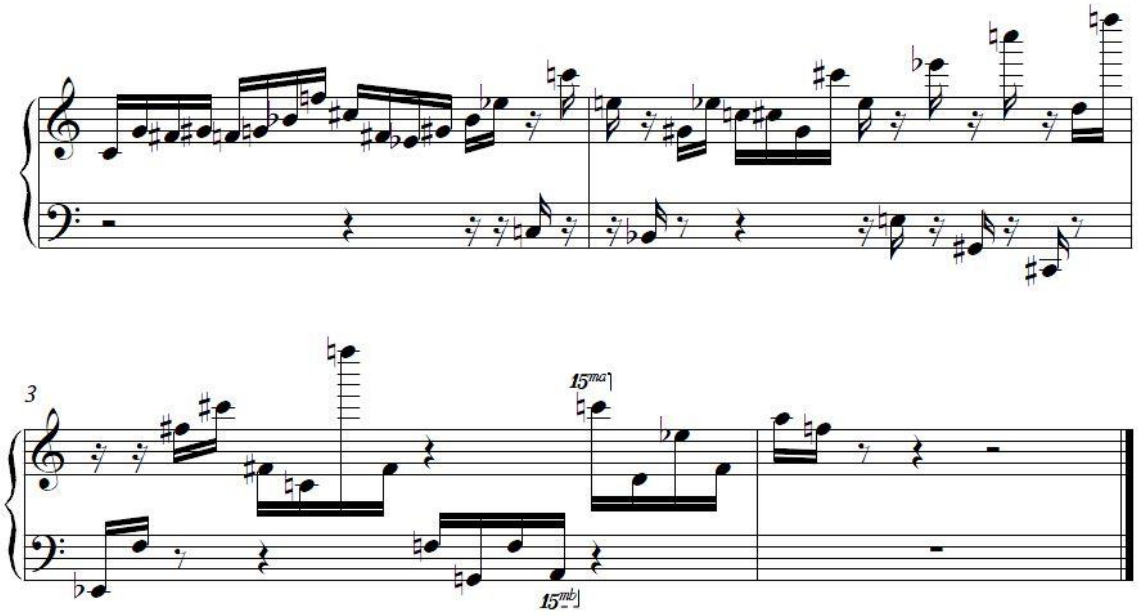

Figure A5. OLS residuals from a regression based on simulated data generated with $\varepsilon \sim N(0, 3 \times time)$ note that the abbreviation 15ma (15mb) indicates a note that is a two octaves, i.e. a fifteenth, above (below) the written note.

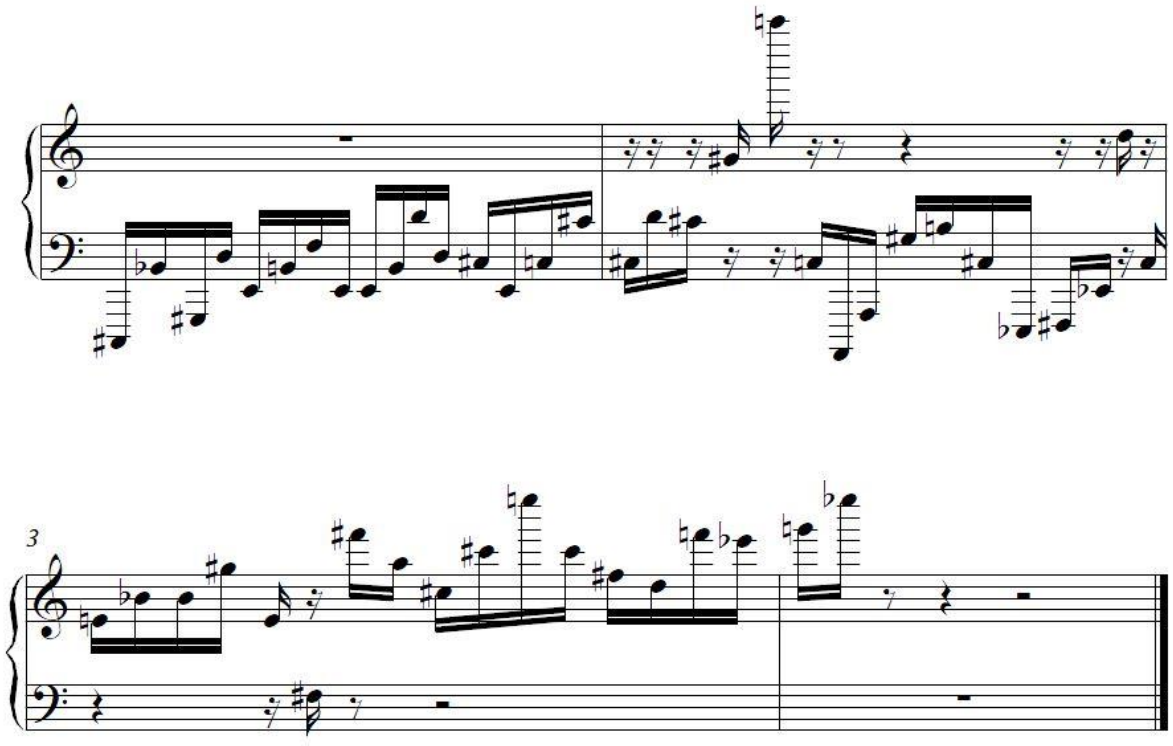

Figure A6. OLS residuals from a regression with the time trend omitted based on simulated data with $\varepsilon \sim N(0,15)$

Appendix B: Data

Pwheat	Pbarley	Ppotatoes	Time
3.8461	2.5960	9.6346	1
3.8490	2.0000	9.1509	2
3.4629	1.8703	6.7592	3
3.4000	1.5090	7.6727	4
3.2678	1.6428	6.7142	5
2.9473	1.4035	7.3333	6
2.9298	1.5610	8.9122	7
2.9322	1.4915	5.5080	8
3.0163	1.4590	4.7377	9
3.0785	1.5150	6.2650	10
2.9375	1.5000	7.4062	11
1.9390	1.3787	6.5606	12
1.9260	1.3970	6.2500	13
2.2142	1.3857	4.0714	14
1.8611	1.4444	5.5694	15
1.5676	1.2432	5.7432	16
1.6800	1.1866	4.6000	17
1.8289	1.2105	5.0921	18
1.7837	1.3108	5.7020	19
2.4743	1.3461	5.7690	20
4.7792	1.7142	8.9220	21
4.8433	2.8430	6.0722	22
4.4102	3.6923	5.1282	23
3.0740	2.8640	5.9012	24
3.3000	2.5875	6.5300	25
3.0048	2.3292	6.2430	26
2.9235	2.3764	5.6230	27
2.7258	2.7978	3.2921	28
3.1932	3.3290	5.4780	29
3.2584	2.9887	5.4166	30
3.8777	2.7667	7.7259	31
3.6304	3.0217	5.5452	32
3.5274	2.7360	6.4570	33
2.6483	2.5710	7.1080	34
2.8172	2.2365	5.2240	35
4.3080	2.3720	4.7535	36
3.8421	2.9052	5.3754	37
2.5918	2.7857	4.7653	38
3.6000	2.6200	7.2608	39
3.4653	2.7426	7.5248	40
2.8173	2.7019	5.6129	41
3.3964	2.4905	5.7303	42
4.1389	2.2407	5.6774	43
3.4695	2.6522	5.2927	44
2.6218	2.6386	5.1519	45
2.0176	2.4955	5.6954	46
2.1891	2.1621	4.5105	47
2.1293	2.2413	4.8491	48
2.1400	2.2700	5.1700	49
2.1678	2.1290	5.0190	50