International Environmental Modelling and Software Society (iEMSs)
2010 International Congress on Environmental Modelling and Software
Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada
David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova (Eds.)
http://www.iemss.org/iemss2010/index.php?n=Main.Proceedings

Set-membership approach for identification of the uncertainty in power-law relationships: the case of sediment yield

Karel J. Keesman[a], Lachlan Newham[b] and Tony Jakeman[b]

[a]*Systems and Control Group, Wageningen University, Bornse Weilanden 9, 6708 WG Wageningen, The Netherlands (karel.keesman@wur.nl)*

[b]*iCAM, The Australian National University, Canberra ACT 0200, Australia*

Abstract: Power laws are used to describe a large variety of natural and man-made phenomena. Consequently, they are used in a wide range of scientific research and management applications. In this paper, we focus on the identification of uncertainty bounds on a power law relationship from experimental data, using a bounded-error characterization. These bounds can subsequently be used as constraints in e.g. optimization and scenario studies. The basic so-called set-membership approach involves outlier identification and removal, feasible parameter set estimation, evaluation of the feasible model output set and tuning of the error bounds. As an example we examine scattered sediment yield versus catchment (or watershed) area data of Wasson, (1994). The key result of this is an appropriate unfalsified relationship between sediment yield and catchment area with uncertainty bounds.

Keywords: identification, set-membership, power laws, sediment

1 INTRODUCTION

In a wide range of science and in many applications power laws are used to describe natural and man-made phenomena in a quantitative way (eg Deng and Jung, 2009; Millington et al., 2009). In particular, power law distributions are widely applied in earth sciences, linguistics, biology, economics and social sciences, when relating sizes to the frequency of occurrence. This distribution describes the phenomenon that large is rare and small is common. In power law distributions, the exponent in the power law function is negative. But it is certainly not limited to this. For instance, an exponent of 0.5 gives a square root function describing the free outflow from a tank, as is commonly derived from Bernouilli's law. Moreover, an exponent greater than 1 gives rise to a kind of exponential growth, as is frequently seen in biology. In fact, many well-known laws in physics are expressed in terms of a power law function, for instance, Stefan–Boltzmann law, Inverse-square laws of Newtonian gravity and Electrostatics, van der Waals force model, Kepler's third law, Square-cube law (ratio of surface area to volume), but also Pareto's principle follows a power law function. In the pre-computer era, scientists plotted all kind of phenomena on a log-log scale to arrive at a linear relationship between the variables, which they most often found. Hence, the power law is a fundamental way of describing a wide range of relationships.

Apart from some fundamental relationships in physics, most frequently power laws are derived from experimental data. Experimental data always contain measurement and sampling errors, which are usually characterized in statistical terms. Consequently, the estimates of the power law parameter are of a stochastic nature. However, for instance, in

case of limited data or after some non-linear transformation of the data, the presumed stochastic characterization is not always valid. Hence, as an alternative to a stochastic characterization a so-called bounded-error characterization, also called set-membership approach, has been proposed in the last decades.

The objective of this paper is to present a set-membership approach to the identification of error bounds on a power law and to provide unfalsified bounds on the data of Wasson (1994) using a power law relationship and a bounded-error characterization. The Wasson (1994) data provides a convenient example but the technique could be applied to any number of similar data sets.

2 BACKGROUND

Power-law functions are polynomials in a single variable, that is

$$f(x) = ax^k + o(x^k) \tag{1}$$

where a, k are real constants and $o(x^k)$ is an asymptotically small function. The parameter k is called the scaling exponent, since a typical property of a power law is the scaling invariance. To show the scaling invariance of power laws, let x be multiplied with a constant c then $f(cx) = a(cx)^k = c^k f(x) \propto f(x)$. In other words, multiplication with a constant does not change the shape of the function. Hence, power laws are explicitly used to describe the scaling behavior of natural processes. Allometric scaling laws, for instance, are frequently used to describe the relation between biological variables and thus are some of the best known power-law functions in nature.

When dealing with an experimental data set (y, x), the term $o(x^k)$ is replaced by a deviation or error term e, so that

$$y = ax^k + e \tag{2}$$

where y is the measured (dependent) variable in (2). In what follows, and from the viewpoint of parameter estimation, Eqn. (2) is also called a nonlinear regression, with unknown parameters a and k.

Notice from (2) that when k is given, a can be simply estimated from the resulting linear regression using ordinary least-squares estimation. On the contrary, the estimation of the exponent k is not so easy. There are many ways of estimating the scaling exponent in a power law from data. However, not all of them yield unbiased and consistent estimates. A commonly applied technique is to apply a (natural) logarithm transformation to the deterministic part of (2), which results in the linear regression

$$\ln y = \ln a + k \ln x \tag{3}$$

It is well-known that logarithmic transformation leads to distortion of the error e (see e.g Barlett, 1947; Box and Cox, 1962). If the original error e is normally distributed, after logarithmical transformation of the data it becomes log-normally distributed. However, in many cases with limited data this assumption about normality of the data is questionable and cannot be thoroughly tested. Considering a stochastic nature of e, and more particular assuming a Gaussian distribution, the most reliable estimation techniques are often based on maximum likelihood methods.

In this paper, given a limited data set - as is quite common in practice - we will follow an alternative route that is based on so-called set-membership estimation (Walter, 1990; Norton, 1994, 1995; Milanese *et al*, 1996). Let us shortly summarize this approach. Consider hereto the following non-linear regression type of model in vector form,

$$\mathbf{y} = \mathbf{F}(\vartheta) + \mathbf{e} \tag{4}$$

where $\mathbf{y} \in \mathbb{R}^N$ contains the observed output data, $\mathbf{F}(\vartheta)$ is a non-linear vector function mapping the unknown parameter vector $\vartheta \in \mathbb{R}^m$ into a noise-free model output $\hat{\mathbf{y}}$. The

error or information uncertainty vector **e** is assumed to be bounded in a given norm. In what follows, we assume that

$$\|\mathbf{e}\|_\infty \leq \varepsilon \tag{5}$$

where ε is a fixed positive number. Hence, a measurement uncertainty set (MUS), containing all possible output measurement vectors consistent with the observed output data and uncertainty characterization, is defined as

$$\Omega_y := \{\tilde{\mathbf{y}} \in \mathbb{R}^N : \| \mathbf{y} - \tilde{\mathbf{y}} \|_\infty \leq \varepsilon\} \tag{6}$$

This set is a hypercube in $\mathbb{R}^N$. Let the set

$$\Omega_\vartheta := \{\vartheta \in \mathbb{R}^m : \| \mathbf{y} - \mathbf{F}(\vartheta)\|_\infty \leq \varepsilon\} \tag{7}$$

define the feasible parameter set. Then, the set-membership estimation problem is to characterize this feasible parameter set (FPS), which is consistent with the model (4), the data (**y**) and uncertainty characterization (5)-(6).

Hence, instead of trying to find the optimal parameter vector as in an ordinary least-squares approach, our goal now is to find the set with feasible parameter vectors that are consistent with the model and the data with related error bounds. Hence, we will not consider the measurements as such but define intervals for each measurement. This approach avoids the distortion of the original probability density function after some non-linear transformation, because only bounds are considered. Furthermore, a symmetric bound in the log-log space introduces automatically skewed error bounding in the original space, which seems to be natural when considering data which most likely can be described by a power law.

From the set-membership literature, it is well-known that for the linear regression case, the FPS is a polytope found from the intersection of N (number of data points) strips in the parameter space. This will also be demonstrated in our example case, when working in the log-log space. It can even be shown that the well-known weighted least-squares techniques can be used to solve the bounded linear regression problem (Milanese, 1995; Keesman, 1997). But, in general, the FPS can be a complex, and even unconnected, set (see Keesman, 2003, for details and possible solutions)

3 APPLICATION

In this study we examine a data set collated by Wasson (1994) of sediment yields versus catchment area. The Wasson (1994) data, was collated from numerous studies of long-term sediment yields in south east Australia. The data shows a high level of scatter – a function of (i) high inherent spatial and temporal variability of sediment yield across south east Australia; and (ii) the use of a range of different underlying methods to estimate sediment yields.

Our particular interest in the analysis of this data was to identify likely upper and lower bounds of plausibility for sediment yield estimates. Such information is invaluable in informing the development and testing of dynamic, semi-distributed (spatially) models of sediment generation e.g. Newham et al. (2004).

The following figures show the Wasson (1994) data in original and log-log space.

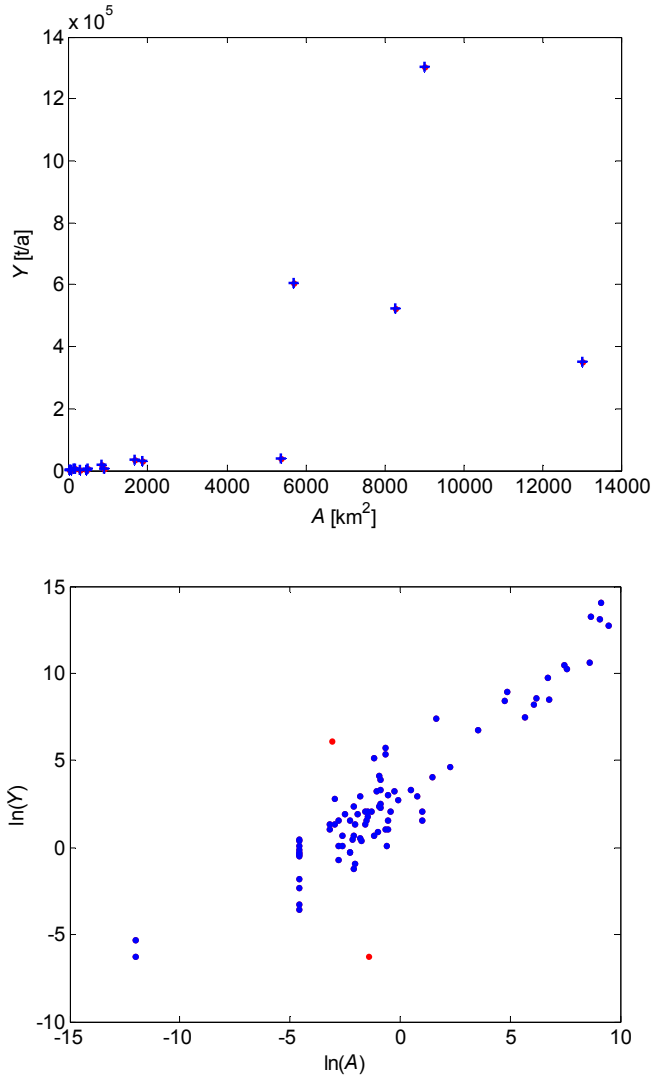

Figure 1. Data set (from Wasson, 1994), upper panel: original space and bottom panel: log-log space.

Presume that the catchment area-sediment yield data of Figure 1a can be described by the power law,

$$Y = aA^k \tag{8}$$

where a and k are unknown parameters. Given these data (Figure 1a,b) and the model (8), our objective is to find an appropriate uncertainty description of the sediment yield Y for a given catchment area A, preferably in a log-log space.

Let us start by applying a natural logarithmic transformation of the power law (8). Consequently,

$$\ln Y = \ln a + k \ln A \tag{9}$$

from which we define $\alpha_1 := \ln a$ and $\alpha_2 := k$. The ordinary least-squares (OLS) estimates (indicated by a hat) and corresponding covariance matrix of the estimation error (Σ) are given by

$$\hat{\alpha}_1 = 3.3125, \qquad \hat{\alpha}_2 = 0.9177; \qquad \Sigma = \begin{pmatrix} 0.0353 & 0.0012 \\ 0.0012 & 0.0020 \end{pmatrix} \tag{10}$$

However, as mentioned before, it is well-known that a logarithmic transformation leads to distortion of the error. Hence, the assumption about normality of the log transformed error is questionable and, because of the limited size of the data set (see Figure 1), cannot even be thoroughly tested. Consequently, the covariance matrix in (3) cannot be directly interpreted, which limits the possibilities for a direct uncertainty analysis. Moreover, due to the error distortion the estimates become biased. Hence, bias correction must be applied to correctly estimate the unknown parameters.

The set-membership approach, as presented in Section 2, avoids these obstacles, since we focus only on the calculation of the bounds and not at all on the probability distributions. However, given the linear regression (9) and the data in Figure 1b, the key question here is how to choose the error bound ε (see (5)). Notice that outliers, with inappropriate bounds, can easily lead to an empty feasible parameter set (FPS). Hence, the first step is to remove possible outliers. Keesman and van Straten (1989a) suggested a re-iterative min-max estimation, where the maximum error is plotted against the iteration number. After each min-max estimation step in a specific iteration, the data point at which the maximum error occurs is removed and the procedure is repeated. The result of such a procedure for the given data set is presented in Figure 2.

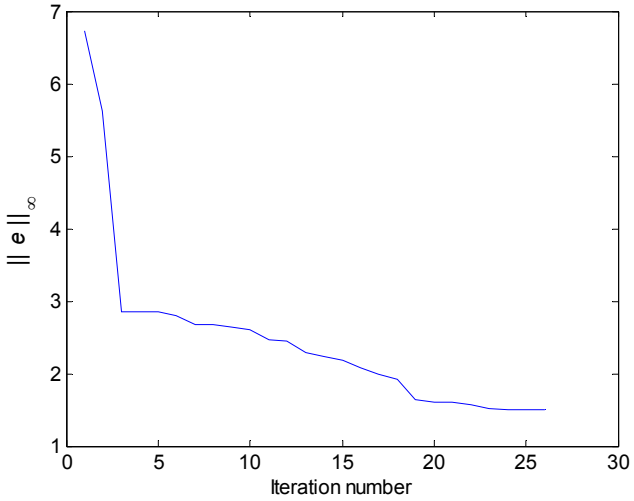

Figure 2. Results of reiterative min-max estimation.

Figures 1 and 2 suggest that the data set basically contains two possible outliers (as indicated in Fig. 1b in red) and that a sufficient error bound would be 3 t/a. The min-max estimate from the third iteration, thus after removal of two possible outliers, is given by: α_1 = 3.5237 and α_2 = 0.9072, with maximum error of 2.85 t/a. Hence, choosing the error bounds on the basis of this maximum error, would degenerate the FPS to a singleton. The estimation results related to a constant error bound of 3 t/a and using exact (see e.g. Walter and Piet-Lahanier, 1989; Mo and Norton, 1990) and approximate (Monte Carlo based) bounding techniques (Keesman and van Straten, 1989b; 1990) indicate that, for this error bound, the feasible model output set (FMOS) does not fully reflect the uncertainty in the measurements (not shown here). This would be acceptable when the data points not covered by the FMOS could be considered as outliers. However, in this case there is no evidence to do so. Hence, in the next step, we will increase the error bounds such that the FMOS contains (most of) the measurements.

The variation in the data set, in particular for ln(A) = -4.6 and on the intervals [–2.12, –1.96], [–0.67, –0.56] (see Figure 1b), can be estimated from the standard deviations in ln(Y). For each of these regions the standard deviations have been estimated as 1.40, 1.49 and 2.33, respectively. Consequently, an error bound of 5 t/a has been chosen to reflect the

3σ-bound for the individual measurements in this data set. The set-membership estimation results are presented in Figures 3 and 4. Notice from Figure 3 that, given the uncertainty in the data, the scaling exponent can possibly be smaller and larger than 1. As expected, the FPS indicated by blue dots and constrained by lower (green) and upper (red) bounds contains the min-max estimate (α_1 = 3.5237 and α_2 = 0.9072). Increasing the error bound will thus lead to a larger FPS. Hence, it reflects the larger uncertainty considered in the data. Notice that the FMOS in Figure 4 does contain almost all of the measurements and thus we may consider these results as appropriate for further evaluation. For instance, the (interpolated) bounds could be used in the identification of an erosion model, using bounded information, i.e. basically taking into account constraints instead of point measurements.

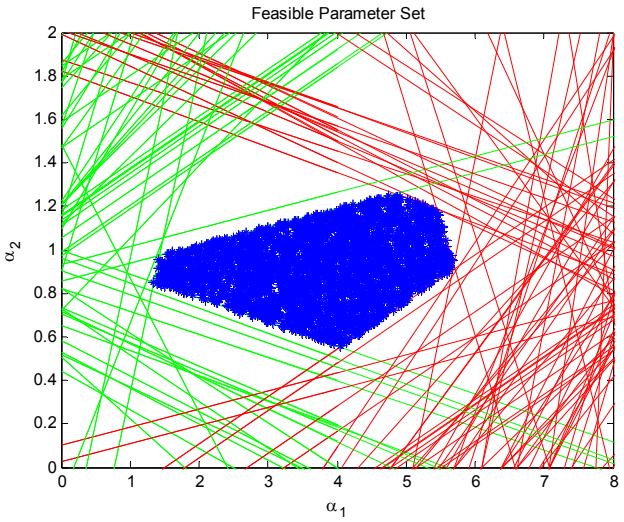

Figure 3. Feasible parameter set related to error bound of 5 t/a.

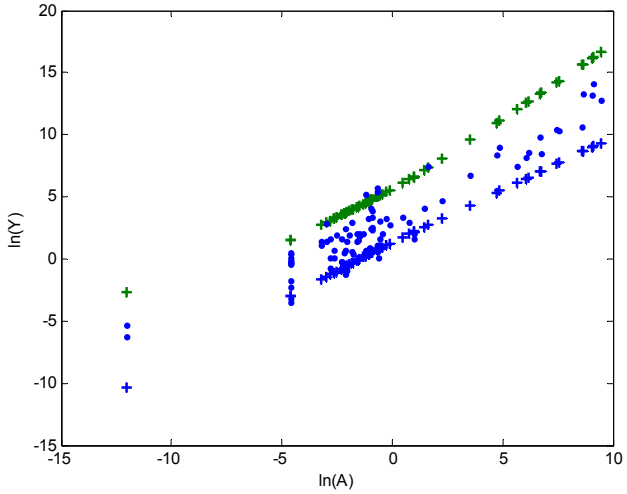

Figure 4. Feasible model output bounds (upper bound: green +; lower bound: blue +) related to error bound of 5 t/a.

4 DISCUSSION

Considerable challenges exist in the identification of feasible parameter sets for modelling environmental data. This is particularly the case for power law relationships such as those applied to water quality data.

It is interesting to see that the power law, $Y = aA^k$, is a solution to the differential equation,

$$\frac{dY}{dA} = k\frac{Y}{A}, \quad Y(0) = 0 \tag{11}$$

In other words, since $\frac{A}{Y}\frac{dY}{dA} = \frac{dY/Y}{dA/A} = k$, the relative or normalized slope of the relation between catchment area and sediment yield is constant and equal to the scaling exponent.

Notice that power law functions, as in (1), describe the static, non-linear relationship between variables. In this paper it has been shown that, given bounded-error data, a Monte Carlo-based bounding technique, also known as the Monte Carlo Set-membership Method (MCSM), can solve the parameter estimation problem. However, approximate (Monte Carlo based) bounding techniques are also applicable to the identification of dynamic, non-linear simulation models (see e.g. Keesman and van Straten, 1990). As for all other non-linear deterministic or stochastic estimation methods, the Monte Carlo-based bounding technique is practically constrained to cases with a limited number of unknown parameters. This curse of dimensionality is, in fact, an issue in many estimation and optimization problems. Hence, reduction of the problem via e.g. time scale decomposition or parameter space decomposition is crucial (Keesman, 2002).

5 CONCLUSIONS

A bounded-error characterization leads either to an empty set, a singleton or a (non-convex, not even connected) set of parameter vectors, using deterministic algorithms. As such, it directly reflects the uncertainty in the model and in the data without statistical computations. In particular for data sets of limited size, for which statistical properties are difficult to verify, the set-membership approach provides a good alternative. Given the Wasson (1994) catchment area-sediment yield data and a power law relationship with unknown coefficients, we were able to derive unfalsified model-based bounds on the data for use in constrained optimization and scenario studies.

REFERENCES

Bartlett, M.S., The use of transformations, *Biometrics, 3*, 39-52, 1947.

Box, G.E.P. and D.R. Cox, The analysis of transformations, *Journal of the Royal Statistical Society, Series B (Methodological), 26*, 211-252, 1964.

Deng, Z-Q. and H-S. Jung, Scaling dispersion model for pollutant transport in rivers, *Environmental Modelling & Software, 24*, 627-631, 2009.

Keesman, K.J. and G. van Straten, Identification and prediction propagation of uncertainty in models with bounded noise, *International Journal of Control, 49*(1), 2259-2269, 1989a.

Keesman, K.J. and G. van Straten, Membership-set estimation using random scanning and principal component analysis, *Mathematics Computers and Simulation, 32*, 535-544, 1989b.

Keesman, K.J. and G. van Straten, Set Membership Approach to Identification and Prediction of Lake Eutrophication, *Water Resources Research, 26*(11), 2643-2652, 1990.

Keesman, K.J., Weighted least squares set estimation from l∞-norm bounded noise data, *IEEE Transactions on Automatic Control , 42*(10), 1456-1459, 1997.

Keesman, K.J., State and parameter estimation in biotechnical batch reactors, *Control Engineering Practice, 10*(2), 219–225, 2002.

Keesman, K.J., Nonlinear-Model Case in Bound-based Identification, in Control Systems, Robotics and Automation, edited by H. Unbehauen, in *Encyclopedia of Life Support Systems (EOLSS)*, UNESCO Publishing-Eolss Publishers, Oxford, UK (http://www.eolss.net), 2003.

Milanese, M., Properties of least-squares estimates in set membership identification, *Automatica*, *31*, 327-332, 1995.

Milanese, M., J.P. Norton, H. Piet-lahanier and E. Walter (Eds.), *Bounding Approaches to System Identification*, Plenum Press, NY., 1996.

Millington, D.A., J. Wainwright, G.L.W. Perry, R. Romero-Calcerrada, and B.D. Malamud, Modelling Mediterranean landscape succession-disturbance dynamics: A landscape fire-succession model, *Environmental Modelling & Software, 24*, 1196-1208, 2009.

Mo, S.H. and J.P. Norton, Fast and robust algorithm to compute exact polytope parameter bounds, *Mathematics Computers and Simulation, 32*, 481-493, 1990.

Newham, L.T.H., R.A. Letcher, A.J. Jakeman and T. Kobayashi, A Framework for Integrated Hydrologic, Sediment and Nutrient Export Modelling for Catchment-Scale Management, *Environmental Modelling and Software, 19*, pp. 1029-1038, 2004.

Norton, J.P., Bounded-error estimation: issue 1, *International Journal on Adaptive Control and Signal Processing, 8*(1), 1-118, 1994.

Norton, J.P. (1995) Bounded-error estimation: issue 2, *International Journal on Adaptive Control and Signal Processing, 9*(1), 1-132, 1995.

Walter, E. (Ed.), Parameter identifications with error bounds. Special Issue *Mathematics Computers and Simulation, 32*(5-6), 447-607, 1990.

Walter, E. and H. Piet-Lahanier, Exact recursive polyhedral description of the feasible parameter set for bounded-error models, *IEEE Transactions on Automatic Control, Vol. 34*, 911-915, 1989.

Wasson, R. J., Annual and decadal variation of sediment yield in Australia, and some global comparisons, Variability in stream erosion and sediment transport (Proceedings of the Canberra Symposium, December 1994) *IAHS Publication no. 224*, 269-279, 1994.