An Approximate Entropy Based Approach for Quantifying Stability in Spatio-Temporal Data with Limited Temporal Observations

J. Piburn[1], R. Stewart[1], A. Morton[1]

[1]Oak Ridge National Laboratory, 1 Bethel Valley Road Oak Ridge, TN 37830
Email: piburnjo; stewartrn; mortonam@ornl.gov

Abstract

Identifying erratic or unstable time-series is an area of interest to many fields. Recently, there have been successful developments towards this goal. These newly developed methodologies however come from domains where it is typical to have several thousand or more temporal observations. This creates a challenge when attempting to apply these methodologies to time-series with much fewer temporal observations such as for spatio-temporal socio-cultural understanding, a domain where a typical time series of interest might only consist of 20-30 annual observations. Identifying instability in spatio-temporal trends is critical for understanding global dynamics and finding areas of potential concern or intervention. In the case of spatio-temporal socio-cultural understanding, instability is marked by two characteristics, 1) how widely varying the values are and 2) how predictable that variance is from one observation to the next. Approaches for simultaneously addressing both of these concerns have been limited. In this paper, we introduce an approximate entropy based method for characterizing the behaviour of a time series with limited temporal observations. For Geocomputation, this methodology represents a novel additional tool for researchers to use for exploring and understanding spatio-temporal data, specifically with limited temporal observations. As a case study, we look at national youth male unemployment across the world from 1991-2014.

Keywords: Time Series, Approximate Entropy, Spatio-Temporal, Exploratory Spatial Data Analysis (ESDA), Data Mining.

1. Introduction

In socio-cultural research, Identifying instability in spatio-temporal trends is critical for understanding global dynamics and finding areas of potential concern or intervention. A common constraint however, is the limited temporal observations that data of interest are typical of having. This leaves many of the advances from time series data mining literature unable to contribute to socio-cultural research, as they rely on having hundreds, if not thousands, of temporal observations for responsible application. Although potentially useful in other fields, this paper represents an attempt to develop a time series data mining technique specifically designed with the application constraints and operational definitions of identifying spatio-temporal instability in socio-cultural research.

For operational definitions in socio-cultural research, instability is marked by two characteristics, 1) how widely varying the values are and 2) how predictable that variance is from one observation to the next. How widely varying the values of a time series are can be considered simply through variance. However, moment statistics such as variance do not consider the order of the observations in their summaries. For example, a vector that alternates regularly between the values of 5 and 10 has the same variance as a vector that takes the value of 5 or 10 each with a probability of ½. Variance is thus a necessary, but not sufficient, measure of instability. To address how predictable changes in a time series are various instantiations of entropy have seen much success. Most of the entropic methods were born out of applications with numerous temporal observations, such as finance or heart rate monitoring, and thus were not designed with observational constraints in mind. A notable example otherwise is approximate entropy (ApEn). Introduced by Pincus (1991), ApEn is a computational approximation of Kolmogorov-Sinai entropy and is used to measure the amount of regularity and unpredictability in a time series. Due to its approximate nature ApEn is not burdened by the need for numerous observations and thus is particularly well suited for time series' consisting

of limited data points. ApEn however is no panacea. While useful in measuring how predictable changes are in a time series, ApEn does not consider how widely varying the changes are, just the regularity of a change. Furthermore, ApEn is a parametric model that puts requirements on the user to know at what value to set the input parameters. Of particular responsiveness is the value of r which sets the threshold of what the model considers a change. Anything above r is counted towards the measure and anything below it passes by as if nothing changed at all. This of course can widely alter the resulting ApEn value of time series if all changes in the series were just below this value compared to just above it.

In this paper we introduce a method, we refer to as the Attribute Stability Index (ASI), that uses both ApEn and variance for characterizing the instability of a time series with limited temporal observations that incorporates both how widely varying the values are and how predictable that variance is from one observation to the next. Additionally, this methodology allows for the removal of r as an input parameter and thus sidesteps the sensitivity concerns that were mentioned above. As a case study, we look at the ASI for national youth male unemployment across the world from 1991-2014.

2. Methodology

To interpret ASI values a basic understanding of ApEn is needed. Along with the time series vector under consideration, the ApEn algorithm takes two input parameters, m the length of comparison windows, and r a comparison distance which can be thought of as a filtering level for distances that are considered unexpected if they occur within m. The higher the resulting ApEn value is the more irregular and unpredictable the series is considered to be. For our purposes, m is set to 2 as we are concerned with immediate change from one observation to the next. As mentioned above, differing values of r can have a dramatic influence on the resulting ApEn values and taking advantage of this behaviour is at the center of how the ASI values are calculated. Thinking of ApEn for given time series as a function of r, the ASI value for that time series is equivalent to the approximation of the definite integral of ApEn from $r = 0$ to a logical maximum value that is defined below. By integrating over all values of r, we accomplish two things 1) we sidestep the problem of setting an arbitrary value of r and thus the sensitivity concern and 2) we are able to get a more complete picture of a time series' instability not only graphically but intuitively as well by incorporating all changes large and small. ApEn is used in equation 4 and is represented by $\Theta(.)$ however, due to space limitations, ApEn will not be discussed in depth, interested readers should see Pincus (1995) and Richman and Moorman (2000) for further details.

Given a time series of interest, $\mathbf{x} = (x_i, x_{i+1}, x_{i+2}, \dots, x_n)$, the first step is to calculate the lagged difference of the vector with a lag of 1, this can be seen in Equation 1.

$$\mathbf{x}_{lag} = \{(x_{i+1} - x_i), (x_{i+2} - x_{i+1}), \dots, (x_n - x_{n-1})\} \quad \text{Equation 1}$$

Once $\mathbf{x}_{lag}$ is defined, the absolute value of the maximum lag is calculated. This value is upper bound for the values of r used in the ApEn calculations, with the lower bound being 0 as stated above. The maximum lag is defined as the upper bound because any value of r that is greater than the largest lag by definition will result in an ApEn value of 0. Equation 2 uses this lag and an integer λ, which is set by the user as an input into the ASI calculations, to determine ρ, how large of a step to take between successive evaluations of ApEn. The larger the value of λ the closer the approximation will be to the definite integral. λ can be thought of as the resolution of the resulting approximation

$$\rho = \frac{max|\mathbf{x}_{lag}|}{\lambda} \qquad \text{Equation 2}$$

Using ρ and a vector of integers from 0 to λ we can construct a vector, $\mathbf{r}$ (Equation 3), that contains all of values of r for which we will evaluate ApEn used in the ASI calculation.

$$\mathbf{y} = (0,1,2,3,\dots,\lambda)$$
$$\mathbf{r} = \rho\mathbf{y} \qquad \text{Equation 3}$$

At this point the final ASI estimation can be written in the form of trapezoidal integration (Equation 4).

$$ASI(\mathbf{x},\lambda) = \frac{1}{2}\sum_{k=1}^{N}(r_{k+1} - r_k)\cdot(\Theta(\mathbf{x},r_{k+1}) + \Theta(\mathbf{x},r_k)) \approx \int_0^{\rho\lambda}\Theta(\mathbf{x})dx \qquad \text{Equation 4}$$

3. Case Study

As a case study of the this methodology we explore the behaviour of youth male unemployment as a percent of male labour force ages 15-24 from 1991-2014 at the national level for 170 countries. For illustration purposes, the raw trends from five example countries are shown in Figure 1 (left). These five countries have a range of different behaviours that provide insight into how ASI calculations behave. Botswana and France both have widely varying values with no clearly identifiable overall trend, Trinidad and Tobago has non trivial changes in value from one observation to the next but with an overall downward trend, and finally Bahrain and China have values that tend to change in an irregular pattern but by smaller amounts. Since the goal of the ASI is to identify trends that are both irregular and widely varying, these examples will provide a better understanding of how individual trends are scored.

The ASI results can be investigated in two ways. First, we can generate a curve over r (filtering level) and their resulting ApEn values as shown in Figure 1 (right) and second we can use the end result ASI values (the integral of those curves) and simply display them spatially (Figure 2) or use them as input into other models such as spatial clustering techniques (Figure 3).

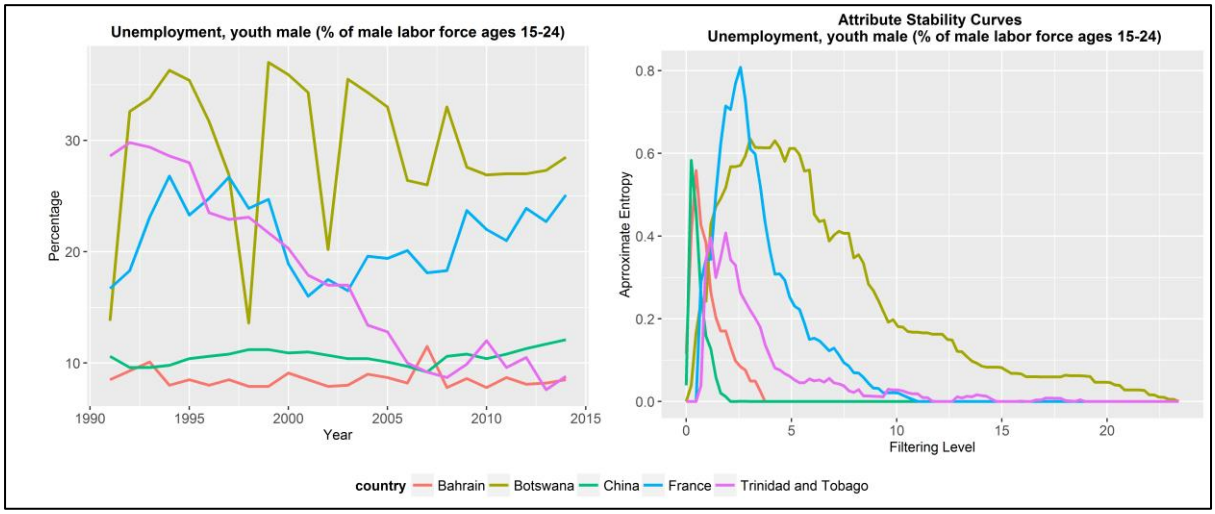

Figure 1. Youth Male Unemployment in Example Countries and Corresponding Attributes Stability Curves (1991-2014).

Figure 1 (right) displays the attribute stability curves for the five example countries from Figure 1 (left). These curves show the behaviour of ApEn values across all values from the $\mathbf{r}$ vector defined in equation 3, also known as the filtering level, and it is the definite integral of these curves that give us the ASI values. However, the total areas under these curves are not the only thing of interest, the shape of the curve also gives us information about the nature of the trend. We can see that at very low values of r, China and Bahrain have higher ApEn scores than France or Botswana, but then drop quickly to zero. This indicates that while the changes from year to year in China and Bahrain may be irregular, they do not vary by a large amount. France has the highest peak ApEn score of the example countries and maintains high ApEn scores across a wider range of r values than that of all countries with the exception of Botswana. Although Botswana's maximum ApEn score is not has high as that of France, the ApEn scores stay higher over a much wider range of r values. The shape of Botswana's stability curve tells us that not only are changes from year to year irregular, the amount by which it changes is also irregular.

Expanding the results to all countries in the case study, the spatial distribution of ASI scores can be seen in Figure 2.

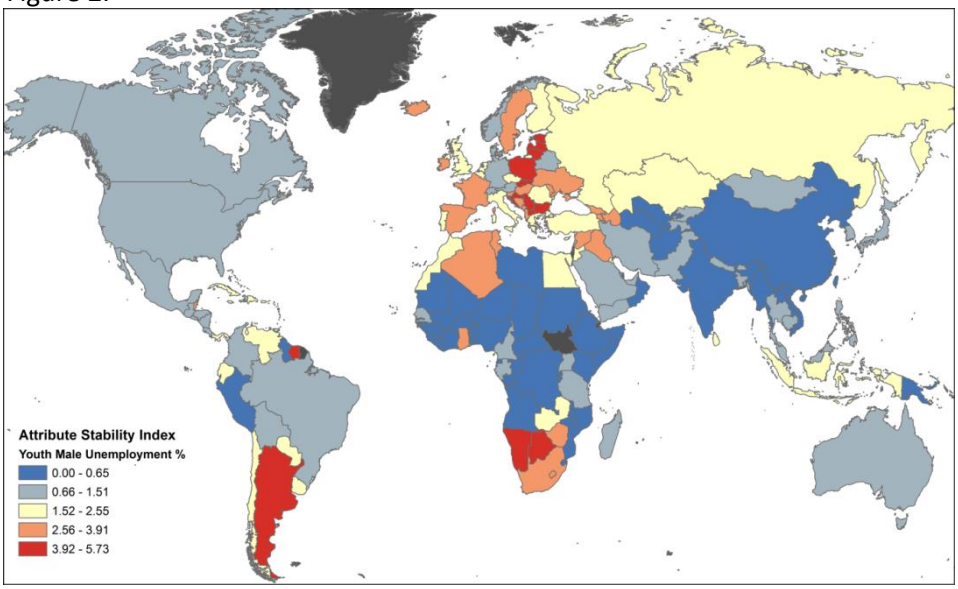

Figure 2. Attribute Stability Index for Youth Male Unemployment (1991-2014).

Spatial patterns immediately emerge including a large cluster of low ASI scores in central Africa and high scores in Eastern Europe. As for the example countries from above, Botswana falls into the highest break along with neighbouring Namibia, France is a member of the second highest grouping, while China is in the lowest category, in line with what we expected from inspecting the attribute stability curves. An important note is that the ASI values represent a temporal behaviour, not just a single value of the attribute in question. By summarising temporal behaviour into a single measure it allows the spatial distribution of temporal behaviours to be visualized on a static map, without the use of animation or multiple visualizations.

Since each country now has an ASI value associated with it, these values can be explored with the same techniques as any standard spatially referenced attribute, such as Moran's I, LISA, or K-means clustering. As an example of this, the Getis-Ord Gi* statistic for youth male unemployment ASI is shown in Figure 3.

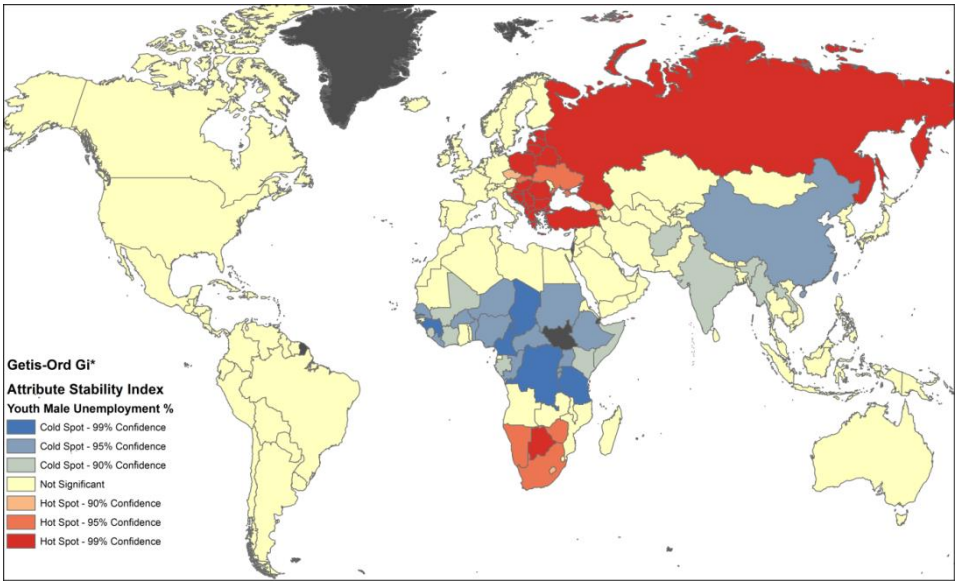

Figure 3. Spatial Clustering of Attribute Stability Index.

3. Conclusion

In this paper, we propose a method for investigating and quantifying instability in time series data, particularly as used in the field of spatio-temporal socio-cultural research where instability is understood to mean widely varying and irregular changes from one observation to the next. This methodology represents an additional tool for exploratory spatial data analysis, principally for spatio-temporal data with limited temporal observations.

4. Acknowledgements

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. Support for this work was provided by the National Geospatial-Intelligence Agency.

5. References

PINCUS, S. M. 1991. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, 88**,** 2297-2301.

PINCUS, S. M. 1995. Approximate entropy (ApEn) as a complexity measure. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 5**,** 110-117.

RICHMAN, J. S. & MOORMAN, J. R. 2000. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, 278**,** H2039-H2049.