Spatio-Temporal Analysis by Frequency Separation: Approach and Research Design

Edward Valachovic[1]
[1]Department of Epidemiology and Biostatistics, School of Public Health, University at Albany, State University of New York, One University Place, Rensselaer, NY 12144

Abstract
When time series data contain frequency specific principal components, including periodic 'signals' such as seasonal or daily cycles, separating these components from interfering frequencies is essential to understand the time and space structures of variation within data. Without properly separating processes operating at different frequencies, statistical analysis can obscure and confound true spatio-temporal relationships. Kolmogorov-Zurbenko (KZ) filters are iterated moving averages and their extensions are well suited to spatio-temporal analysis by frequency separation (STAFS). With guided parameter selection, KZ filters permit finely separating adjacent uncorrelated frequencies and enable analysis of factors within each independent component time scale. This work derives formulas for the separable spectral distance between any two frequencies given data constraints as well as sequence length requirements for frequency separation within research design. Finally, simulations demonstrate proper guided spatio-temporal component frequency separation, the consequences of incomplete signal separation, and effectiveness of this method in spatial, temporal, and spatio-temporal analysis.

Key Words: Spatio – Temporal Analysis Frequency Separation, Periodic Signals Seasonality, Time Series Temporal Spatial, Kolmogorov-Zurbenko Filter, Parameter Selection, Research Design Sample Size Sufficient Sequence Length

1. Introduction

Time Series Analysis, or the observation of data across time, commonly involves variation that exhibits periodicities, or cyclic fluctuations. While periodic changes frequently relate to changes over time, with spatio-temporal data analysis periodic changes can occur in any measured dimension, whether temporal or spatial. These spatio-temporal cycles can result from natural phenomenon, such as seasonal or daily rhythms, to manmade processes such as work weeks. The variation associated with one cycle may be smaller than that associated with different component factors, such as the overall mean trend across time, random variation, system shocks, or other cycles. In the conventional spatial-temporal measurement domain, for example measurements in unit time, each observation is the collective sum of all factors at that time point. Smaller factors can be obscured. However, these sources of variation may operate on different time scales, such as the case with cycles operating at an associated frequency. Therefore, within the frequency domain, the spectral representation of spatio-temporal data provides an opportunity to separate and investigate different time and space scales without the entanglement in the original domain.

Kolmogorov-Zurbenko (KZ) filters and their extensions are able to separate portions of the frequency domain to exclude interfering frequencies [1, 2]. The frequency separated spatio-temporal components can then independently be used to reveal important details about patterns and processes often hidden within the original data, as well as associations with possible factors operating at similar spatio-temporal scales. This is the idea behind spatio-temporal analysis by frequency separation. These filters are used to isolate frequencies in a variety of fields such as the environmental sciences, meteorology, and climatology [3, 4, 5]. They have also been used to separate and model pollution and public health [6, 7]. They can be applied to time series and spatio-temporal data of higher dimensionality [8]. More recently their use was extended to epidemiological surveillance data in a multivariate analysis of the frequency separated uncorrelated components of variables thereby greatly improving model fit [9]. Many of these examples highlight the use of Kolmogorov-Zurbenko filters to smooth data, reduce random variation, interpolate missing observations, and necessarily separate portions of the frequency domain prior to analysis [9,10].

Some of the prior examples address filtering only one portion of the spectrum, or components widely separated in the frequency domain. However, with the increasing use of KZ filters in various research fields and the use of filters to split closely adjacent frequencies or where data is scarce, guidance for minimum data requirements is necessary to guarantee filter performance. First, this study derives a mathematical expression for what number of observations is necessary to separate two different frequencies. Next, this novel development is extended to provide rules for the minimum number of observations, or the sufficient sequence length, necessary to separate any given number of different frequencies. Third, for research design, this work proves the closest that two frequencies may be in an analysis and still be separated given a fixed set of observations. Computer simulations of component signals, combined with strong random errors and missing data are used to model real world unprocessed spatio-temporal data. Finally, these simulations of spatio-temporal analysis by frequency separation (STAFS) illustrate the importance of following the proposition guidelines proven here by comparing results under scenarios of sufficient and insufficient observational data requirements. These demonstrations illustrate applications, limitations, and outcomes of KZ filters to isolate, separate, and reconstruct the signals from the original observed data.

2. Methods

2.1 Statistical Analysis Tools

The Kolmogorov-Zurbenko (KZ) filter is an iteration of a moving average of length m, a positive odd integer [1]. It is a filter with two parameters. The parameter m is the filter window size and k is the number of iterations. KZ filters are low pass filters that strongly attenuate signals of frequency $1/m$ and higher while passing lower frequencies. Applied to a random process $\{X(t): t \in \mathbb{Z}\}$ a KZ filter with m time points, and k iterations is defined as:

Equation 1: Kolmogorov-Zurbenko Filter

$$KZ_{m,k}\big(X(t)\big) = \sum_{u=-k(m-1)/2}^{k(m-1)/2} \frac{a_u^{m,k}}{m^k} X(t+u)$$

The coefficients $a_u^{m,k}$ are the polynomial coefficients from:

$$\sum_{r=0}^{k(m-1)} z^r a_{r-k(m-1)/2}^{m,k} = (1+z+\cdots+z^{m-1})^k$$

One advantage of the KZ filter is the computational ease with which statistical software can apply it in an iterated form. As an iterated application of a moving average filter of *m* time points, *k* times, the Kolmogorov-Zurbenko filter can be produced:

Equation 2: Kolmogorov-Zurbenko filter as an iterated algorithm

$$KZ_{m,1}\big(X(t)\big) = \sum_{u=-(m-1)/2}^{(m-1)/2} \frac{a_u^{m,1}}{m^1} X(t+u) = \frac{1}{m} \sum_{u=-(m-1)/2}^{(m-1)/2} X(t+u)$$

$$KZ_{m,2}\big(X(t)\big) = \frac{1}{m} \sum_{u=-k(m-1)/2}^{k(m-1)/2} KZ_{m,1}\big(X(t+u)\big)$$

$$\vdots$$

$$KZ_{m,k}\big(X(t)\big) = \frac{1}{m} \sum_{u=-k(m-1)/2}^{k(m-1)/2} KZ_{m,k-1}\big(X(t+u)\big)$$

The transfer function is the linear mapping that describes how input frequencies are transferred to outputs. The energy transfer function is the square of the transfer function and as such is symmetric about zero. The energy transfer function of the KZ filter at frequency λ is:

Equation 3: Kolmogorov-Zurbenko energy transfer function

$$|B(\lambda)|^2 = \left(\frac{\sin(\pi m \lambda)}{m \sin(\pi \lambda)}\right)^{2k}$$

The cutoff frequency is a limit or boundary at which the energy transferred through a filter is suppressed or diminished rather than allowed to pass through. A cutoff frequency is used in many fields such as physics, communications, and electrical engineering, and selection depends upon the application. The point where output power is $\alpha \epsilon (0,1)$ times that of the input can be used as the boundary, and it is common to use $\alpha = 1/2$ or the half power point, a power ratio in $10 * log_{10}$ of -3 decibels units. The cutoff frequency, where the transfer function for a KZ filter is [11]:

Equation 4: Kolmogorov-Zurbenko cutoff frequency

$$\lambda_0 \approx \frac{\sqrt{6}}{\pi} \sqrt{\frac{1-(1/2)^{\frac{1}{2k}}}{m^2-(1/2)^{\frac{1}{2k}}}}$$

Where the KZ filter is a low pass filter, strongly filtering signals of a frequency at or above the frequency equivalent to $1/m$, the related Kolmogorov-Zurbenko Fourier Transform (KZFT) filter is a band pass filter. KZFT is a filter applied to a random process

$\{X(t): t \in T\}$ that has parameters m time points, and k iterations but is shifted to center at a frequency ν and is defined:

Equation 5: Kolmogorov-Zurbenko Fourier Transform

$$KZFT_{m,k,\nu}\big(X(t)\big) = \sum_{u=-k(m-1)/2}^{k(m-1)/2} \frac{a_u^{m,k}}{m^k} e^{-i2m\nu u} X(t+u)$$

The coefficients $a_u^{m,k}$ are the polynomial coefficients from:

$$\sum_{r=0}^{k(m-1)} z^r a_{r-k(m-1)/2}^{m,k} = (1 + z + \cdots + z^{m-1})^k$$

Where the KZ filter is symmetric around zero, the KZFT is a symmetric band pass filter around frequency ν. Practical use of the KZFT filter is similar to the KZ filter since it can be produced in statistical software. The energy transfer function of the KZFT filter at a frequency λ with parameters m, k, and ν is given below.

Equation 6: Kolmogorov-Zurbenko Fourier Transform energy transfer function

$$|B(\lambda - \nu)|^2 = \left(\frac{\sin(\pi m(\lambda - \nu))}{m \sin(\pi(\lambda - \nu))}\right)^{2k}$$

It follows that the cut off frequency is:

Equation 7: Kolmogorov-Zurbenko Fourier Transform cutoff frequency

$$|\lambda_0 - \nu| \approx \frac{\sqrt{6}}{\pi}\sqrt{\frac{1 - (1/2)^{\frac{1}{2k}}}{m^2 - (1/2\,)^{\frac{1}{2k}}}}$$

For these filters, the cutoff frequency boundaries then become useful to determine the region of the spectra that is passed and that which is suppressed or filtered.

2.2 Statistical Theory

With a sufficiently large spatio-temporal sequence length or number of observations, *n*, this study proves that two frequencies can be separated by Kolmogorov-Zurbenko (KZ) filters with appropriate chosen filter parameters, so that each frequency is outside of the filter cutoff from the other frequency. In practice this does not mean that different frequencies are separable for any set of data. However, the cutoff frequency can be used to derive a set of conditions necessary so that appropriate KZ filters can be assured of separating frequencies, while minimizing interference between filtered spectral components subject to the limitations of the data. This research derives how many observations are minimally necessary in order to separate two given frequencies and then it details what separation is possible given a certain quantity of data observations.

Proceeding with the outline of the theory for the following propositions and proofs, it is through the choice of filter parameters that control is exercised over the KZ filters, and their extensions such as KZFT. With the goal to separate and filter each of two or more different given frequencies and control the bandwidth of the cutoff frequency, it is possible to create two filters that center or pass one frequency while selecting window size, *m*, so that 1/*m* is less than or equal to one half the separation range or bandwidth between the two given frequencies. The cutoff frequency would then be closer to the central target frequency, thereby attenuating the other target frequency enough so that interference is kept below a predetermined arbitrarily small level controlled. For simplicity, the proofs use the KZFT filter which can center the band-pass filter over a given frequency, and attenuate other frequencies based on the choice of the window size m and number of iterations.

This idea only requires that there is some sufficiently large number of observations. It does not indicate that this is in some way a wise choice for n, or for that matter a practical choice. In practice it is unlikely someone is able to choose any large number n of observations with which to separate frequencies. For this reason, it is interesting to know what a lower limit of n observations that would be necessary to again be certain that given two frequencies, they can be separated outside of filter cutoff boundaries. This necessitates adjusting filter parameters so that bandwidth is not wasted with unnecessarily larger choices of m or k. The target frequencies are unchanged, thus the only parameters remaining are the filter windows size and iterations. Larger numbers of iterations narrow the band-pass filter, but requires higher numbers of observations because time points are discarded from the beginning and end of the available data due missing data outside the filter window time range. The only remaining parameter to adjust is the filter window size. Adjusting the window size of two filters centered at different frequencies so that the respective cutoff frequencies approach but do not overlap should separate with the minimum number of observations required. The following figure illustrates KZFT filters centered over different frequencies and how lowering the choice of m for each filter should decrease the number of observations required while still separating the frequencies. As the filters widen, band-pass regions do not overlap up to the point that cutoff boundaries equal. This attenuates the interference caused by the other frequency with a minimum number of observations.

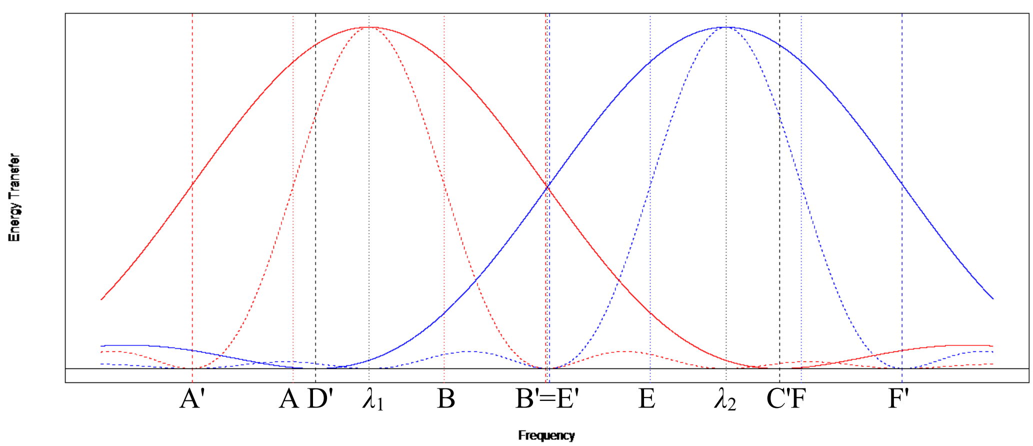

Figure 1: Illustration of 2 different frequencies λ_1 and λ_2, and the reduction of filter window sizes so A and B cutoffs shift to A' and B' from a KZFT filter centered at λ_1 and E and F cutoffs shift to E' and F' from a KZFT filter centered at λ_2. Window size is reduced until cutoff B' equals cutoff E'. C' and D' provide the new frequencies to set window size in the respective KZFT filters.

While time series typically separate signal frequencies in time, there is no difference in signal separation in any spatio-temporal dimension, either a temporal or a spatial dimension. Because of this, the following propositions make no differentiation between time and space scales and the propositions hold in either case. Frequency is equal to one divided by the period, where the period is the length of the cycle in what ever units that dimension happens to be measured.

With this motivation we proceed to the proposition on the theoretical lower limit of the number of spatio-temporal observations necessary to separate two given frequencies so that the cutoff frequencies of KZFT filters are near equal and attenuated regions do not overlap. The first proposition can be considered to provide the minimal required or 'sufficient sequence length' (SSL) of spatio-temporal data for separation. Proofs of propositions and corollaries are included in the Appendix.

Proposition 1: Given $\lambda_i = 1/d_i$ and $\lambda_j = 1/d_j$ different spatio-temporal frequencies where $\lambda_i \neq \lambda_j$, and given k_i, k_j parameters of KZFT filters, where $m_i = m_j = m_{i,j} \equiv$

$$max\left(ceiling\left(\sqrt{(1/2)^{1/2k_i} + \frac{1-(1/2)^{1/2k_i}}{\frac{\pi^2}{6}\left(\frac{|\lambda_i-\lambda_j|}{2}\right)^2}}\right), ceiling\left(\sqrt{(1/2)^{1/2k_j} + \frac{1-(1/2)^{1/2k_j}}{\frac{\pi^2}{6}\left(\frac{|\lambda_j-\lambda_j|}{2}\right)^2}}\right)\right)$$

if $\left|\frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2k_i}}{m_{i,j}^2-(1/2)^{1/2k_i}}}\right| \leq \frac{|\lambda_i-\lambda_j|}{2}$ and $\left|\frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2k_j}}{m_{i,j}^2-(1/2)^{1/2k_j}}}\right| \leq \frac{|\lambda_i-\lambda_j|}{2}$ then

$$n_{i,j} \geq max\left(m_{i,j}, ceiling\left(\frac{1}{\lambda_i}\right), ceiling\left(\frac{1}{\lambda_j}\right)\right).$$

One immediate corollary to the first proposition provides the rule for the sufficient sequence length, or minimum *n*, needed to separate any finite number, *h*, of different frequencies.

Corollary 1: Given h different spatio-temporal frequencies, $\lambda_1 < \ldots < \lambda_h$, and $k_1, k_2, \ldots, k_h$ given parameters of KZFT filters, where $m_{i,j} \equiv$

$$max\left(ceiling\left(\sqrt{(1/2)^{1/2k_i} + \frac{1-(1/2)^{1/2k_i}}{\frac{\pi^2}{6}\left(\frac{|\lambda_i-\lambda_j|}{2}\right)^2}}\right), ceiling\left(\sqrt{(1/2)^{1/2k_j} + \frac{1-(1/2)^{1/2k_j}}{\frac{\pi^2}{6}\left(\frac{|\lambda_j-\lambda_j|}{2}\right)^2}}\right)\right)$$

for $i < h$ and $j = i+1$, if $\left|\frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2k_i}}{m_{i,j}^2-(1/2)^{1/2k_i}}}\right| \leq \frac{|\lambda_i-\lambda_j|}{2}$ and $\left|\frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2k_j}}{m_{i,j}^2-(1/2)^{1/2k_j}}}\right| \leq \frac{|\lambda_i-\lambda_j|}{2}$

for all $i < h$ and $j = i+1$, then

$$n \geq max\left(\{m_{i,j} \mid i < h \text{ and } j = i+1\}, \left\{ceiling\left(\frac{1}{\lambda_1}\right), \ldots, ceiling\left(\frac{1}{\lambda_h}\right)\right.\right).$$

Proposition 1 provides guidance for the smallest *n* possible to separate two given frequencies with KZ filters and Corollary 1 extends this describing the SSL, *n*, to separate any finite number of different frequencies with KZ filters. Generally, much larger numbers of observations are desired to more accurately represent the spatio-temporal patterns in data. In practice, it is often the case that *n* is not chosen but is fixed with the data available. Waiting for additional future observations to be recorded to extend the dataset may not be practical or possible. A subsequent question therefore is, with a fixed spatio-temporal sequence length, n, what is the closest that two frequencies may be and still be separated with given KZ filters. This proposition can be considered as the rule or guideline to provide the minimum separable frequency difference between any number of frequencies.

Proposition 2: If n is the given number of observations, and $\lambda_i = 1/d_i$ and $\lambda_j = 1/d_j$ are two frequencies so that $\frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2k_i}}{m_i{}^2-(1/2)^{1/2k_i}}} \le \frac{|\lambda_1-\lambda_2|}{2}$, $i = 1, 2$ where m_1, k_1, λ_1 and m_2, k_2, λ_2 are parameters of KZFT filters, then $|\lambda_1 - \lambda_2| \ge \frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2k_1}}{n^2-(1/2)^{1/2k_1}}} + \frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2k_2}}{n^2-(1/2)^{1/2k_2}}}$.

Recall that the propositions extend to both temporal and spatial data as well as higher dimensioned mixed spatio-temporal frameworks. Spatial frequency, or the reciprocal of distance (1/*d*), is suitable for the more common temporal frequency often used in research.

The results of these propositions provide for the robust application of Kolmogorov-Zurbenko filters and their extensions to separate spatio-temporal components in multidimensional time series data. In real world datasets, only with sufficient observations and appropriate choices of KZ parameters can the separation between different frequencies be effective so that time scale components can treated as independent.

3. Simulations

The theoretical conclusions of this study are supported by the use of simulations under assumed conditions and settings comparable to real world spatio-temporal data analysis. Simulations help understand and illustrate the performance of multidimensional Kolmogorov-Zurbenko filters to recover signals from original unprocessed observed data, conditions that may require the separation, isolation and recovery of signals of different frequencies, from a high degree of noise and high missing data rates.

3.1 Simulation Methods

Analysis is performed in R version 3.6.1 statistical software using the KZA and KZFT packages [11,12] with datasets in arrays with two spatial dimensions and one time dimension. Arrays are constructed with 100 x-axis and 100 y-axis spatial units, and 100 time units. These arrays are populated by the sum of two spatio-temporally dependent sine wave signals with different frequencies and spatial patterns, where time and a combination of x and y coordinates determine the phase of the sine wave. The result is a motion picture in time of moving and interacting waves entangled in the time domain. While it is not possible to display the motion picture here, an example sequence of 5 equally spaced time points of one of the spatio-temporally changing signals used to construct the motion picture of moving and interacting entangled waves is seen in the following figure.

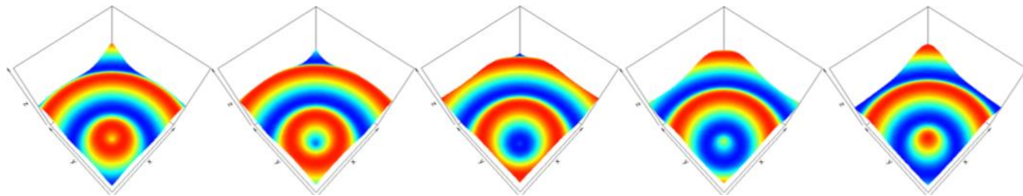

Figure 2: Example sequence of 5 equally spaced time points of one spatio-temporally changing signal used to construct the motion picture of moving and interacting entangled waves.

Next, random variation is introduced by generating equal size arrays of elements randomly selected from a uniform distribution with a range of five times the amplitude of the original signals. These arrays of random variations are then combined with the array of the original pure signals. Finally, each (x, y) coordinate within the array is assigned a uniformly distributed randomly generated number from which a fixed percentage are selected and discarded as missing. This simulates the geographic or spatial scarcity often present in observed data. In this example the chance of being selected missing is 50 percent.

The resulting three dimensional array of data is composed of the pure signals obscured by noise, and then with randomly selected observations discarded as missing data to form the final observed data to be processed. This observed data is all that would be available prior to analysis with no knowledge of the signals. One frame, or time point, of the final observed motion picture of the simulated data array can be seen in the following figure.

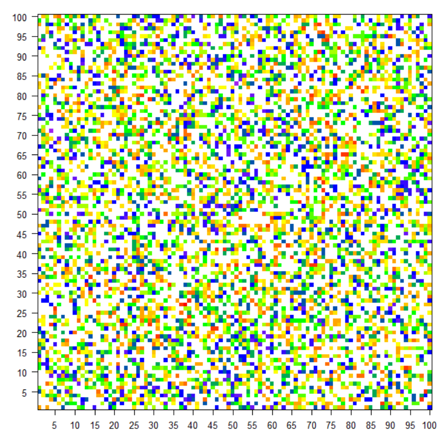

Figure 3: Simulated data of two signals, with noise and missing observations at one example time point.

This simulation design is used in two simulation scenarios to demonstrate the propositions, formulas, and guidance described above for separating frequencies. In each scenario, KZFT filters are centered above the original signal frequency, while choosing parameters to exclude the other frequency outside the cutoff boundary for that filter. A combination of KZFT filters removes the longer period, lower frequency, signal to reconstruct the shorter period, higher frequency, signal. The resulting reconstructed high frequency signal can be compared to the original true high frequency signal initially used in construction of the data. The separation, filtering, and signal reproduction is then repeated with the role of high and low frequency signals reversed. An illustrative visual comparison is made at one sample time point in the figures of the Results section.

The signals in the first simulation, 'Scenario 1: Frequency Separation with Insufficient Sequence Length', have frequencies 0.020 and 0.025, a frequency separation of 0.005. In the second simulation, 'Scenario 2: Frequency Separation with Sufficient Sequence Length' the frequencies are farther apart at frequencies 0.010 and 0.025, a separation of 0.015. With this data, KZ filters with identical methodologies attempt to smooth and separating the two signals within the two scenarios. To illustrate the importance of the propositions and performing frequency separation with a sufficient sequence length, the results are compared showing the effect of signal separation with and without a spatio-temporal SSL for the desired frequencies to be separated.

3.2 Assessment of the Quality of Fit

Correlation is a normalized measure of the association between random processes. Correlation measures the similarity in how one random process varies in time relative to a different process. It assumes a value between -1, implying perfect negative correlation, and positive 1, implying perfect positive correlation between random variables. Zero implies that the two random variables are not correlated.

The coefficient of determination, calculated as the square of correlation, is a measure that indicates the quality of fit of a given time series to another by the fraction of the variance of one that is explained by another. In classical statistics, particularly linear regression through ordinary least squares, a typical assumption is that observations are independent and identically distributed. Time series processes are unlikely to be independent, violating these assumptions, but the use of R^2 for time series does not require the assumption of independence of observations and is mathematically identical in calculation to that in classical statistics. This means that functions for calculating the R^2 provided in statistical software can be used in time series, with care to interpret it as a measure of the goodness of fit between two time series. $R^2 * 100$ gives the percentage of variance of one time series that is explained by another.

Here coefficient of determination, R^2, measured in percent is used after simulation between one reconstructed component against another, as well as against the known true original signals to assess the fit, revealing the ability separate and to reconstruct, respectively.

4. Results

According to Proposition 1, in a time series with *n*=100 observations and with an $\alpha = 0.5$, or half power, the closest two frequencies may be is approximately 0.0062, with Kolmogorov-Zurbenko (KZ) filters having parameters $k_1, k_2 = 2$. We note here one KZ iteration does not completely interpolate all missing data, and more than two iterations require filter windows with wider support than the number of observations given, making two the natural choice. As may be the case, there are times when some parameters are dictated by the particular application or research. A minimum frequency separation of 0.0062 is more than the frequency separation in scenario one, 0.005. This indicates that in the first scenario the frequencies are too close to each other for 100 observations to sufficiently separate them. A minimum frequency separation of 0.0062 is less than that in scenario two, 0.015, indicating 100 observations is a least sufficient. Indeed, according to Proposition 2, a frequency separation of 0.005 should require 125 observations at a minimum. A frequency separation of 0.015 should require 100 observations at a minimum.

What results in the following figures after filtration and signal reconstruction are images that have smoothed noise and interpolated missing observations, but in Scenario 1 where there is an insufficient sequence length for a signal separation of 0.005, the two signals are not well separated (Figure 4A and Figure 5A). Both high and low frequencies are still present and somewhat visible, looking like a mix of the true high and low frequency component (Figure 4C and Figure 5C). The filters left the two signals entangled. In the images corresponding to the second scenario, there is improved signal separation (Figure 4B and Figure 5B). When the frequencies are farther apart as in Scenario 2 where there is a SSL for a signal separation of 0.015, the reconstructed higher frequency signal looks increasing like the true high frequency signal and exhibits less remnants of the low frequency signal. When the signal frequencies are very close, given limited data, the filters capture more of the adjacent frequencies including the other interfering signal, resulting in a reconstructed image that is more a blend of the higher and lower frequency signals. When the signals are close there is confusion as to what the reconstructed signal indicates is the true pattern at that frequency. In a real world scenario, where the true signals are not known, the reconstructed signals can easily be mistaken as arising from the other component, or without indicating a given pattern at all (Figure 5A).

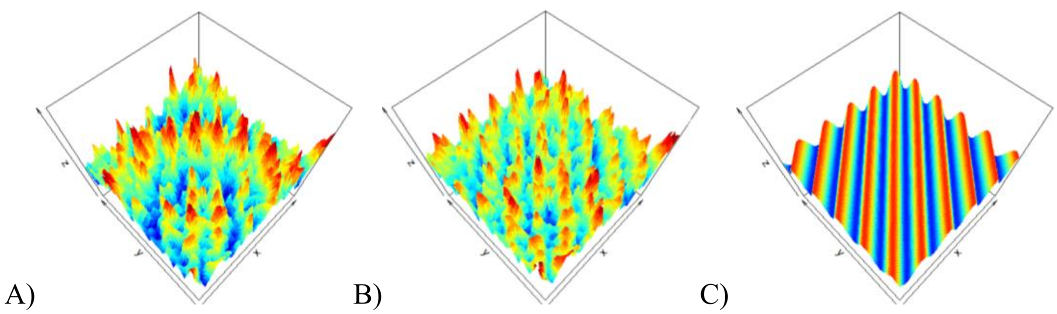

Figure 4: Reconstructed high frequency signals when separation is (A) 0.005, (B) 0.015, and (C) the true high frequency component signal.

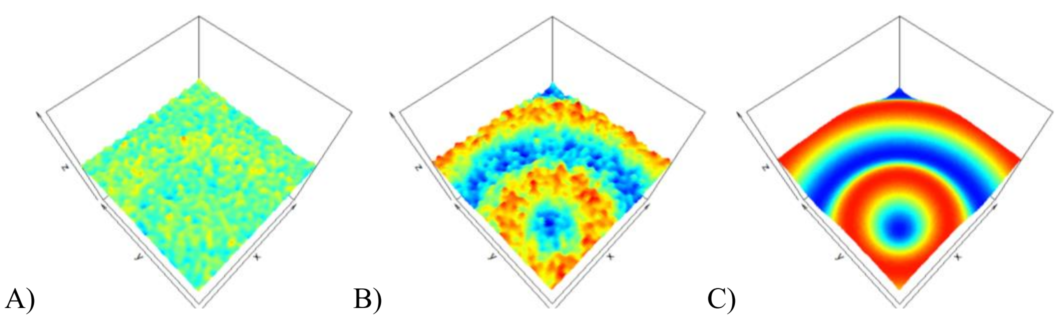

Figure 5: Reconstructed low frequency signals when separation is (A) 0.005, (B) 0.015, and (C) the true low frequency component signal.

The following table displays calculated fit statistics for the models produced. In the first scenario, with a small separation between signals, the reconstructed high and low frequency signal along with the original observed data, with noise and missing observations, are fit against the true component high and low signal. The same is done for the scenario when the signal separation is larger and correlation and coefficient of determination are provided.

Table 1: Correlation and Coefficient of Determination between original, true and recovered signals.

Simulation Scenario 1: 0.005 Frequency Separation with Insufficient Sequence Length		
R^2 * 100 (Percent)	True High	True Low
Original Data	1.76%	1.75%
Recovered High	23.27%	25.73%
Recovered Low	0.80%	11.68%
Simulation Scenario 2: 0.015 Frequency Separation with Sufficient Sequence Length		
Original Data	1.78%	1.82%
Recovered High	28.72%	6.86%
Recovered Low	0.11%	72.58%

Results indicate that in both scenarios, the observed data did not fit either the high or low frequency component well, where the original data explains less than 2% of the variation of each of the component signals. This is not surprising given that the original observed data was composed of both signals but only in the presence of severe noise and missing observations. This illustrates the challenge in the analysis. While in both simulation scenarios the recovered signals were an improvement in fit to their respective targeted true signals, the recovered high and low signals modeled the true high and low frequency signals better when the signal separation was greater. The R^2 as a percentage of variance explained for the recovered high frequency signal improved the correlation from 23.27% to 28.72%, with the true high frequency signal. The success of the recovered low frequency signal was even more striking, with R^2 increasing from 11.68% to 72.58%, and explaining over sixty percent more of the variation in the true low frequency signal. An additional failure caused by not following the proposition guidelines in scenario one, where frequency separation is below the minimum number of required observations, is the recovered signal for a given targeted frequency had higher R^2 fit statistics to the wrong component frequency. In the simulated scenario one of an example of not following the proposition guidelines, the recovered high frequency signal had higher correlation and R^2 with the true low frequency component, 25.73%, than it did with the intended high frequency target signal, 23.27%. This indicates that failing to follow these proposition guidelines likely result in poor signal separation and could be prone to model misspecification.

5. Discussion

This study illustrates the importance of understanding the applications and limitations of Kolmogorov-Zurbenko (KZ) filters and their extensions as well as spatio-temporal analysis by frequency separation (STAFS). It extends the theory of separating component signals by proving propositions to guide what separation may be expected given a set of data, and similarly what data is required to investigate the separation of two or more targeted signals. This helps to understand what questions these tools can answer retrospectively given a set of data, and assists the design of future research that uses this class of filters as an investigatory tool.

Noise exceeding many times the strength of component signals or missing data rates of half or more of all observations may seem impossible obstacles given other statistical analysis techniques. The simulations in this study were not only intended to illustrate the use of KZ filters to handle these difficulties, but were intentionally designed with scenarios chosen to stress KZ methods for signal separation both in line with and outside of guidance from the propositions about frequency separation with a sufficient sequence length. The purposely low number of simulated observations coupled with signals with smaller and then greater separation, as guided by the previous propositions, illustrated signal reconstruction where data was theoretically inadequate verses minimally sufficient. Correlation analysis between the reconstructed signals and the original component signal in both scenarios indicates the superiority of KZ filters where there is sufficient data to more effectively separate the given signals. The assessment of the quality pf fit also indicated the negative consequences of not following the proposition guidelines. Besides poor quality signal reconstruction fit, there was the potential for misidentification of signals and model misspecification. In practice, far greater numbers of observations are desirable, several multiples of the longest signal period. Still, in these challenging simulated conditions above the successful signal reconstruction and quality of fit was visibly and measurably evident when following the proposition guidance.

The use of KZ filters has increased as demand increases to meet statistical analysis challenges such as multidimensional spatial and temporal data analysis, large random errors, high rates of missing observations, signal interference, and situations where other statistical analysis methods are inadequate. As these statistical analysis tools find wider use in a variety of scientific fields the theoretical results discussed here are necessary to ensure performance in spatio-temporal analysis by frequency separation.

References

1. Zurbenko, I. (1986). The Spectral Analysis of Time Series. North-Holland. 248 pages
2. Yang, W., & Zurbenko, I. (2010). Kolmogorov-Zurbenko Filters. Computational Statistics, 2(3), 340-351. DOI: 10.1002/wics.71
3. Wise, E. K., & Comrie, A. C. (2005). Extending the Kolmogorov–Zurbenko filter: application to ozone, particulate matter, and meteorological trends. Journal of the Air & Waste Management Association, 55(8), 1208-1216. DOI: 10.1080/10473289.2005.10464718
4. Kang, D., Hogrefe, C., Foley, K. L., Napelenok, S. L., Mathur, R., & Rao, S. T. (2013). Application of the Kolmogorov–Zurbenko filter and the decoupled direct 3D method for the dynamic evaluation of a regional air quality model. Atmospheric environment, 80, 58-69. https://doi.org/10.1016/j.atmosenv.2013.04.046
5. Anh, V., Duc, H., & Azzi, M. (1997). Modeling anthropogenic trends in air quality data. Journal of the Air & Waste Management Association, 47(1), 66-71. DOI: 10.1080/10473289.1997.10464406
6. Valachovic, E., & Zurbenko, I. (2014). Skin Cancer, Irradiation, and Sunspots: The Solar Cycle Effect. Biomed Red Int 2014:1-9. doi:10.1155/2014/538574.
7. Gupta, I., Salunkhe, A., & Kumar, R. (2010). Modelling 10-year trends of PM 10 and related toxic heavy metal concentrations in four cities in India. Journal of hazardous materials, 179(1), 1084-1095. https://doi.org/10.1016/j.jhazmat.2010.03.117
8. Zurbenko, I., & Ming, L. (2017). KZ Spatial Waves Separations. Journal of Research in Applied Mathematics, 3(4), 1-7.
9. Valachovic, E. & Zurbenko, I. (2017). Multivariate analysis of spatial–temporal scales in melanoma prevalence. Cancer Causes Control, 28: 733. https://doi.org/10.1007/s10552-017-0895-x
10. Tsakiri, K., & Zurbenko, I. (2013). Explanation of fluctuations in water use time series. Environ Ecol Stat 20: 399. https://doi.org/10.1007/s10651-012-0225-0.
11. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.
12. Close, B., & Zurbenko, I. (2013). Kolmogorov-Zurbenko adaptive filters. (Version 3). Retrieved from http://cran.r-project.org/web/packages/kza/index.html.

Appendix

Proposition 1: Given $\lambda_i = 1/d_i$ and $\lambda_j = 1/d_j$ different spatio-temporal frequencies where $\lambda_i \neq \lambda_j$, and given k_i, k_j parameters of KZFT filters, where $m_i = m_j = m_{i,j} \equiv$

$$max\left(ceiling\left(\sqrt{(1/2)^{1/2k_i} + \frac{1-(1/2)^{1/2k_i}}{\frac{\pi^2}{6}\left(\frac{|\lambda_i-\lambda_j|}{2}\right)^2}}\right), ceiling\left(\sqrt{(1/2)^{1/2k_j} + \frac{1-(1/2)^{1/2k_j}}{\frac{\pi^2}{6}\left(\frac{|\lambda_j-\lambda_j|}{2}\right)^2}}\right)\right)$$

if $\left|\frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2k_i}}{m_{i,j}^2-(1/2)^{1/2k_i}}}\right| \leq \frac{|\lambda_i-\lambda_j|}{2}$ and $\left|\frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2k_j}}{m_{i,j}^2-(1/2)^{1/2k_j}}}\right| \leq \frac{|\lambda_i-\lambda_j|}{2}$ then

$n_{i,j} \geq max\left(m_{i,j}, ceiling\left(\frac{1}{\lambda_i}\right), ceiling\left(\frac{1}{\lambda_j}\right)\right)$.

Proof of Proposition 1: Define $c = \frac{\pi^2}{6}\left(\frac{|\lambda_1-\lambda_2|}{2}\right)^2$. For a given *k* it follows that:

$$m = \sqrt{(1/2)^{1/2k} + \frac{1-(1/2)^{1/2k}}{\frac{\pi^2}{6}\left(\frac{|\lambda_1-\lambda_2|}{2}\right)^2}} \Leftrightarrow m = \sqrt{(1/2)^{1/2k} + \frac{1-(1/2)^{1/2k}}{c}}$$

$$\Leftrightarrow m^2 = (1/2)^{1/2k} + \frac{1-(1/2)^{1/2k}}{c} \Leftrightarrow m^2 = \frac{c(1/2)^{1/2k} + 1 - (1/2)^{1/2k}}{c}$$

$$\Leftrightarrow cm^2 = c(1/2)^{1/2k} + 1 - (1/2)^{1/2k} \Leftrightarrow cm^2 - c(1/2)^{1/2k} = 1 - (1/2)^{1/2k}$$

$$\Leftrightarrow c = \frac{1-(1/2)^{1/2k}}{m^2-(1/2)^{1/2k}} \Leftrightarrow \frac{\pi^2}{6}\left(\frac{|\lambda_1-\lambda_2|}{2}\right)^2 = \frac{1-(1/2)^{1/2k}}{m^2-(1/2)^{1/2k}}$$

$\Leftrightarrow \frac{|\lambda_1-\lambda_2|}{2} = \frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2k}}{m^2-(1/2)^{1/2k}}}$, the cutoff for a filter with parameters *m* and *k*.

Then $m_i = m_j = m_{i,j} \equiv$

$$max\left(ceiling\left(\sqrt{(1/2)^{1/2k_1} + \frac{1-(1/2)^{1/2k_1}}{\frac{\pi^2}{6}\left(\frac{|\lambda_1-\lambda_2|}{2}\right)^2}}\right), ceiling\left(\sqrt{(1/2)^{1/2k_2} + \frac{1-(1/2)^{1/2k_2}}{\frac{\pi^2}{6}\left(\frac{|\lambda_1-\lambda_2|}{2}\right)^2}}\right)\right)$$

has $\frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2k_i}}{m_{i,j}^2-(1/2)^{1/2k_i}}} \leq \frac{|\lambda_1-\lambda_2|}{2}$ and $\frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2k_j}}{m_{i,j}^2-(1/2)^{1/2k_j}}} \leq \frac{|\lambda_1-\lambda_2|}{2}$. By definition of KZ filters, $m_i \leq n$ for all *i*, thus $n_{i,j} \geq max\left(m_{i,j}, ceiling\left(\frac{1}{\lambda_i}\right), ceiling\left(\frac{1}{\lambda_j}\right)\right)$. ■

Corollary 1: Given h different spatio-temporal frequencies, $\lambda_1 < \ldots < \lambda_h$, and $k_1, k_2, \ldots, k_h$ given parameters of KZFT filters, where $m_{i,j} \equiv$

$$max\left(ceiling\left(\sqrt{(1/2)^{1/2k_i} + \frac{1-(1/2)^{1/2k_i}}{\frac{\pi^2}{6}\left(\frac{|\lambda_i-\lambda_j|}{2}\right)^2}}\right), ceiling\left(\sqrt{(1/2)^{1/2k_j} + \frac{1-(1/2)^{1/2k_j}}{\frac{\pi^2}{6}\left(\frac{|\lambda_j-\lambda_j|}{2}\right)^2}}\right)\right)$$

for $i < h$ and $j = i + 1$, if $\left|\frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2k_i}}{m_{i,j}^2-(1/2)^{1/2k_i}}}\right| \leq \frac{|\lambda_i-\lambda_j|}{2}$ and $\left|\frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2k_j}}{m_{i,j}^2-(1/2)^{1/2k_j}}}\right| \leq \frac{|\lambda_i-\lambda_j|}{2}$ for all $i < h$ and $j = i + 1$, then $n \geq max\left(\{m_{i,j} | \ i < h \text{ and } j = i + 1\}, \{ceiling\left(\frac{1}{\lambda_1}\right), \ldots, ceiling\left(\frac{1}{\lambda_h}\right)\right)$.

Proof of Corollary 1: Without loss of generality, we can assume that $\lambda_1 < \lambda_2 < \cdots < \lambda_h$. Assuming the givens, by Proposition 1 we have

$n_{i,j} \geq max\left(m_{i,j}, ceiling\left(\frac{1}{\lambda_i}\right), ceiling\left(\frac{1}{\lambda_j}\right)\right)$ so that $\left|\frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2k_i}}{m_{i,j}^2-(1/2)^{1/2k_i}}}\right| \leq \frac{|\lambda_i-\lambda_j|}{2}$ and $\left|\frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2k_j}}{m_{i,j}^2-(1/2)^{1/2k_j}}}\right| \leq \frac{|\lambda_i-\lambda_j|}{2}$ for all $i < h$ and $j = i + 1$. Then

$$n \geq max(n_{1,2}, n_{2,3}, \ldots, n_{h-1,h})$$

$$= max\left(\left\{m_{1,2}, ceiling\left(\frac{1}{\lambda_1}\right), ceiling\left(\frac{1}{\lambda_2}\right)\right\}, \ldots, \left\{m_{h-1,h}, ceiling\left(\frac{1}{\lambda_{h-1}}\right), ceiling\left(\frac{1}{\lambda_h}\right)\right\}\right)$$

$$= max\left(\{m_{i,j} | \ i < h \text{ and } j = i + 1\}, \{ceiling\left(\frac{1}{\lambda_1}\right), \ldots, ceiling\left(\frac{1}{\lambda_h}\right)\right) \ \blacksquare$$

Proposition 2: If *n* is the given number of observations, and $\lambda_i = 1/d_i$ and $\lambda_j = 1/d_j$ are two frequencies so that $\frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2k_i}}{{m_i}^2-(1/2)^{1/2k_i}}} \leq \frac{|\lambda_1-\lambda_2|}{2}$, $i = 1, 2$ where m_1, k_1, λ_1 and m_2, k_2, λ_2 are parameters of KZFT filters, then $|\lambda_1 - \lambda_2| \geq \frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2k_1}}{n^2-(1/2)^{1/2k_1}}} + \frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2k_2}}{n^2-(1/2)^{1/2k_2}}}$.

Proof of Proposition 2: Assume the given statements. By definition a KZFT filter has $m_i \leq n, i = 1, 2$ so it follows ${m_i}^2 - (1/2)^{1/2k_i} \leq n^2 - (1/2)^{1/2k_i}$ and the cutoff frequency $\frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2k_i}}{n^2-(1/2)^{1/2k_i}}} \leq \frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2k_i}}{m_i^2-(1/2)^{1/2k_i}}}$. Then, $\frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2k_1}}{n^2-(1/2)^{1/2k_1}}} + \frac{\sqrt{6}}{\pi}\sqrt{\frac{1-(1/2)^{1/2k_2}}{n^2-(1/2)^{1/2k_2}}} \leq$

$$\frac{\sqrt{6}}{\pi}\sqrt{\frac{1-\left(\frac{1}{2}\right)^{1/2k_1}}{{m_1}^2-\left(\frac{1}{2}\right)^{1/2k_1}}} + \frac{\sqrt{6}}{\pi}\sqrt{\frac{1-\left(\frac{1}{2}\right)^{1/2k_2}}{{m_2}^2-\left(\frac{1}{2}\right)^{1/2k_2}}} \leq \frac{|\lambda_1-\lambda_2|}{2} + \frac{|\lambda_1-\lambda_2|}{2}$$
$$= |\lambda_1-\lambda_2| \blacksquare$$