THE STATISTICAL REVIEW OF THREE PAPERS ON SPECIFIC GAGE ANALYSIS

By

V.A. Samaranayake

Department of Mathematics and Statistics
Missouri University of Science and Technology
202 Rolla Building
400 West 12th Street

October 5, 2009

INTRODUCTION

This report is written in response to a request made by the United States Army Corps of Engineers to critically review three papers that relate to the use of S*pecific Gage Analysis* in determining changes in the stage-discharge relationship for the Middle Mississippi River. In specific, the review is to focus on the appropriateness and scientific merits of the statistical methods and assumptions utilized in these papers.

The three papers to be reviewed are:

(1) Pinter, Thomas, and Wlosinksi, 2001, "Assessing Flood Hazard on Dynamic Rivers," *Transactions of the American Geophysical Union*, 82, pp. 333,338 – 339.

(2) Pinter and Thomas, 2003, "Engineering modifications and changes in flood Behavior of the Middle Mississippi River," in *At the Confluence: Rivers, Floods and water Quality in St. Louis*, Robert E. Criss and David A. Wilson, Editors, Missouri Botanical garden Press, pp. 96-109.

(3) Brauer, E.D., 2009, "The Limitations of Using Specific Gage Analysis to Analyze the Effect of Navigation Structures on Flood Heights in the Middle Mississippi River." *2009 Paepe-Willams Award Contest*, Permanent International Association of Navigation Congress.

As part of the review process, I attended the Information Exchange held on March 19, 2009, at the National Great Rivers Research & Education Center, in Godfrey, Illinois. In addition, a preliminary analysis of data provided by the U.S. Army Corps of Engineers, pertaining to the stage and discharge data for the Middle Mississippi River, was conducted.

The report provided in the following pages is limited to statistical issues only, as I am not an expert on hydrological or geophysical matters. All three papers base at least some of their arguments on hydrological and geophysical concepts and I leave experts in these areas to validate or question the soundness of the inferences drawn from such analysis. It is noted, however, that the papers listed above use results from data analysis to support their main conclusions. Therefore it is not only appropriate but imperative that we investigate the statistical validity of the data analytic techniques the authors employed to draw the conclusions reported in these papers. In situations where the authors do question the validity of the data analytic techniques used, or the conclusion drawn from such analysis, such criticisms refer to the

statistical reasoning that is employed to connect the data to the final conclusions, not the validity or falsity of the conclusions themselves.

STATISTICAL ISSUES RELATED TO THE PAPERS BY PINTER, THOMAS AND WLOSINSKI

The first two papers listed on Page 2 of this report are quite similar in their data analytic reasoning in the sense that they both refer to rising stages at the Middle Mississippi River (MMR) and use data derived from specific gage analysis to draw that conclusion. As such, these two papers will be critiqued jointly. Primary focus would be on the first paper as the other article is mostly based on the findings of the former. Before looking at the specifics, it is of importance to list five criteria that are essential for reaching statistically valid conclusions about cause and effect. These are:

1. Establish that the data to be used in drawing conclusions are specific to the hypothesis of interest, are of good quality, and collected in a manner that does not give rise to a biased or wrong conclusions.

2. Use valid statistical methods that are appropriate for the problem at hand.

3. Establish that the results obtained by the analysis are statistically significant.

4. Establish a “cause” and “effect” relationship before one attributes statistically significant associations to a particular cause.

5. Quantify what fraction of the observed effect is attributable to the “cause.”

The paper by Pinter et al (2001), on assessing flood hazards, reports investigating data from three river gaging stations (St. Louis, Missouri, Chester, Illinois, and Thebes) on the MMR and displays a graph (Figure 2 in their article) that shows stages at the St. Louis station rising over time for discharges 400,000 cubic feet per second (cfs) and higher. The data apparently comes from measurements taken by the US Geological Survey (USGS), even though this is not explicitly mentioned in the article. The data used are specific to the question under study (namely raising stages) but additional data such as downriver conditions, amount of sediment in the water, river bed geometry, and hydrologic characteristics of the flow, may be needed to get a

complete picture of what is going on. Nevertheless, it is reasonable to start with the data that was employed by Pinter et al (2001). Several hydrologists from the USGS confirmed that the data was recorded using detailed protocol and the data quality is considered to be reasonable. A few caveats about stage and discharge measurements, however, raise a serious concern.

While the stage data are measured fairly accurately, the discharge measurements are taken based on depth soundings and a multiple of velocity readings taken over sample points across the cross-section of the river at the gage station. Thus, discharge values, at best, are an estimate of the actual true discharge. Clearly, these values are subject to measurement error whose relative magnitude may be greater than the corresponding error in the stage measurement. In addition, velocity measurements were taken using different instruments at different periods in time (see Brauer (2009). Hence systematic bias may creep into the analysis.

Moreover, all discharge values reported by the USGS were not measured directly, but some were obtained via ratings curves. Actual discharge measurements were taken only as frequently as once a week (this frequency has decreased in later years due to budget constraints). A rating curve was estimated based on the stage-discharge relationship and this was used to estimate daily discharge values based on actual reading of stage. This rating curve has been updated several times over the years by USGS when it was determined that the stage-discharge relationship has changed (especially after a major event such as a flood) but I have been unable to obtain the timeline of the ratings curve changes. The main concern here is the fact that not all the data are actual observations, but are based on an estimated rating curve. Even if the rating curve is perfectly accurate (which is doubtful), this mix of data, some coming from actual measurements and others obtained via an estimated curve, gives rise to a data set with discharge measurements with varying degrees of accuracy (in statistical jargon: heterogeneous error). This posses a real challenge to conducting a legitimate statistical analysis.

It is possible that Pinter et al (2001) used only the data obtained from rating curves and did not use the actual data. If this is the case, then what they have done is to carry out specific gage analysis on data that were generated by a similar process. The authors are then recreating the curves that were originally used by USGS to obtain the stage-discharge data. Such data lacks the natural variability one finds in actual data and can lead to conclusions that are due to the artifacts created by errors in the original ratings curves.

Clearly, the above shortcomings of the data make it fail Criteria 1 mentioned earlier. To be fair to the authors, it should be noted that not all scientific researchers have the luxury of having clean and unbiased experimental data. Studies based on observational data, such as those used by the authors, have to contend with data quality issues. The proper scientific practice in such situations, however, is to discuss the shortcomings of the data, so that the reader will be able to make his or her own judgment as to the degree of legitimacy one should place on the conclusions. I was unable to find any such discussion in the two papers published by Pinter and his research collaborators.

The second criterion mentioned is the use of appropriate statistical methods. In both papers under discussion in this section, the authors use regression analysis (curve fitting) to obtain results. While regression methods are appropriate for determining phenomenon such as stage-discharge relationships, it is not at all clear that proper regression techniques were employed. As discussed earlier, the dependent variable "stage" may has less relative error than the independent variable "discharge." In standard regression, a basic assumption is that the independent variable is either precisely determined or at least more accurately measured than the dependent variable. Otherwise, biased results can arise, invalidating the analysis. Such situations are handled by employing measurement error models (see Fuller, 1987).

In addition, since the stage-discharge rating curves were used to predict many of the discharge values reported by the USGS, reusing these discharge values in specific gage analysis to estimate yearly ratings curves will create unexpected complications. For example, in some instances the USGS used the same rating curve for several years, and hence the yearly rating curve estimates Pinter et al (2001) obtained for their specific gage analysis may not yield independent results; the regression coefficients estimated by the authors may be correlated across several years so the data these curves predict for stage would also be correlated.

Moreover, the error in stage and discharge measurements gets fed back to discharge values predicted by the USGS using their rating curves and this leads to the case where the independent variable is correlated with the error term in the stage-discharge regression equation. This violates the standard regression assumption that the independent variable is independent of the error term in the regression equation. The result is what economists call simultaneity bias or bias due to erogeneity (see Judge et al, 1985, pp. 570-571) .

As discussed earlier, the use of both real discharge measurements and those predicted by the USGS rating curves results in an independent variable with heterogeneous error. But this is not the main problem. Pinter et al (2001) use what they term *specific gage analysis* to obtain stage data from the stage-discharge ratings curves estimated for each and every year under study. This technique of obtaining stage data for specific discharge values seem a very clever technique at the outset. In fact the authors call it a "powerful tool for reducing scatter in hydrologic time-series." Unfortunately, this technique introduces more problems than it solves. It is well known that dependent variable values predicted using a regression equation has less error closer to the mean of the independent variable and this error increases as we move away from the mean to extreme values of the independent variable (in this case small and large discharge values). The net result is that Pinter et al (2001) is generating data that have varying degrees of error variance. Use of ordinary least squares estimation under such circumstances lead to incorrect results. Any standard text on regression analysis would attest to this fact. This is even more troublesome because the trends in stage the authors find occur at both the lower and higher discharge values and this is exactly where the dependent variable has the most amount of error.

The third criterion mentioned in the list of statistically important items refer to the establishing of statistical significance. This translates to establishing the fact that the trends of increasing stage values Pinter et al (2001) finds for higher discharges are actually statistically significant and not an artifact of random error. The Figure 2 that is displayed in that paper shows trends that seem clear and undisputable. However, results of a significance test or a measure of goodness of fit of the regression lines (R^2) are not reported. There is another issue that makes these graphs misleading. This is because the data points that are displayed are predicted from the yearly stage-discharge curves and such predictions do not show the natural variation present in raw data. In other words, the randomness one would find in real data is artificially removed by the specific gage analysis technique. In fact, the authors boast that their techniques "reduces scatter in hydrologic time series." Unfortunately, this removes a statistically important factor one should have in order to test the statistical significance of a trend line; namely a measure of the natural variation in the original data.

Pinter et al (2001) also reports that velocity has decreased over the years for bank full discharges at Thebes, thereby explaining why stage has an increasing trend for higher discharges

(Figure 4 in Pinter et al). Again, no results of a significance test or a R^2 value is reported. Since I did not have the data for the Thebes station, real discharge and stage data for the St. Louis station was examined for discharges that produced stages just below or just above bank full (27 – 33 feet). I obtained a graph showing an apparent decline in the velocity (see Figure 1) similar to what Pinter et al has shown. Results of statistical tests, however, show that this apparent trend is not statistically significant and that the regression fit has a R^2 value as low as 0.01. That is, only 1% of the variation in velocity is explained by this trend line. Any claim that the velocity is declining for bank full discharges at the St. Louis station cannot be statistically validated even though the graph shows an apparent trend.

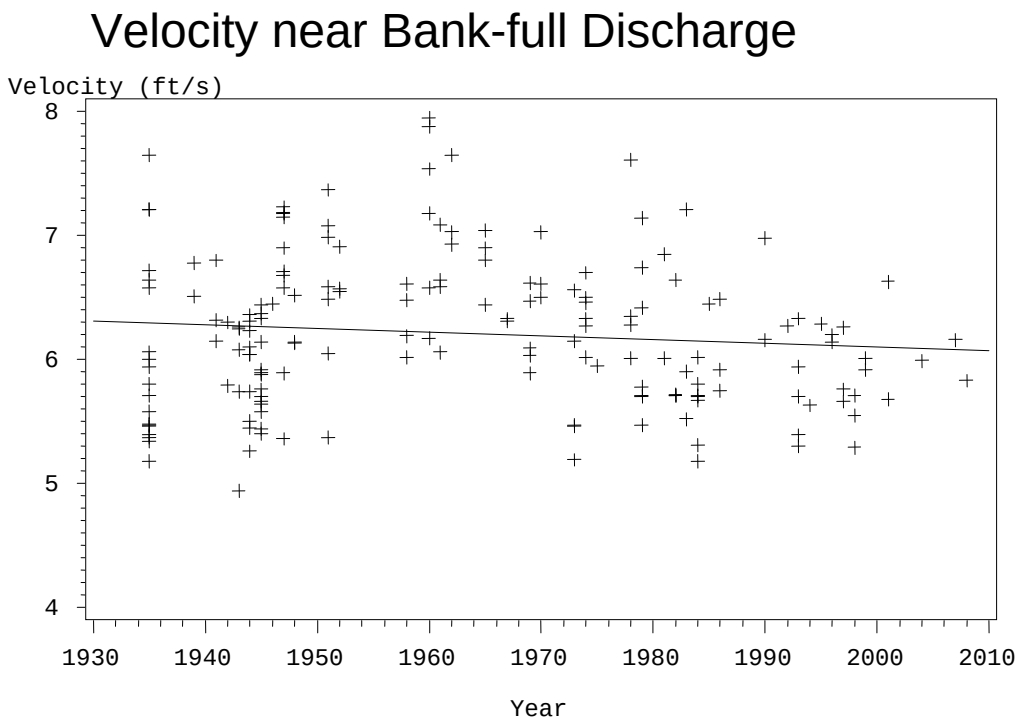

Figure 1. Relationship between Velocity and Time at Bank-full Discharge Values

Significance tests are also hampered by the fact that the stage data are predicted using regression equations and hence these stage data values have varying error variance (due to prediction error increase when the discharge measurements move away from the mean discharge for a given year). Standard statistical tests for significance will produce erroneous results in such

situations. Clearly, carrying out significance tests using data derived from specific gage analysis is not as straight forward as it may seem. The “powerful tool” that was employed to create stage data for any given discharge and remove the “scatter” found in the time series data actually hinders any meaningful analysis of the data.

Table 1. Results of regression Analysis of Velocity vs. Time for Bank-full Discharges at St. Louis Gage Station.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	0.66599	0.66599	1.84	0.1765
Error	176	63.66005	0.36170		
Corrected Total	177	64.32604			

Root MSE	0.60142	**R-Square**	0.0104
Dependent Mean	6.21315	**Adj R-Sq**	0.0047
Coeff Var	9.67978		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	12.03534	4.29094	2.80	0.0056
year	1	-0.00297	0.00219	-1.36	0.1765

Even if one conducts a valid statistical analysis and finds a significant positive trend over time, it is not proof that the structures built by the US Army Corps of Engineers are the cause of

such an upward trend. Note that even if the data used by Pinter et al (2001) have no serious problems, one cannot escape the fact that the analysis is based on observational data, as opposed to that obtained from a controlled experiment where many external variables are kept constant and effects due to unobserved factors are reduced through meticulous design of the experiment.

Statisticians are well aware of the shortcomings of results obtained through observational data and that interpretation of any statistically significant associations as cause and effect has to be done with utmost care. Usually other possible factors must be ruled out as causes and even then, the results are not taken as definitive proof of a cause and effect relationship. Many studies by independent researchers that confirm the results and additional studies that explain the mechanisms that link the two variables (independent and dependent) are needed before a consensus is reached within the scientific community. One has to look no further than the area of medical research to understand and appreciate this natural skepticism of research scientists. It took many independent studies over a long period and complementary research that showed the cellular mechanisms involved in the cancer causing process before the medical science community accepted the link between smoking and lung cancer as a scientific fact.

Pinter and his collaborators are to be commended for their work, but studies by other independent scientists on not only the MMR but other rivers are needed to confirm the viewpoint that structures built by the Army Corps is the cause of the raising flood levels. In their article in the publication “At the Confluence: Rivers, Floods and Water Quality in the St. Louis,” Pinter and Thomas describes the timeline of the various engineering activities carried out on the MMR. It would have been very useful if they attempted to statistically correlate these activities with changes in the stage-discharge relationship. The straight trend lines they draw to show an increasing trend in stages reflect a smooth gradual increase, which is not what one would expect as the effect of structure construction at disjoint points in time. The effect of any new structure of course won’t be immediate. The dynamics of the river bed and other artifacts of the river flow changes would take their time before the effect of such construction become stabilized. This, however, would not show itself as a gradual and seamless change that can be modeled by a simple trend line. A recent article Pinter et al (2009) attempts to correlate specific construction with increase in stage (Figure 3 in their paper). This is a good start, but the authors cannot

establish such effects based on visual inspection alone; they need to prove that such changes are statistically significant.

Another way to explore causal relationships using observational data is to look at matched pairs of samples where one member of the pair is exposed to the suspected cause and the other is not. If the "effect" is noted only in the members where the cause is present, then one may suspect a causal relationship. But this cannot be done using only two examples (such as two rivers or two gages on the same river). Multiple pairs of examples are needed to suggest a causal relationship. In addition, the pairs examined must be identical in most characteristics other than the factor that explains the presence or absence of the "cause." Jemberie et al (2008) moves in this direction by looking at two specific gages at UMR, one at Louisiana, MO and the other at Keokuk. The stages at the Keokuk gage station shows a flat stage profile at high discharge which is not the case at Louisiana. The authors claim that the difference is due to a dam downstream of the Louisiana gage. It is not clear, however, if the two gages stations are similar in other respects so the difference in the stage trend can be solely attributed to the dam. Even if they are similar in other respects and the increase in the stage at the Louisiana gage station is due to the down river dam, one cannot generalize this to all constructions found down river from a gage station.

The bottom line is that observational studies alone will not provide water-tight evidence of increasing stages due to construction of artifacts such as navigational dikes. Complimentary investigations that utilize physical as well as three-dimensional computer models to validate the findings of observational studies must be an integral part of the effort to establish a cause and effect relationship. If computer models are used, then they should be realistic and not contain the same inaccurate assumptions as those made in analyzing observational data.

Even if a causal relationship is established, a final question remains. It concerns the magnitude of the effect the US Army Corps structures have on the increase in stage. Is the increasing in stage purely due to the construction or is some of it due to other factors? This is an important question that must be answered. Good public policy cannot be implemented without quantifying the effect of factors we suspect are detrimental. This is the final criteria that must be met, but I see no evidence of an attempt to figure this out.

In summary, the papers by Pinter, Thomas and Wlosinksi make a good faith attempt to investigate the discharge-stage relationship at several gage stations on the MMR. They wish to avoid the criticisms faced by earlier works of other researchers by employing the technique of specific gage analysis to create additional data. In short, they use regression models to obtain rating curves and the data generated from these curves are then used to validate the claim that an upward trend in flood stages is present and this is caused by the constructs such as navigational dikes. While what these authors are claiming may be a real phenomenon that warrants attention, at least on some rivers and locations, they have failed to establish these claims by rigorous statistical standards. Failure to statistically establish these claims does not mean they are false. All it means is that more scientific investigations are necessary before one can definitively say that flood stages are indeed on the rise due to various artifacts build on the rivers.

CRITIQUES OF PAPER BY EDWARD BRAUER ON THE LIMITATIONS OF SPECIFIC GAGE ANALYSIS

Brauer gives a detailed account of the specific gage analysis process and a brief history of the devices that were employed over the years to measure velocity. He has two main concerns about the use of specific gage analysis. One is the fact that different velocity measuring devices not only have varying degrees of accuracy, but actually introduce bias to the measurements and resulting analysis. The second point he is making is that discharge measurements are subject to error and this error is directly related to the accuracy of the instruments and methodologies used to measure velocity, depth, and cross-sectional area.

The potential bias that may be introduced by using velocity measurements obtained by different devices could, as Brauer points out, seriously affect the validity of results obtained through specific gage analysis. He attempts to show the pitfalls associated with the use of a non-homogeneous data set by showing a sudden jump in the stage-discharge relationship in 1930, when the velocity measurements were switched from double floats to Price current meter or ADCP. Unfortunately, this jump is demonstrated visually (Figure 10 in his paper) with no statistical test reported to show that this is not an artifact of random noise. Brauer claims that once the data are separated into two groups, demarcated by the year 1930, one gets two

approximately horizontal lines. Again, no statistical tests to determine whether the slopes are zero or not are absent. Results of such tests are needed before one can conclude that the increasing trend lines Pinter et al (2001) found were an artifact of the change in the velocity measuring device.

Brauer brings out a valid concern about the accuracy of the discharge measurements done in the early years. He also raises the point that seasonal vegetation and timing of the floods can affect the stage-discharge relationship. Other factors of concern he raises are the effect of water temperature and sediment load. These are valid concerns and must be taken into account when looking at the dynamics of stage-discharge relationship over time.

One of the most important contributions in Brauer's paper is his analysis of stage-discharge relationship based on actual observed discharges. These data points are more homogeneous than those obtained from specific gage analysis as far as their variance is concerned and can be considered devoid of simultaneity bias and other such artifacts. Brauer goes one step further and actually fit regression lines to the data and reports results of statistical significance tests, which is a vast improvement from what Pinter et al (2001) have done. One drawback in this analysis is the fact that no information is given as to how the discharge values were grouped into the categories such as 100,000 cfs and 150,000 cfs. According to his data, no statistically significant trend can be found for discharges over 300,000 cfs. These results are more believable than those reported by Pinter et al because the underlying data do not have some of the drawbacks of data obtained via specific gage analysis.

REFERENCES

Brauer, E.D., 2009. "The Limitations of Using Specific Gage Analysis to Analyze the Effect of Navigation Structures on Flood Heights in the Middle Mississippi River," *2009 D Paepe-Willams Award Contest*, Permanent International Association of Navigation Congress.

Fuller, W.A., 1987. *Measurement Error Models*, New York, John Wiley.

Jemberie, A.A., Pinter, N. and Remo, J.W.F., 2008. "Hydrologic history of the Mississippi and Lower Missouri Rivers based upon a refined specific-gage approach," *Hydrological Process*, Wiley InterScience, <www.interscience.wiley.com>, DOI: 10.1002/hyp.7046.

Judge, G.G, Griffiths, W.E., Hill, R.C., Lutkepohl, H., Lee, T., 1985. *The Theory and Practice of Econometrics*, 2nd Edition, New York, John Wiley.

Pinter, Thomas, and Wlosinksi, 2001. "Assessing Flood Hazard on Dynamic Rivers," *Transactions of the American Geophysical Union*, 82, pp. 333,338 – 339.

Pinter and Thomas, 2003. "Engineering modifications and changes in flood Behavior of the Middle Mississippi River," in *At the Confluence: Rivers, Floods and water Quality in St. Louis*, Robert E. Criss and David A. Wilson, Editors, Missouri Botanical garden Press, pp. 96-109.

Pinter, N. Jemberie, A.A., Remo, J.W.F, Hein, R.A., and Ickes, B.A., 2009. "Cumulative Impacts of River Engineering, Mississippi and Lower Missouri Rivers," *River Research and Applications*, Wiley InterScience, <www.interscience.wiley.com>, DOI: 10.1002/rra.1269