Indian J.Sci.Res. 16 (1): 151-159, 2017 ISSN: 0976-2876 (Print)
ISSN: 2250-0138(Online)

AN INTRODUCTION TO TIME SERIES ROTTING LONG-TERM AND OTHER TRENDS OF WASTE WATER CHEMICAL IN RIHAND DAM

DILEEP KUMAR PANDEY[1]

Department of Management, ABRPG College, Anpara , Sonbhadra.U.P.,India

A recent paper published time series of concentrations of chemicals in drinking water collected from the bottom of Rihand Dam, a major Indian water supply reservoir. Data were compared to water level using only linear regression. This creates an opportunity for students to analyze these data further. This article presents a structured introduction to time series decomposition that compares long-term and seasonal components of a time series of a single chemical (meprobamate) with those of two supporting datasets (reservoir volume and specific conductance). For the chemical data, this must be preceded by estimation of missing datum points. Results show that linear regression analyses of time series data obscure meaningful detail and that specific conductance is the important predictor of seasonal chemical variations. To learn this, students must execute a linear regression, estimate missing data using local regression, decompose time series, and compare time series using cross-correlation. Complete R code for these exercises appears in the supplementary information. This article uses real data and requires that students make and justify key decisions about the analysis. It can be a guided or an individual project. It is scalable to instructor needs and student interests in ways that are identified clearly in this article.

KEYWORDS: Cross-correlation, Local regression, Water quality, Waste water contaminant

During the first decade of the 21st century, many water researchers studied the occurrence and treatment of pharmaceuticals, personal care products, and other organic chemicals in natural waters, wastewater, and drinking water. These chemicals captured widespread attention because of their potential to disrupt human endocrine function, because they arouse disgust in water consumers, and because analytical chemistry technologies advanced so that these chemicals could be measured reliably at low concentrations . we found that, as the water level of Rihand DAM declined during the years 2003–2007, inclusive, the concentration of a group of pharmaceuticals and other endocrine-disrupting chemicals (e.g., the herbicide atrazine) steadily increased. This led the authors to suggest that less water in Rihand Dam implied less dilution of chemicals discharged to it and that this was evidence of climate change altering water quality, since the decreasing water level of Rihand Dam can be partially attributed to climate change reached their conclusions regarding the opposite five-year trends of chemical concentrations and water level by conducting two separate linear regressions on nonlinear time series data. From an environmental chemistry and water management standpoint, these authors contributed valuable data to the literature and offered a reasonable interpretation of those data. However, their limited statistical analyses offer students a valuable opportunity to explore additional richness in a real, published, analytically costly, and practically relevant dataset. Structured investigation of this and related datasets provides a rigorous introduction to time series decomposition, which is described fully by Yafee and McGee(2000Yafee, R.A., and McGee, M.(2000), Intro duction to Time Series Analysis and Forecasting: with Applications of SAS and SPSS and references therein. Doing so also requires an introduction to local regression and cross-correlation as well as critical thinking by students regarding statistical decision making. These techniques, their coding in R, and the presentation of figures that they generate are the focus of this article.

The author uses this exercise at the end of the time series unit of his Data Science B: Time Series and Multivariate Statistics course. Students come to this course having had either Statistics 1: The Practice of Statistics, AP Statistics in high school, or a strong quantitative background. The first major unit of Data Science B reviews prerequisite content in an applied context. It covers hypothesis testing, comparisons of related datasets, and linear regression. It also introduces students to writing scripts in R. Students in the course are usually either upper-year undergraduates who expect to generate significant data in their undergraduate theses or students focusing on applied math. The author's institution does not have majors, but nonmath students in this course tend to focus on ecology, health sciences, environmental science, or data science.

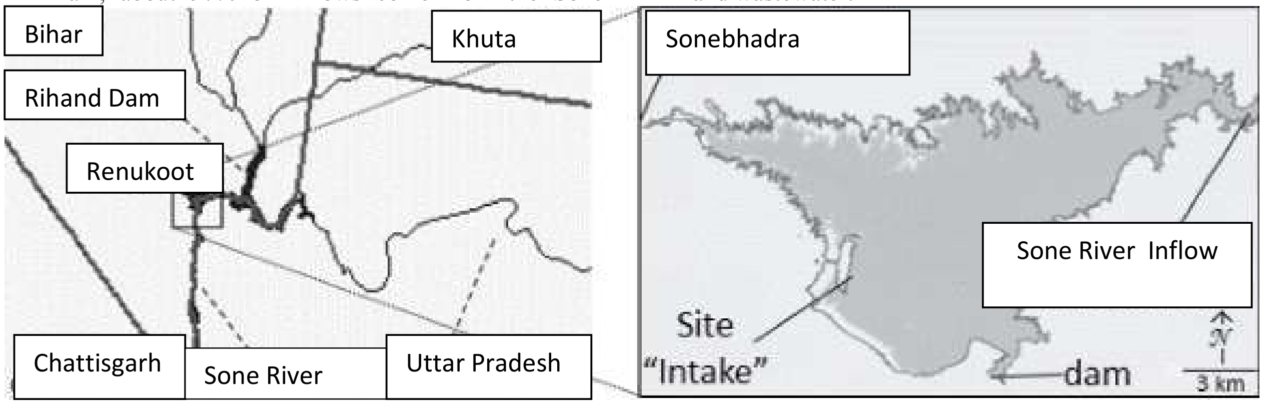
THE DATASETS AND THE STORY

Background

The Rihand River which is the tributary of the Son River flows more than 160 km from the Rocky Mountains to the Matiranga hills. The majority of its water comes from mountain near its headwaters. Much of its watershed is arid. Major infrastructure projects allow diversion of water from the river to agricultural and municipal users. This consumptive use of water is highly regulated.More water has been promised to users than flows down the Son River in an average year, and so the water storage of reservoirs on the river has decreased in most years of the 21st century. This unsustainable situation is exacerbated by an on-going, long-term decrease in annual flow volumes of the Son River, a phenomenon that may be caused by climate change

[1] **Corresponding Author**

Chattisgarh draws all its sone River water from a location deep in Rihand Dam, the large reservoir formed by the NTPC , and distributes it to municipal and industrial users in Uttar Pradesh ,Bihar, Delhi, Telangana, Tamil Nadu and Karanataka.This accounts for less than 2% of the inflows to Rihand Dam; about 97% of inflows come from the Sone River. Therefore, a small fraction of the water taken by Southern Chattisgarh for municipal supply is highly diluted former wastewater. Because of this, the Southern Chattisgarh Water Authority (SCWA), which coordinates seven utilities in southern Chattisgarh, has developed extensive expertise in the treatment of water and wastewater.



The SCWA has developed particular expertise in the measurement of wastewater-derived organic chemicals (WDOCs). These chemicals, which range from pharmaceuticals to personal care products (e.g., shampoo ingredients or insect repellent) to pesticides ingested with food, exist in treated wastewater because wastewater treatment facilities are not designed to remove them (e.g., Sedlak 2014Sedlak, D. (2014), *Water 4.0*. These chemicals have been measured in a variety of aquatic systems, such as groundwater. However, the SCWA practice of discharging treated wastewater to its drinking water source creates the possibility that concentrations of WDOCs in Rihand Dam could increase over time despite considerable dilution because chemicals could be added to water more rapidly than they are degraded as water makes successive passes through the urban system.

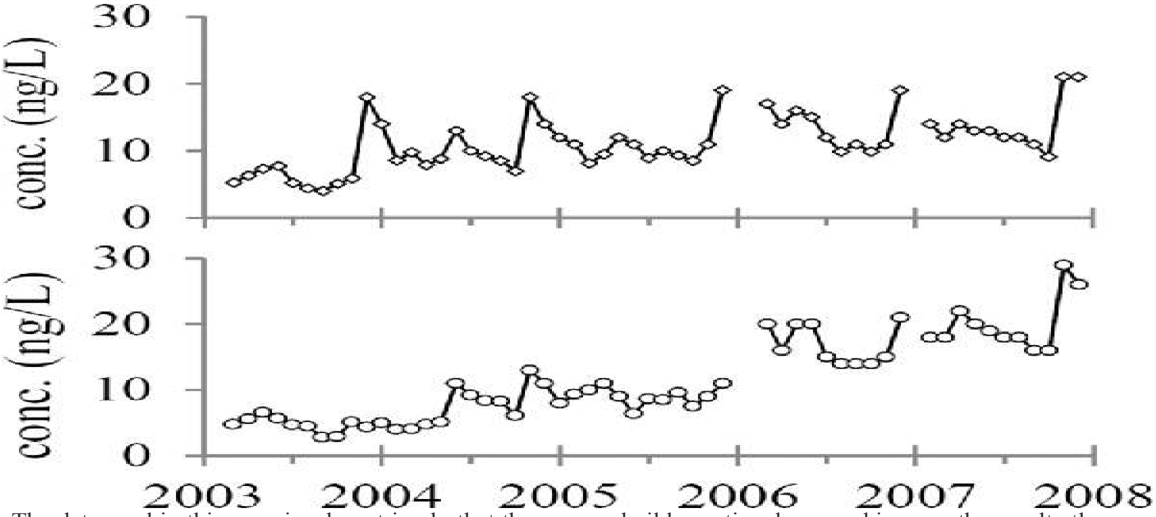
Most studies of WDOCs in natural waters characterize their occurrence; few have explored connections between their concentrations and hydrological processes. They described repeated sampling at a specific deep-water location (the intake structure for the Southern Chattisgarh water supply system), reported time series data of the summed concentration of the WDOCs analyzed, and evaluated trends over five years (2003–2007, inclusive) using linear regression. After controlling for other variables and possible explanations, they reported a long-term, significant, inverse relationship between water level and concentration of WDOCs in Rihand Dam. They attribute decreasing water level to climate change and thus suggest a link between this phenomenon and water quality for a major city.

However, analyzing their data only with linear regression and exploring only long-term trends does not exploit the full richness of their dataset. Time series analysis provides a useful tool to explore the data collected in much greater detail, and the data contain multiple complexities that enable a thorough introduction to time series for students. This exercise is designed to occur after an introduction to time series decomposition and local regression. It can be used as a thorough introduction to of time series analysis before students broach the complex topics of time series modeling and forecasting.

Data

The primary dataset in this investigation consists of monthly concentrations of a suite of WDOCs collected took care to divide their chemical data into two sets because of an improvement in their analytical technique at the end of 2005 that allowed lower detection limits for most chemicals. Because they reported the sum of the 16 chemical concentrations that were measured above detection limits, this value increases after the change in analytical method. However, two chemicals, meprobamate (an antianxiety drug) and sulfamethoxazole (an antibiotic) are measured above detection limits in every sample using both methods. As a result, they are unaffected by the change in analytical method, and so their concentrations can be treated as single datasets with three missing datum points. To focus on data analysis rather than chemistry, this exercise analyzes only the meprobamate time series (subsequently referred to as the "chemical time series"). The sulfamethoxazole time series, which leads to similar yet not identical results, can also be studied. Meprobamate concentrations range from 4 to 21 ng/L with a median of 11 ng/L and an interquartile range (IQR) of 9–14 ng/L, and sulfamethoxazole concentrations range from 3 to 29 ng/L with a median of 10 ng/L and an IQR of 6–16 ng/L (Figure 2).

Figure 2. Observations of meprobamate (upper panel) and sulfamethoxazole (lower panel) concentrations measured in untreated water sampled from the drinking water intake in Rihand Dam.



The data used in this exercise do not imply that the water supply of the people of Southern Chattisgarh has ever been in danger of exceeding health standards or the best practices of the water supply industry. All water samples discussed here were measured before that water was treated for public consumption.

Research Tasks

This investigation focuses on five research tasks:

1. Reproduce the linear regressions and create an overlay graph of observations and their linear fit.

2. Estimate values for missing data in the chemical time series using local regression.

3. Decompose the chemical time series and assess goodness of fit.

4. Decompose the volume time series and compare its long-term trend to that of the chemical time series.

5. Decompose the SC time series. Compare the seasonal trend of the chemical data to the seasonal trends of the volume and SC time series using both overlay plots and cross-correlation.

Completing these tasks allows for the reproduction of the analysis conducted by the authors who collected the dataset, the demonstration of the shortcomings of fitting a linear model to seasonally variant time series data, and the use of time series decomposition to provide a more nuanced understanding of the data.

Helpful hint

The instructor might choose to have students act as reviewers for the manuscript as a lead-in to this exercise. This can help students see why time series analyses are necessary, despite the ease of applying simpler techniques like linear regression inappropriately. It may also help them build emotional ownership over the results they produce in this exercise. When students have completed their reviews, a group discussion can be used to generate the list of tasks presented here with the instructor steering the group to create a complete list and perhaps improvising to allow students to explore any other analyses that they suggest.

Potential pitfall

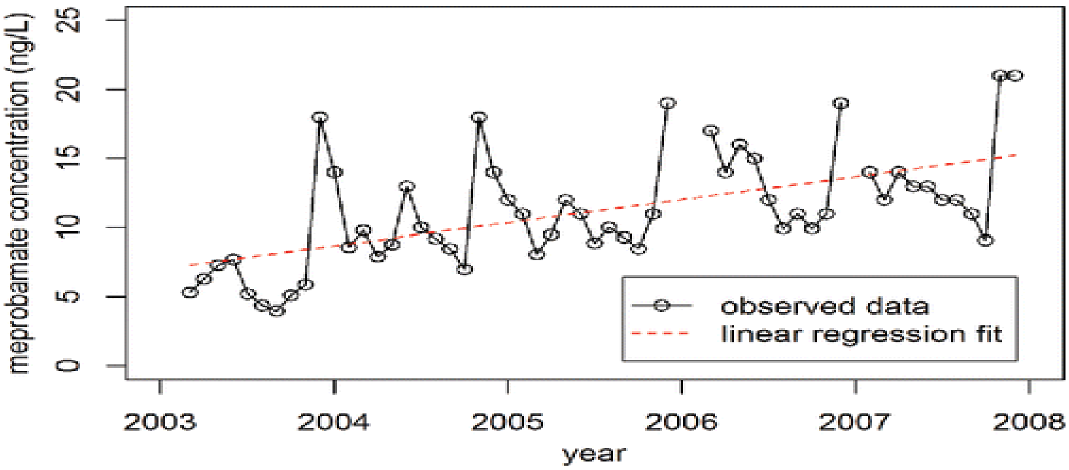
In the experience of the author, students often require substantial amounts of time to review papers, especially when this is an entirely new activity to them. This can be because they seek to understand all parts of the content of the paper, which may be new to them. The instructor can facilitate a review assignment by explaining the Methods section carefully, providing a sheet defining strange terms (e.g., those about lakes or chemical names) and introducing all the variables that the students will encounter. Additionally, instructors might enforce a time limit of 2 to 3 hr to ensure that this prelude assignment does not exhaust students before they reach the assignment(s) described in this article, which can be more difficult.

The Analyses and Conclusions

The following subsections explain the statistical analyses that complete the research tasks listed above. Major statistical techniques used are as follows: linear regression, local regression, prediction (or imputation) of missing data, seasonal decomposition of time series by loess, and cross-correlation.

Research Task 1

Regression of the chemical time series as a dependent variadcble on time as an independent variable is fairly straightforward with the lm command in R (all commands described here are for R version. Plotting an overlay of the regression line (as represented by the set of fitted values that exists as an element of the object to which the lm command is assigned) on the original dataset shows the poor fit.Meprobamate concentrations measured in Rihand Dam.



This task can be excluded for groups that have extensive experience in R, but, for other groups, it may be valuable because it introduces students to overlay plots in R (which are enabled by executing the command par(new = T) between plotcommands). Additionally, this task, which is likely conceptually straightforward to students who are studying time series analysis, can be used to verify that students can use R for the basic tasks of reading in data, conducting a linear regression, and plotting actual and fitted values.

Helpful hint

This linear regression presents an opportunity to discuss with students that the viability of a linear regression, especially with regard to time, does not imply its appropriateness. There are two criticisms of the linear regression of chemical concentrations on time presented .First, time series analysis is a much more appropriate approach because of the obvious temporal correlations and the seasonality and long-term trend in these data. Second,Invoke decreasing dilution as an explanation for the increasing concentration over time, since the water level and, implicitly, the volume of water in the reservoir was decreasing during this time. Therefore, a more appropriate linear regression analysis would regress chemical concentrations on Rihand Dam volume (or elevation), since this is the relationship at the core of their interpretation.

Potential Pitfall

If students are not comfortable writing an R script to read in data, execute a linear regression, and plot data, then this will be a very difficult assignment. For inexperienced students, the instructor may choose to divide this assignment into smaller assignments with individual deadlines so that students are forced to finish tasks and get feedback on whether their R code is producing expected results. Students who need hints to create overlay plots or address the error described in the next potential pitfall should be able to complete this assignment, but those who are uncomfortable with the basic concepts of data manipulation and coding in R are likely insufficiently experienced for the subsequent tasks in this assignment.

Potential pitfall: If one completes a linear regression of the chemical time series against time, R will omit by default the three rows that contain "NA" values in the chemical data. Therefore, the fitted element in the object to which the regression was assigned with have 55 rows, not 58 as in the original dataset. This will induce an error when the fitted values are plotted against the "year" column of the original dataset. This error can be avoided by constructing a modified version of the original dataset that does not contain the rows with NA values. There are multiple ways to accomplish this; one is to use the rbind command to join together only the rows in the original dataset that do not contain NAs. Syntax is

```
chemicals.no.NAs<-rbind(chemicals[1:34,],
                        chemicals[37:46,],
                        chemicals[48:58,])
```

where "chemicals" is the object into which the original dataset was read. Complete R code for this step appears in the supplementary information.

Research Task 2: Estimate Values for Missing Data in The Chemical Time Series

Before the chemical time series can be decomposed, the three missing datum points should be estimated because the stl command used for time series decomposition (see below) does not tolerate missing observations in a time series. Several methods are available to replace missing observations in a dataset . The means or medians of either the dataset, the period of the missing observations, or the adjacent observations can be used. One can use linear regression or autoregressive-integrated-moving average (ARIMA) modeling to forecast forward from the previous observation or backward from the following observation. In this dataset, local regression offers a simple and robust alternative. Local regression creates separate linear regression models for each point in a time series to create a smoothed curve through a two-dimensional scatterplot via a more sophisticated method than a running average. Regressions occur in windows that are centered on the points to which the regressions apply. Points in a given window do not contribute equally to the regression calculation; rather, points nearest the center of the window are given a greater weight via a "tricube" weighting function such that

W(xi)=(1−(xi−xxi−xmax)3)3,

W(xi)=(1−(xi−xxi−xmax)3)3,

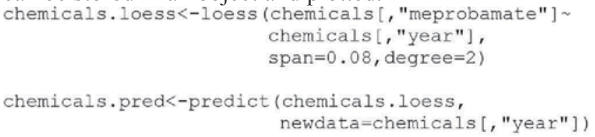
where x is the point at the center of the window in which the regression is performed, x_i is another point in the window, $W(x_i)$ is the weight of point x_i in the calculation of the regression, and $x_{\max}$ = the point at an extreme end of the window.

In addition to common arguments like formula, data, and na.action, the R command for local regression, loess, accepts arguments of span and degree, which control much about the regression model produced by this function. The span is a value greater than 0 that represents the fraction of datum points to be included in the window about a specific point. A value greater than or equal to 1 implies the inclusion of the entire dataset; when span is greater than 1, the weighting function becomesW(xi)=(1−(xi−x(xi−xmax)s(1/p))3)3, W(xi)=(1−(xi−x(xi−xmax)s(1/p))3)3,

where s is the span and p is the number of explanatory variables for the response variable being modeled .

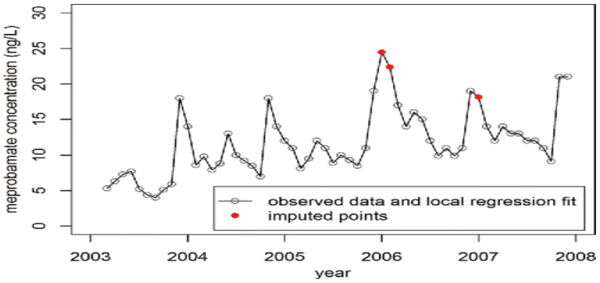
Consequently, larger span values lead to smoother regression curves that do not contain detail that is part of the dataset, and progressively smaller values yield curves that reflect some of the minor variations in the time series.

The degree argument specifies the degree of the regression functions that are created in each window. Possible values are 0, 1, or 2; most time series are modeled well with a degree of 1 .The loess command is often paired with the predictcommand so that a new array of predicted values can be stored in an object and plotted.



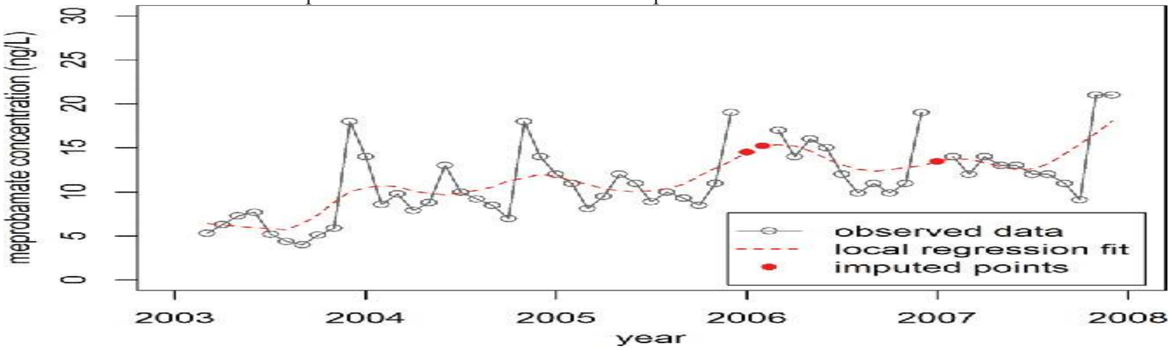
```
chemicals.loess<-loess(chemicals[,"meprobamate"]~
                       chemicals[,"year"],
                       span=0.08,degree=2)

chemicals.pred<-predict(chemicals.loess,
                        newdata=chemicals[,"year"])
```

Plotting the predicted data as an overlay above the original data shows that the fit of this time series can be achieved precisely and that the loess function imputes peak values in this dataset.



Helpful hint

The imputation of the missing data points presents an excellent opportunity to let students make a difficult decision and justify it. If one uses more customary values for the span and degree (e.g., 0.2 and 1), the chemical time series is fit with a smooth curve. Students can rightly argue that they have no way of knowing what the chemical concentrations would have been had water samples been collected and analyzed during the three missing months in 2006 and 2007. Additionally, values of the span less than 0.11 lead to values for the missing 2006 data that exceed all other values in the chemical time series. An argument can be made that one should not use imputed values that fall outside the range of observed values since there is no basis in the dataset for concentrations of that magnitude in water samples from Lake Mead. However, fitting the dataset with the loess and predict commands only leads fitted data to match observed data when the span and degree equal 0.08 and 2, respectively. Thus, one can also argue that using the imputed data that exceed observed values is appropriate because the modeled data match the observed data exactly. This dilemma allows the instructor to reinforce that statistics is often about judgment and critical thinking. The author allows students to continue with their assignments using any imputation of the missing datum points so long as they can explain and defend their choice.

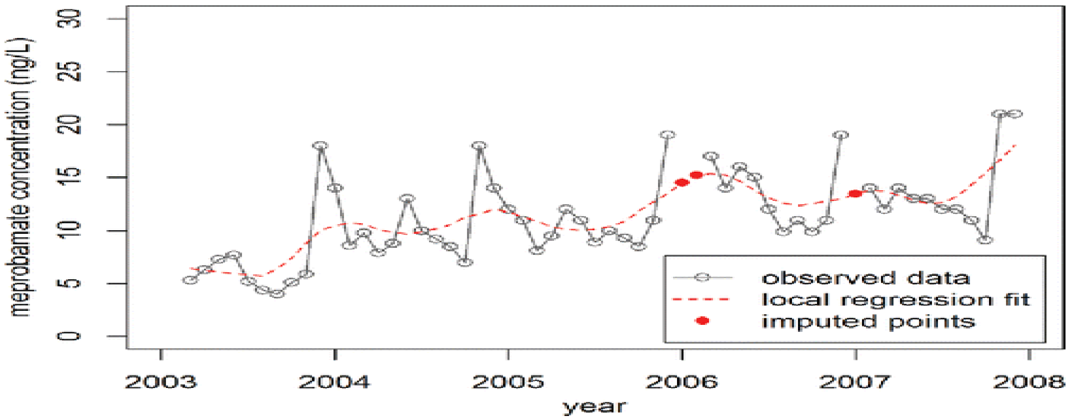


Helpful hint

The imputation of the missing data points presents an excellent opportunity to let students make a difficult decision and justify it. If one uses more customary values for the span and degree (e.g., 0.2 and 1), the chemical time series is fit with a smooth curve. Students can rightly argue that they have no way of knowing what the chemical concentrations would have been had water samples been collected and analyzed during the three missing months in 2006 and 2007. Additionally, values of the span less than 0.11 lead to values for the missing 2006 data that exceed all other values in the chemical time series. An argument can be made that one should not use imputed values that fall outside the range of observed values since there is no basis in the dataset for concentrations of that magnitude in water samples from Rihand Dam. However, fitting the dataset with the loess and predict commands only leads fitted data to match observed data when the span and degree equal 0.08 and 2, respectively. Thus, one can also argue that using the imputed data that exceed observed values is appropriate because the modeled data match the observed data exactly. This dilemma allows the instructor to reinforce that statistics is often about judgment and critical thinking. The author allows students to continue with their assignments using any

imputation of the missing datum points so long as they can explain and defend their choice.

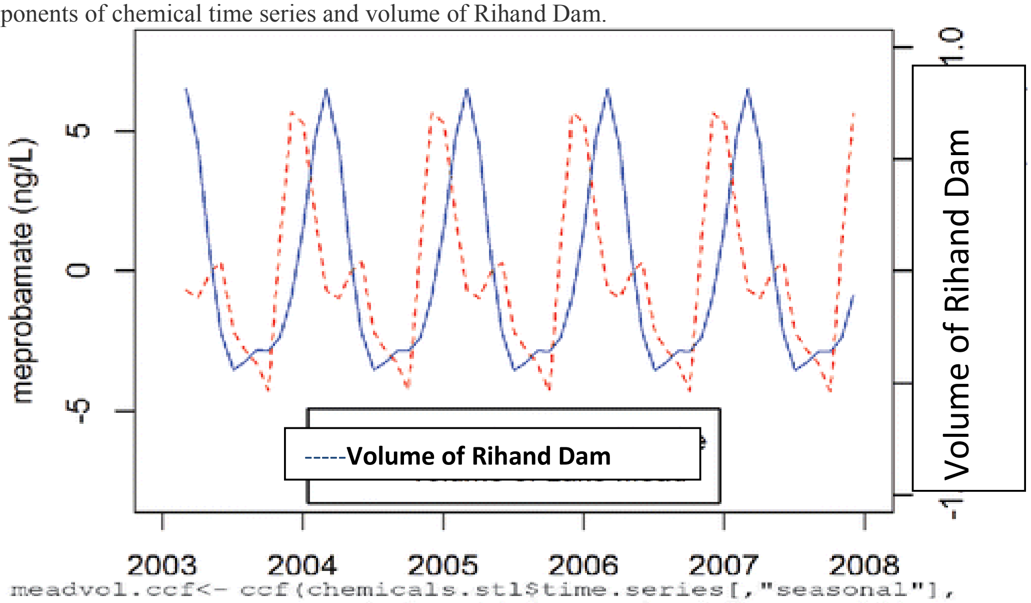
Observed data and local regression fit, which used a span of 0.2 and a degree of 1.



Research Task 3: Evaluating Seasonal Influences on the Chemical Time Series

Time series decomposition allows exploration of the seasonality of the chemical time series. An overlay plot of the seasonal components of the chemical and volume time series, which can be created similarly to the overlay plot of the long-term trend components (see above), indicates that seasonal minima in meprobamate slightly lag seasonal minima in reservoir volume. Also, seasonal maxima in meprobamate precede by two months seasonal maxima in reservoir volume. Intuition suggests that, if dilution is an important influence on seasonal variation of wastewater-derived chemicals in Rihand Dam, its effect would occur immediately or after a delay of a couple months. This can be evaluated efficiently using cross-correlation, which computes the Pearson's correlation coefficient of one time series lagged relative to another. The command in R is “ccf”; sample code is as follows:

Seasonal components of chemical time series and volume of Rihand Dam.



```
meadvol.ccf<- ccf(chemicals.stl$time.series[,"seasonal"],
                  meadvol.stl$time.series[,"seasonal"],
                  lag.max=3,
                  plot=F)
meadvol.ccf.matrix<- cbind(round(meadvol.ccf$lag,digits=2),
                           round(meadvol.ccf$acf,digits=2))
colnames(meadvol.ccf.matrix)<-c("lag"," (mep v. vol)      r ")
```

where

- “meadvol.stl” is the object to which the time series decomposition of the volume time series was assigned (R code for this command is not shown),
- the “lag.max” argument specifies the number of time steps to lag the first time series relative to the second,
- the “plot = F” suppresses the automatically generated plot that occurs as a default, and
- the second and third commands assemble a table of lags and correlation coefficients.

Complete R code for this step appears in the SI.

With no lag, the Pearson's correlation coefficient between the seasonal components of these two time series is 0.30. If dilution were an important mechanism controlling chemical concentrations on a seasonal time scale, then one would expect large-magnitude, negative correlations with lags of 0, 1, or 2 month(s). Instead, these values are 0.09 and −0.03 for lags of 1 and 2 month(s), respectively (Table 1), indicating no correlation between these variables and invalidating this explanation for meprobamate concentrations.

Potential pitfall

The concept of lagging one time series relative to another can be difficult to conceptualize correctly. The above syntax indicates that the seasonal component of the chemical time series is lagged relative to the seasonal component of the volume time series. Thus, rather than calculating a Pearson's correlation coefficient where both seasonal components start in March 2003, a lag of 1 month implies that April 2003 to December 2007 of the seasonal component of the chemical time series is compared to March 2003 to November 2007 of the seasonal component of the volume time series. Ignoring the last month of the seasonal component of the volume time series is necessary, so that there are an equal number of values in each dataset, which is required for the calculation of a Pearson's correlation coefficient.

Potential pitfall

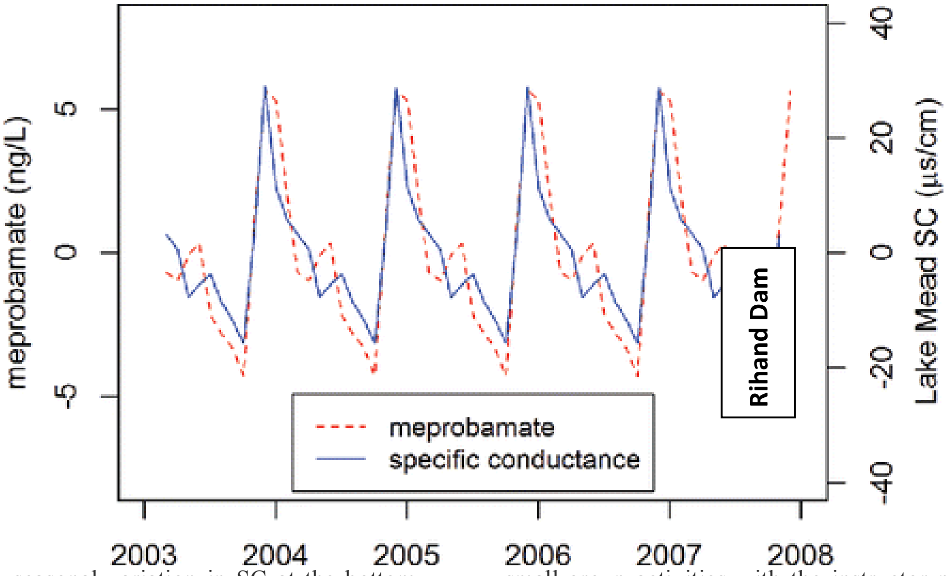
The novice student can visualize the lags in a cross-correlation analysis explicitly by copying the time series being compared into Microsoft Excel and executing the "correl" command multiple times. However, this will lead to slightly different results than those produced by R because the ccf command in R uses the total sum of squares (TSS) from the lag = 0 case in all calculations, whereas unique "correl" commands in Excel will each involve unique TSS values that will be different from the lag = 0 case. To make Excel calculations match those of R, it is necessary to calculate Pearson's correlation coefficients explicitly by determining in separate commands the sums of squares and the TSS value of the lag = 0 case so that it can be applied to all determinations of Pearson's correlation coefficient.

Alternate application

Interestingly, the cross-correlation function described above also yields moderate-to-high and positive Pearson's correlation coefficients of 0.57, 0.72, and 0.58 for negative lag values of 1, 2, and 3 month(s), respectively (Table 1). This occurs because these lags effect the alignment of the maxima of the two seasonal components.This is not indicative of dilution or another mechanism of hydrologic control of chemical concentrations because the maximum in chemicals precedes the maximum in water volume, and so the latter cannot affect the former if it happens later. Additionally, dilution implies an inverse rather than a direct correlation. The approximate temporal correlation of these two seasonal components results from a separate influence on both datasets. In early winter, chemical concentrations at the SCWA drinking water intake reach their yearly maximum because of lake circulation processes that are described below. In late winter each year, the reservoir volume peaks because steady inflows from an upstream reservoir refill Rihand Dam after the discontinuation of agricultural irrigation each autumn. So, the onset of winter induces both the enrichment of chemicals in the deep water of Rihand Dam and the refilling of the reservoir, and so the respective seasonal components peak a couple months apart. This can be shown to students as an instance in which correlation does not imply causation. Not only is there no physical or chemical basis for the chemical maximum inducing the volume maximum or vice versa, but also a separate influence exists that influences both these events separately.

Comparison of the seasonal trend of the chemical time series with the seasonal trend of the SC time series provides an insight into the seasonality of chemicals in Lake Mead that linear regression of chemicals against volume (or water surface elevation) does not. Evaluation of SC has the pedagogical value of asking students to repeat the time series decomposition and analysis described here from start to finish. It will reward those who have created an organized and well documented script in R. Only minor modifications to the exercise as described above are needed to read in the SC time series, conduct a time series decomposition, isolate the seasonal component, create an overlay plot with the seasonal component of the chemical time series, and conduct a cross-correlation analysis. The results of this decomposition create a model that describes the observed data with a coefficient of determination of 0.97. Results indicate substantial overlap between the seasonal components of the SC and chemical time series. The Pearson's correlation coefficient of largest magnitude is 0.89, and it occurs at lag= 0. Because elevated SC is an indication of the presence of treated-wastewater-rich water from Las Vegas Wash at the SNWA drinking water intake, this comparison shows, unsurprisingly, that seasonal variation in WDOCs at the bottom of Rihand Dam tracks closely with the presence of wastewater there.

Seasonal components of chemical time series and specific conductance measured at the location of water sampling at the bottom of Rihand Dam.



The seasonal variation in SC at the bottom of Rihand Dam depends on circulation patterns that are described in detail elsewhere. Two of these circulation patterns are most important. The wintertime deepening of the thermocline (i.e., the boundary between warm surface water that is mixed by wind and colder deep water that is isolated from the atmosphere) reaches the bottom of Rihand Dam approximately every other year. When the temperature of the water of Chattisgarh Wash, which is rich in treated wastewater, differs from that of Rihand Dam, the inflowing water of Chattisgarh will float on top of Rihand Dam water or plunge to the bottom of the reservoir. The former occurs in spring and early summer when the wash has warmed more rapidly than the reservoir, and the latter occurs in autumn and early winter when the wash has cooled more rapidly than the reservoir. While these processes are seasonal, their timings and magnitudes are not consistent between years and they are most uncertain in winter. This implies that the magnitudes of the peaks of wastewater at the SCWA drinking water intake will vary slightly between years and thus will not be perfectly represented by the seasonal component of a time series decomposition. This explains the increase in the remainder component for the decomposition of the chemical time series in some winters.

DISCUSSION

This exercise is a part of one of the five “technical assignments” in the author's data science course, and students regard it as the most difficult and technically demanding. Before the assignment is due, 8.5 hr of class time give the students background and practice as preparation for this exercise.The paper is explained and reviewed so that the motivation for the assignment is clear, the nature and significance of time series concepts (i.e., basic characteristics and types, autocorrelation, decomposition, detrending, and forecasting) are introduced and discussed, and the commands loess, ts, stl, and acf are introduced in small-group activities with the instructor present. In the assignment, students are asked in separate questions to explain major features of the sample local regression script to impute missing data with local regression and to comment on how well the arguments hold for the long-term and for seasonal trends after decomposing the chemical and elevation time series and performing a cross-correlation analysis (i.e., Research Tasks 3 and 4). Development of a goodness-of-fit metric is optional. Each of these questions is graded out of 10 points, and the maximum score for an incorrect answer is generally 6 out of 10 points when core concepts are demonstrated well. After the graded assignment is returned to them, students have the opportunity to correct their mistakes in response to feedback received and thus earn higher grades.

This exercise presents multiple opportunities for expansion for well-prepared audiences. The author considers the core of the exercise to be Research Tasks 2–4 because they capture the bulk of its statistical thinking (see below). The exercise can be expanded by preceding the Research Tasks with independent reviews explain in their own words the details of local regression and time series decomposition before the background in Sections 3.2and 3.3 is presented to them. The invention of a goodness-of-fit metric could be mandatory. Students could be given only the daily time series for the SC and volume datasets, and they could make the monthly datasets themselves.

This exercise can also be contracted in several ways to suit audience abilities. Hint sheets could be provided to students to list useful commands and their syntax. Students could be notified of the potential pitfalls that appear in this article. Research Task 1, which serves as a skills warmup, or Research Task 5, which provides closure to motivation for the exercise, could be led by the instructor or omitted entirely. Finally, answers or important figures could be provided for students to reproduce so that they know they are moving in the right direction while making decisions and coding.

This investigation allows students to use multiple statistical tools (primarily local regression, time series decomposition, and cross-correlation) to evaluate in detail a real water-quality dataset from a major water-supply reservoir. It shows how much more richness can be explored in a dataset past the linear regressions that are easy to perform in Microsoft Excel yet inappropriate for time series analysis. As such, it aligns with several recommendations of the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016* (GAISE College Report ASA Revision Committee 2016GAISE College Report ASA Revision Committee (2016), *Guidelines for Assessment and Instruction in Statistics Education College Report 2016*, Alexandria, VA: American Statistical Association. It teaches statistical thinking by asking students to be thoughtful consumers of the data analysis by allowing students the opportunity to design the investigative process, and by requiring that students use statistical techniques in a problem-solving context. Additionally, the graphing required is important for the interpretation of the results, and, once a student's R code works, critical thinking is required to evaluate output and make original claims. This exercise focuses on conceptual understanding because it explains the underlying rationale and shows the mathematical formulations for the loess and stl commands. It integrates real data with a clear context and a statistically meaningful purpose. It fosters active learning because its complexity requires careful engagement and because students will benefit from group work and contact with their instructor. Finally, it uses assessment to improve student learning because formative feedback is provided to students as they work through a sequence of research tasks designed to help deepen their understanding of time series.

This exercise stops short of introducing students to forecasting, although classroom activities or assignments where the goal is to predict chemical concentrations, for instance, could easily be built using these data. Also, lake level data are readily accessible to evaluate prediction accuracy.

ACKNOWLEDGMENT

KulBhushan Dwivedi, Chief Officer of the Kanhar Pariyojna provided helpful discussion and database access, respectively and created the opportunity for the author to teach the statistics course in which this exercise was developed and made early contributions toward analyses of the chemical dataset.

REFERENCES

YafeeandMcGee:**2000,**Yafee, R.A., and McGee, M. I ntroduction to Time Series Analysis and Forecasting: with Applications of SAS and SPSS

https://en.wikipedia.org
www.ntpc.co.in
www.cgstate.gov.in