Robust analysis of trends in noisy data

G.Verdoolaege[1,2], A.Shabbir[1,3] and G.Hornung[1]

[1] *Department of Applied Physics, Ghent University, B-9000 Ghent, Belgium*
[2] *Laboratory for Plasma Physics, Royal Military Academy, B-1000 Brussels, Belgium*
[3] *Max Planck Institute for Plasma Physics, D-85748 Garching, Germany*

Abstract. Detection and quantification of trends of key quantities in terms of a set of 'predictor' variables is a common task in fusion data analysis, as in many other areas of science, for model building and experimental planning. In fusion science, the standard way to handle the corresponding regression analysis problem is by means of a linear or power-law regression function and ordinary least squares (OLS) to perform the fit. There are essentially two issues with this common approach that we intend to address in the present work. On the one hand, OLS is a very simple technique that is not suitable in the presence of complex uncertainties on the measured data. Its assumptions can be overly simplifying, e.g. when the measurements originate from multiple diagnostics or experiments, when the predictor variables are affected by considerable uncertainty, or when the data contain outliers. This often leads to erroneous estimates of the regression parameters, which, moreover, greatly depend on the adequateness of the proposed regression function. On the other hand, the measurements used in the regression analysis are often averages over a time window or over multiple occurrences of the phenomenon under study (e.g. certain plasma instabilities). Effectively, this means that potentially valuable information in the data is discarded. Whenever a measured quantity is subject to considerable fluctuation or measurement noise – a common situation in fusion – it can be very beneficial to consider the probability distribution of the quantity instead of its average. Indeed, whereas the average provides a first-order summary of the distribution, the actual distribution contains information about the typical spread of the measurements, symmetry around the mean, frequency of extreme values, etc. We have developed the method of geodesic least squares regression (GLS) that does not depend on the overly simplifying assumptions of OLS, by exploiting the full probability distribution of the regression variables. In the present contribution, the method is applied to regression analysis of plasma energy confinement and energy of edge-localized modes (ELMs). We demonstrate the strongly improved robustness of GLS compared to conventional OLS, with respect to both uncertainty in the data set and in the regression model. We show that this is because GLS performs regression between probability distributions. Finally, we highlight the significant potential and ease of use of GLS for general-purpose analysis of trends in fusion science.