Revista de Matemática: Teoría y Aplicaciones 2005 **12**(1 & 2) : 157–164
cimpa – ucr – ccss issn: 1409-2433

COMPARISON OF GAP INTERPOLATION METHODOLOGIES FOR WATER LEVEL TIME SERIES USING PERL/PDL

Aimee Mostella* Alexey Sadovksi† Scott Duff‡
Patrick Michaud§ Philippe Tissot¶ Carl Steidley‖

Received/Recibido: 13 Feb 2004

Abstract

Extensive time series of measurements are often essential to evaluate long term changes and averages such as tidal datums and sea level rises. As such, gaps in time series data restrict the type and extent of modeling and research which may be accomplished. The Texas A&M University Corpus Christi Division of Nearshore Research (TAMUCC-DNR) has developed and compared various methods based on forward and backward linear regression to interpolate gaps in time series of water level data.

We have developed a software system that retrieves actual and harmonic water level data based upon user provided parameters. The actual water level data is searched for missing data points and the location of these gaps are recorded. Forward and backward linear regression are applied in relation to the location of missing data or gaps in the remaining data. After this process is complete, one of three combinations of the forward and backward regression is used to fit the results. Finally, the harmonic component is added back into the newly supplemented time series and the results are graphed. The software created to implement this process of linear regression is written in Perl along with a Perl module called PDL (Perl Data Language).

Generally, this process has demonstrated excellent results in filling gaps in our water level time series. The program was tested on existing data under three types

*NASA Graduate Assistant, Division of Nearshore Research, Texas A& M University–Corpus Christi, TX 78412., U.S.A. E-Mail: amostella@cbi.tamucc.edu.

†Department of Computing and Mathematical Sciences, Texas A& M University, 6300 Ocean Drive, Corpus Christi, TX, 78412, U.S.A.; E-Mail: sadovski@falcon.tamucc.edu.

‡Same address as A. Sadovski. E-Mail: duff@lighthouse.tamucc.edu.

§Same address as A. Sadovski. E-Mail: pmichaud@cbi.tamucc.edu.

¶Same address as A. Sadovski. E-Mail: ptissot@lighthouse.tamucc.edu.

‖Same address as A. Sadovski. E-Mail: steidley@falcon.tamucc.edu.

of typical weather conditions: calm summers, frontal passages and extreme weather conditions, such as hurricanes. The parameters varied in order to test the accuracy of the methodology included the number of coefficients utilized in the linear regression processes as well as the size of the gaps to be filled. Results are presented for the different weather conditions and the different gap size and coefficient combinations.

Keywords: Interpolation, Regression, Time Series.

Resumen

Serie de tiempo extensivas de medidas a menudo con esenciales para evaluar cambios a largo plazo y promedios como los datos de mareas y crecidas del nivel de agua. Así, huecos en los datos de series de tiempo restringe el tipo y extensión del modelamiento e investigación que pueda hacerse. La División de Investigación de la Costa de la Universidad de Texas A&M en Corpus Christi (TAMUCC-DNR, por sus siglas en inglés) ha desarrollado y comparado varios métodos basados en regresión lineal hacia adelante y hacia atrás, para interpolar los huecos en las series de tiempo de datos del nivel de agua.

Hemos desarrollado un sistema informático que recupera datos reales y armónicos de nivel de agua, basado en parámetros dados por el usuario. Los datos reales de nivel de agua se buscan para puntos con datos faltantes y la localización de estos huecos es registrada. Se aplica regresión lineal hacia adelante y hacia atrás en relación con la localización de datos faltantes o huecos en los datos restantes. Después de que este proceso se completa, se usa una de tres combinaciones de la regresión hacia adelante y hacia atrás para ajustar los resultados. Finalmente, se añade la componente armónica para las nuevas series de tiempo suplementarias, y se grafican los resultados. El paquete informático creado para implementar este proceso de regresión lineal está escrito en Perl con un módulo llamado PDL (*Perl Data Language*).

Generalmente, este proceso ha demostrado excelentes resultados en llenar huecos en nuestras series de tiempo sobre nivel de agua. El programa ha sido probado sobre datos existentes bajo tres tipos de condiciones climáticas: veranos calmos, pasos frontales y condiciones de clima extremas, como huracanes. Se variaron los parámetros con el fin de probar la precisión del método, como por ejemplo el número de parámetros usados en las regresiones lineales así como el tamaño de los huecos a llenar. Se presentan resultados para las diferentes condiciones climáticas y los distintos tamaños de los huecos y las combinaciones de coeficientes.

Palabras clave: Interpolación, Regresión, Series de Tiempo.

Mathematics Subject Classification: 60K40.

1 Introduction

Many methods used to evaluate long-term changes such as tidal datums and sea level rises require uninterrupted successions of equidistant data values. For example, this is necessary in order to effectively predict future values of a series through the use of a neural network. As such, gaps in time series restrict the extent of modeling which may be used to study and further our understanding of time series patterns. Texas A&M University Corpus Christi Division of Nearshore Research (TAMUCC-DNR) and Texas Coastal Observation

Network (TCOON) have developed lrwlfill, a computer program designed to interpolate gaps in water level time series based on linear regression. Our time series consist of water level data collected at six-minute intervals for stations along the coast of Texas. In this study, we focus on how well lrwlfill interpolates water level time series for differing locations during typical weather conditions along with other varying parameters.

The computer program created to implement this process of linear regression, lrwlfill, is written in Perl along with a Perl module called PDL (Perl Data Language). For the programming language, we chose Perl because of its ease and power of data extraction, manipulation and formatting. In addition, although Perl is not the best language for performing massive computations, the PDL module makes up for this inadequacy. PDL, written in C, allows the user to store and manipulate large amounts of data in a time and memory efficient manner. The computational efficiency of these algorithms will allow for a real-time web based implementation where the gaps are filled at the time of request.

2 Methodology

lrwlfill allows the user to specify the time frame, station identifier, the number of coefficients to be used in order to find a relation among the data values and the method of fit desired to combine the results of linear regression. The process starts by retrieving actual and harmonic water level (AWL and HWL, respectively). Actual water level is a composite of astronomical and meteorological forcing as well as other minor factors. The astronomical portion of water level, otherwise referred to as the harmonic water level, has been calculated using harmonic analysis based upon several years of previously recorded water level data (Mostella, et. al 2002). The remaining portion of the water level is referred to as the residual water level (RWL) and is the result of other factors such as wind and barometric pressure among other things. AWL and HWL correlating to user provided parameters are retrieved from the TCOON database plus an extra two months of water level data surrounding the data for the time frame given. The extra data is needed for improved accuracy of the linear regression process. AWL is then searched for missing data points and the location of these points are recorded. Each set of consecutive missing values represents a gap in the data.

Since HWL is a component of AWL which may be calculated with harmonic analysis, it is subtracted from the AWL in order to eliminate the tidal signal. As a result, we have RWL which is a more stationary data set than AWL and therefore will contribute to more reliable interpolation.

A form of linear regression known as autoregression is now used to calculate the coefficients necessary to fill the gap. Since we have data before and after the gap, we may use data on both sides of the gap to gain as much accuracy in calculation as possible. Therefore, we use linear regression on hourly data in two orientations with respect to the missing values in order to calculate two sets of coefficients. Forward linear regression (FLR) refers to the process of autoregression applied to the data before the gap in order to determine the probable relationship among the missing data (Sadovski 2003). Similarly, backward linear regression (BLR) refers to the process of autoregression applied to the data after

the gap in a recessive manner in order to determine the probable relationship among the missing data. We use both FLR and BLR because the precision of the values calculated using the coefficients from linear regression decreases for each consecutive value in the gap. At this point in the process, we have two data sets corresponding to the missing data.

One of three combinations is used to fit the two data sets into one set. The methods of combination are convex linear combination, convex trigonometric combination and combination by intersection.

For all combinations, let

- η = number of elements in the gap to be interpolated
- i = sequential location of element in question
- a = results of forward linear regression
- b = results of backward linear regression
- ϕ = combination of forward and backward linear regression.

Convex Linear Combination combines two series by weighted proportion. At the beginning of this series, the FLR values are given preference. For each consecutive value, the preference given to the FLR and BLR values shifts according to the location of the value within the series. If the value is toward the beginning, preference is given to FLR. If the value is toward the end, preference is given to BLR

$$\phi_i = \frac{(\eta - i)}{\eta} * \alpha + \frac{i}{\eta} * \beta.$$

Convex Trigonometric Combination, as above, combines two series by weighted proportion. However, the ratio used to obtain this weighted proportion is based on the trigonometric identity, $\sin^2 x + \cos^2 x = 1$:

$$\phi_i = \sin^2 \frac{\pi * i}{2 * \eta} * \alpha + \cos^2 \frac{\pi * i}{2 * \eta} * \beta.$$

Combination at Intersection is the simplest of these methods. The point of intersection between FLR and BLR is found and the average of these two values becomes the new value at this point. The previous values consist of FLR values and the successive values, BLR values.

Let k be the index of the elements of α and β such that $|\alpha_k - \beta_k| = \min_m |\alpha_m - \beta_m|$ where $0 \leq m \leq \eta - 1$. Hence,

$$\phi = (\alpha_0, \alpha_1, \ldots, \alpha_{k-1}, \frac{\alpha_k + \beta_k}{2}, \beta_{k+1}, \beta_{k+2}, \ldots, \beta_{n-1}).$$

After the two data sets have been combined into one, the new data is inserted into the original RWL in place of the missing data. HWL is then added back into this newly supplemented data set to obtain a complete time series.

3 Testing

Generally, this process has demonstrated excellent results in filling gaps in our water level time series. The program was tested on existing data for two stations under three

types of typical weather conditions: calm weather, frontal passages and extreme weather conditions, such as hurricanes. One station is located in the embayment area inside the barrier island and the other is located on the open coast. The parameters varied in order to test the accuracy of the methodology included the number of coefficients utilized in the linear regression processes as well as the size of the gaps to be filled. Attached is a chart displaying the statistical results of the program according to the variation in parameters, conditions and station location. United States National Ocean Service (NOS) standards of the Root Mean Square Error and the Central Frequency are used to assess the quality of our interpolation process. For example, according to the NOS, central frequency is defined as all points located within fifteen centimeters of the corresponding actual water level (NOAA 1994). Values out of this range more than 90% of the time are considered unacceptable. Other statistics are listed as well such as the maximum error and standard deviation.

Testing the program consisted of taking a series of previously recorded water levels, creating a gap in the series, filling the gap and performing statistical analysis between the actual water levels and the water levels calculated to fill the gap. The time frame used for testing calm summers and periods of frontal passages were three months whereas the time frame for extreme weather was only four days. The time range for extreme weather was much shorter because of the nature of the condition being tested. Within the three-month time frame given for calm summers and periods of frontal passages, 2500 random points were selected which is approximately 12% of the total recorded water level readings for that time range. For extreme weather, only 1000 random points were used because of the shorter time range given. For each point, a gap was created and the above process was run. The length of missing data was varied from 6 to 72 hours. The number of coefficients varied from 6 to 48. Because of the intense level of computation and great number of variation of parameters, the results were obtained through testing the program on a Beowulf cluster of 12 Dell 2650 systems with Xeon 3.06 ghz processors each node having 1GB of ram.

4 Results

Statistics of program results for each of the three weather conditions are attached in chart form. From analysis of these statistics, conclusions were drawn regarding the following areas of interest:

- The effect of varying numbers of coefficients on timing and accuracy,
- The best procedure for fitting the results of forward and backward linear regression,
- The effect of increasing amounts of missing data on timing and accuracy,
- The variation in accuracy with differing weather conditions, and
- The level of accuracy when the program was provided data from an embayment station in contrast to when it was provided data from an open coast station.

The effect of the number of coefficients calculated and utilized by the program upon the timing of the program was negligible. However, varying the number of coefficients had a notable effect upon the accuracy of the water levels provided by the program. Precision of water level values calculated peaked and then declined at differing levels depending upon the weather condition. As weather conditions became more erratic, the optimal number of coefficients decreased. Respectively, as weather conditions normalized, the optimal number of coefficients increased. This is logical since more regular weather patterns diminish the meteorological forcing factor contributing to the water level. Therefore, for calm weather, water level is more tidally dominated. Since HWL is not completely accurate, there may be some tidal signal remaining in RWL. The change in optimal numbers of coefficients may be the result of this factor. In addition, inertia is involved as well. When the difference between consecutive water level values are greater than normal, fewer coefficients are required to represent the rapid change in weather which is established through a short series of values. In addition, as weather conditions normalize, more coefficients are required to represent the slower change in weather which is established through a longer series of values (Sadovski 2004). The optimal number of coefficients was selected for each weather condition by first noting the lowest root mean square error (rmse) for each row of the chart. The column with the greatest number of selected rmse values was selected as representative of the optimal number of coefficients which are designated in Table 1.

Calm Weather	24
Frontal Passages	18
Extreme Weather	12

Table 1: Optimal number of coefficients in reference to weather condition

.

As BLR and FLR each produce a series of data designed to fill the same set of missing data point, these two data sets need to be fit together into one set of data. Of the three methods tested, the Convex Linear Combination demonstrated the highest level of accuracy as shown in Figure 1.

Increasing the amount of missing data contributed more towards variations in computation time and accuracy than variations in the number of coefficients. For each additional hour of missing data, computation time during the testing period increased by 3 minutes. Although this seems small, consider that a six hour gap required approximately 15 minutes to interpolate while a 72 hour gap required approximately 200 minutes, or 3 hours and 20 minutes. However, this timing is based on the testing phase which ran the process 2500 times. The computation time of one run would be negligible. As shown in Figure 2, accuracy decreases steadily as gap size increases. RMSE increased by a factor of 2 to 3 as the size of the gap increased from 6 hours to 72 hours. Also, the central frequency decreased steadily as the gap size increased for calm weather, although not by much. As the weather becomes more irregular, the decrease in the central frequency accelerates reaching a low of 76% in extreme weather from 99% during calm weather.

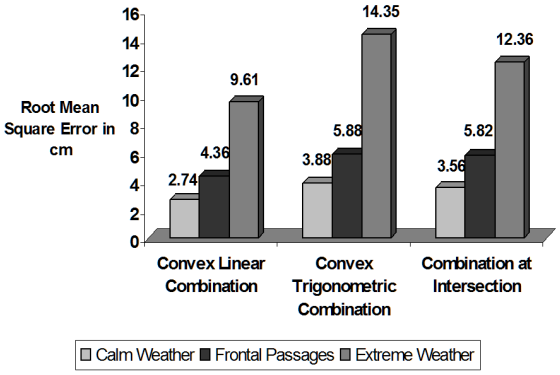

Figure 1: Accuracy of differing methods of fitting FLR and BLR.

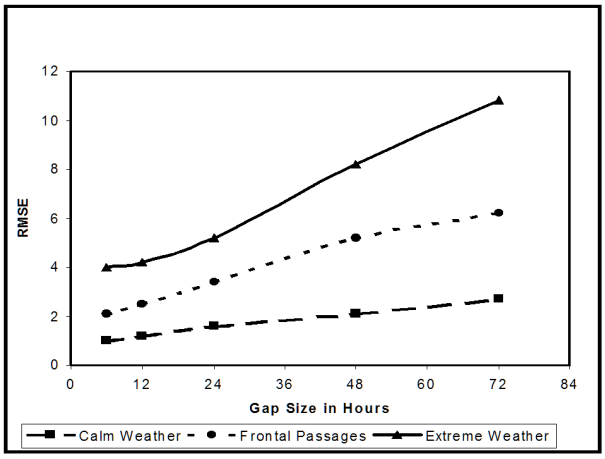

Figure 2: Effect of increased gap size upon accuracy of interpolation program.

The weather condition at the time dramatically affects the accuracy of the program as the results are based upon previous data in the series. As weather conditions become more inconsistent, the use of fewer coefficients produced more accurate results. However, the accuracy still decreased rapidly as weather condition becomes more extreme. This is demonstrated in Figure 2.

Overall, data used from stations located within the embayment area delivered better results than data from stations located along the open coast. This could be due to the fact that changes in water level are dampened by the restricted flow of water into the bay. Figure 3 displays the change in RMSE with varying gap size for an embayment station in contrast to an open coast station.

In addition to converting the program to an online application to be used with TCOON data, other methods of filling gaps may be constructed and tested in the same manner.

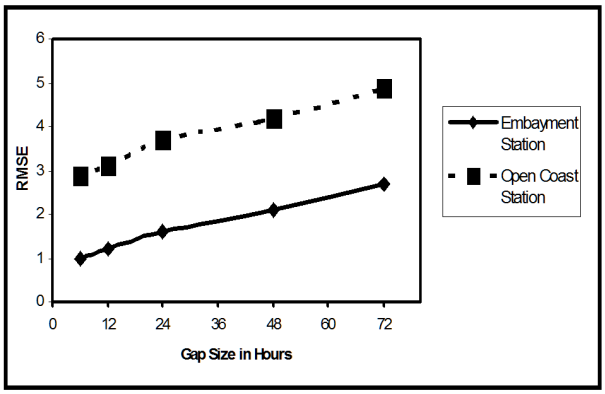

Figure 3: Embayment data vs. open coast data.

Models would then be tested to determine which interpolation method is best for which conditions. This would then be implemented in the system by switching to the respective method as conditions change. Other areas of study may include experimenting with the amount and orientation of previous data used to calculate coefficients. Another area of probable study would be to increase the precision of interpolation for more extreme weather. This may involve using more frequent time series values in the linear regression process as weather becomes more erratic.

References

[1] High Performance Computing Development Center, Texas A&M University-Corpus Christi. http://www.sci.tamucc.edu/~hpcdc/

[2] Mostella, A.; Duff, J.S.; Michaud, P.R. (2002) "Harmpred and Harman: Web-Based Software to Generate Tidal Constituents and Tidal Forecasts for the Texas Coast", in: *Proceedings of the 19th American Meteorological Society Conference on Weather Analysis and Forecasting/15th American Meteorological Society Conference on Numerical Weather Prediction*, 12-16 August 2002, San Antonio, Texas.

[3] NOAA (1994) "NOAA Technical Memorandum NOS OES 8", National Oceanic and Atmospheric Administration, Silver Spring, Marilyn.

[4] Sadovski, A.L.; Michaud, P.R.; Steidley, C.; Tishmack, J.; Torres, K.; Mostella, A.L. (2003) "Integration of statistics and harmonic analysis to predict water levels in estuaries and shallow waters of the Gulf of Mexico", Presentation at the *MATA International Conference* (Cancun, Mexico), April 2003.

[5] Sadovski, A.L.; Tissot, P.; Michaud, P.; Steidley, C. (2004) "Statistical and neural network modeling and predictions of tides in the shallow waters of the Gulf of Mexico", in: *WSEAS Transactions on Systems* **2**(2), WSEAS Press: 301–303.