

The Gamma statistic converges to the noise relative to an unknown nonlinear function



## A New Era in Statistical Reasoning

The *winGamma*<sup>TM</sup> software is a suite of nonlinear modeling and analysis tools, organized around the Gamma Test. The Gamma Test estimates the noise in a data set relative to all smooth functions. It brings together two important ideas - it is a *mean squared error estimator*, that applies to all *smooth nonlinear models*. Please take a moment to consider each of these two concepts.

### Error Estimation

Error estimation plays a central role in all statistical reasoning. For example:

- In sampling and polling theory, the number of samples that must be collected is decided using the variance, an estimate of the MSE of samples.
- In experimental design, the statistical significance of the results is decided based on t-tests and other statistics, which are calculated from estimates of the error variance.
- In linear regression studies, the correlation coefficient and similar MSE estimators are used to calculate “goodness of fit” of the model. If MSE is zero, the data lies along the regression line, so there is clearly a strong relationship between input and output. If the error estimates are high, the regression line is meaningless.

Statisticians use error estimators to decide what is important and what is not. How accurate are these polling results? Is there really a difference between the control and experimental groups? Which of these variables are important for prediction? Do I have enough data? These are some of the questions that are answered by mean squared error estimators.

### Smooth Nonlinear Models

Intuitively, smooth means what it sounds like - not jagged. Mathematically, a smooth function is one whose first and second derivatives are bounded (finite at all points).

Smooth nonlinear models are important because there are so many of them. Theoretically speaking, there are infinitely many smooth curves which could relate any two variables. Practically speaking, straight lines are rare in nature. Natural processes almost always produce smooth curves.

The infinite variety of the natural world is largely created by smooth nonlinear processes. The variety is possible because control operates on a process of growth or change. Slight differences in initial conditions may lead to infinite variety in the outcomes. These process may be completely impossible to predict in the long term, but quite predictable in the short term. This is the world of chaos theory - the study of change processes that are governed by nonlinear functions.

Although straight lines are rare in nature, measuring goodness of fit through error estimation is such a powerful technique that linear statistics have changed the world. The Gamma Test brings the power of standard statistical reasoning to the vastly larger realm of smooth models. One academic reviewer described it as “the holy grail of nonlinear modeling”. Another reviewer wrote, “The Gamma Test will play a significant role in every area of statistical signal processing.”

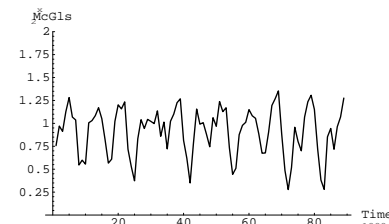
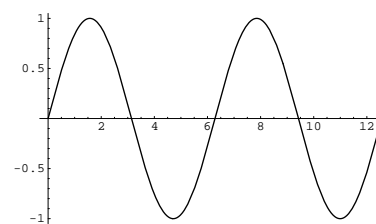
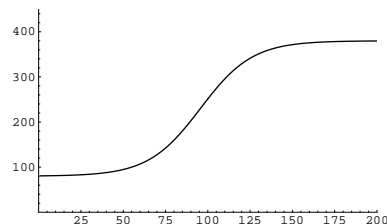
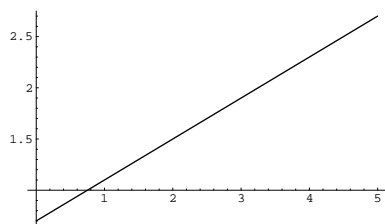
# Nonlinear Goodness of Fit

For nonlinear modeling, there are numerous techniques for curve fitting, including neural networks, local linear regression, and support vector machines. But in the past, there has been no way to estimate goodness of fit for nonlinear models without knowing the model.

Now there is. If gamma is small there is a strong predictive relationship between the input variables and the output. If gamma is large there is no smooth predictive

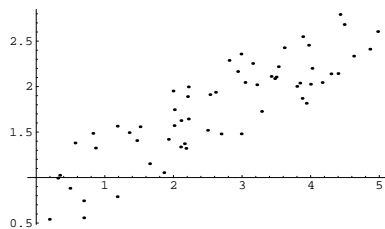
relationship between inputs and output - the inputs are irrelevant to the output. In this way Gamma is similar to the MSE around a regression line.

However, linear regression measures goodness of fit to a line. Gamma measures goodness of fit to any and all smooth curves. Here are four simple examples of functions that gamma can find in data:

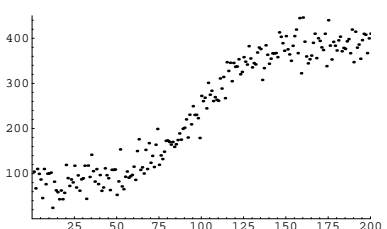


All four of these graphs represent deterministic functions - they are a line, a logistic function, a sine curve, and the Mackey-Glass time delay differential equation. *The Gamma Test recognizes all of these relationships, and uncountable others.* Given data sets generated by these functions, in all four cases the Gamma Test correctly estimates the noise to be zero. What is remarkable about the Gamma Test is that it can compute this noise variance even though the underlying smooth function is unknown.

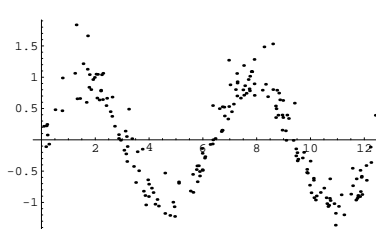
If a random variable is added to any of these functions to simulate noise, the Gamma Test will produce an accurate estimate of the noise variance if given sufficient data. Gamma does not measure how closely the data fits a line, or any predetermined shape. Instead it distinguishes between noise and smooth relationships, so it does not depend on the shape of the function. This allows it to simultaneously measure goodness of fit against the entire class of smooth functions.



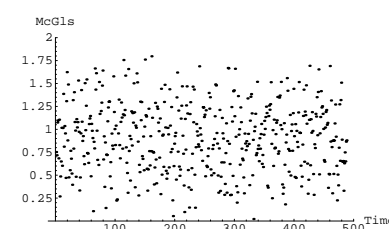
60 data points generated by adding a uniformly distributed random variable, with variance .075, to the line  $y = .4x + .7$ . The Gamma estimate of noise variance is .079



100 points generated from a logistic function, adding normally distributed noise with variance 625. The Gamma estimate of noise variance is 597



200 points generated by adding a lognormally distributed random variable with variance .075 to a sine curve. Gamma estimate of noise variance - .077



500 point series generated from the Mackey-Glass time-delayed differential equation, by adding normally distributed output noise with variance .075. Gamma estimate of noise variance - .073

# Building the Model

The Gamma Test estimates the noise in a data set, relative to the best smooth model that can be built using that data. Having this estimate makes it much easier to discover the model itself.

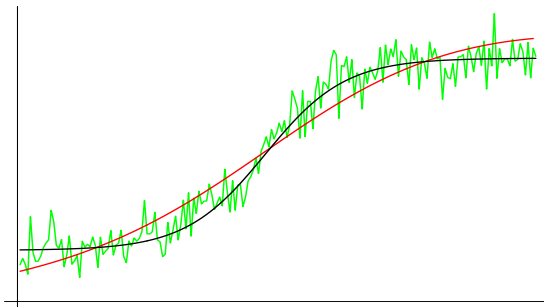
Neural networks are a leading method for building nonlinear models. A neural net is “trained” on a data set. What this means is that the net is presented with training data for which the inputs and outputs are known. The inputs are fed to the network and it calculates a predicted output. If the prediction is wrong, the training adjusts the network. As this process is repeated, the net’s predictions get more accurate. Eventually, the network can be made to predict the training data exactly.

Unfortunately, this approach has a serious weakness in the presence of noise - the problem is called overtraining. As the network is trained, it grows folds and

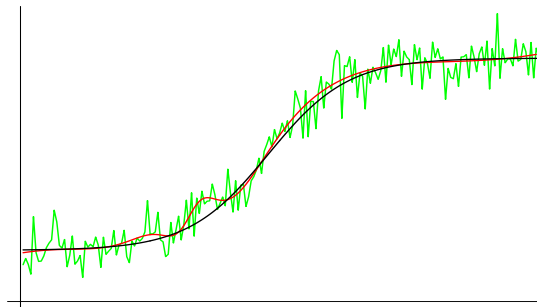
wiggles to fit itself to the noisy data. The predictions become better and better on the training data, but worse and worse when it is applied to new data.

This problem is not unique to neural networks - it appears in all forms of nonlinear model building. We are trying to draw a curve through the data points, that captures the essence of the data. But there are an infinite number of curves that can fit any finite data set. How do we know which is the right one? In particular, if the data is noisy, how do we avoid drawing our curve to fit the noise, rather than the underlying predictive function?

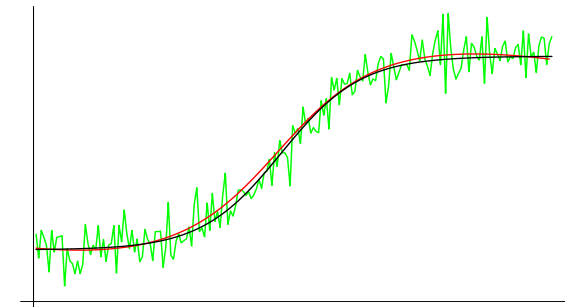
The Gamma Test solves this problem, by giving us an estimate of how closely the model *should* fit the data. Without the Gamma Test, there is an element of guesswork in any nonlinear curve fitting technique. With the Gamma Test, this guesswork is eliminated. It is possible to build models of unprecedented accuracy.



*Too Little Training:* This shows the performance of a neural network at the beginning of training. The training data is shown in green. The data was created by adding a measured amount of noise to a logistic curve. The original curve is shown in blue, so that is what we are trying to model. The predictions of the neural net are shown in red. After a few rounds of training, the network has got the general shape, but it hasn't yet modeled the curve very accurately.



*Too Much Training:* This shows the performance of a neural network that has been overtrained. As the network is adjusted to fit the data, it begins to fit the noise rather than the underlying function.



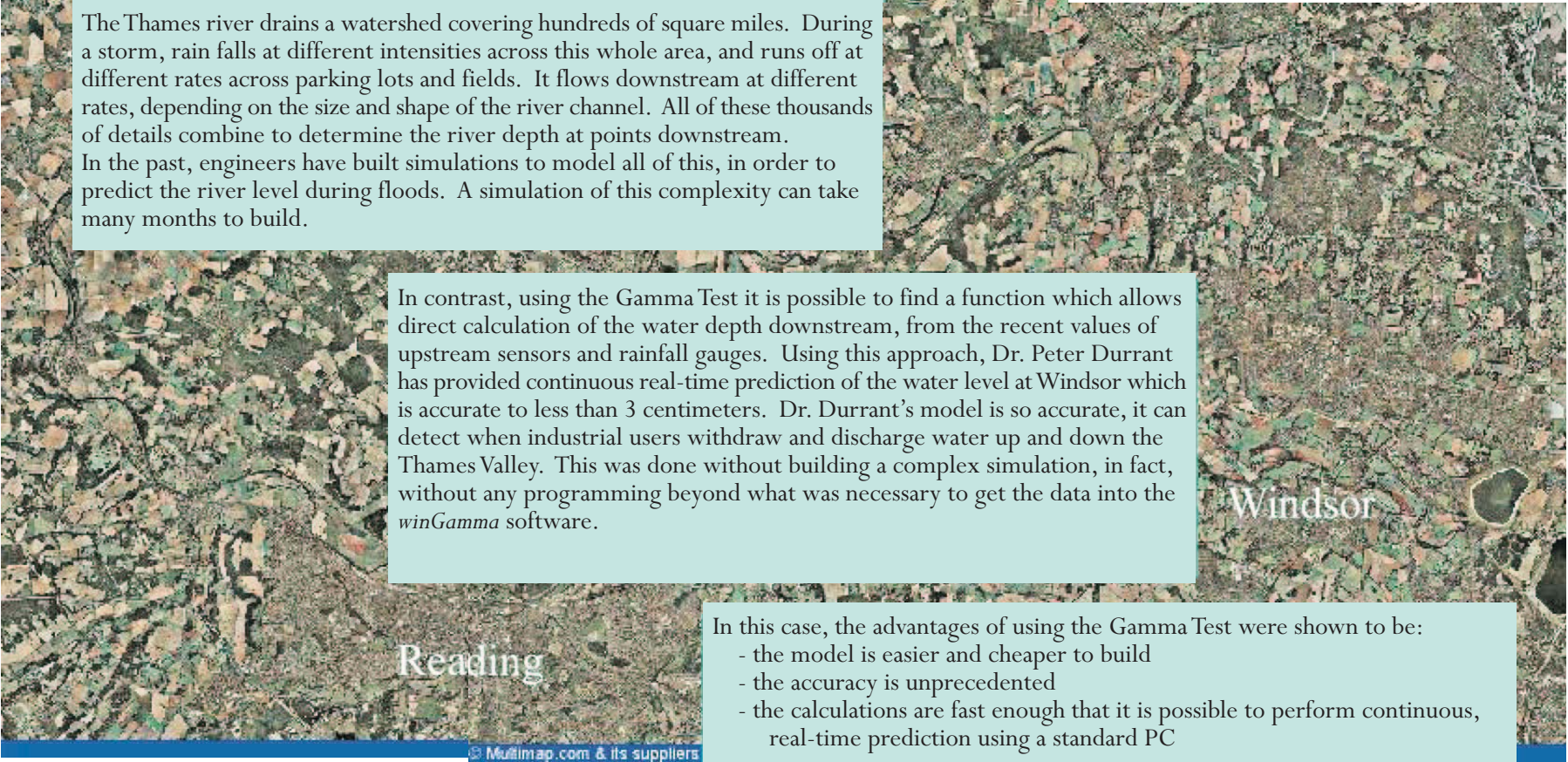
*Just Right Training:* This shows the performance of an optimally trained neural net. The original underlying function has been modeled almost exactly. **This is the model we want to build.** This model was built by training the network until its error was equal to the noise estimate provided by the Gamma Test

# Gamma Test Applications

The Gamma Test can be applied in place of traditional statistics whenever the variables have a reasonably continuous range of measurement, and the underlying causal dynamics are smooth. It has been successfully applied to flood prediction, marketing research, economic forecasting, signal processing, analysis of gene expression data, and control of dynamic systems.

Because the Gamma statistic is formally a noise variance, it is a fundamental value which can be used in many forms of statistical reasoning. The most

direct application is analogous to the MSE, or residuals in linear regression. But there are many other ways to use this statistic. For example, it has been used to detect weak localized signals against a background of heavy noise. It has been used in this way in astronomical research, to examine the infrared map of the universe. This application has discovered 50% more galaxies than had been previously identified in the IR map.



The Thames river drains a watershed covering hundreds of square miles. During a storm, rain falls at different intensities across this whole area, and runs off at different rates across parking lots and fields. It flows downstream at different rates, depending on the size and shape of the river channel. All of these thousands of details combine to determine the river depth at points downstream. In the past, engineers have built simulations to model all of this, in order to predict the river level during floods. A simulation of this complexity can take many months to build.

In contrast, using the Gamma Test it is possible to find a function which allows direct calculation of the water depth downstream, from the recent values of upstream sensors and rainfall gauges. Using this approach, Dr. Peter Durrant has provided continuous real-time prediction of the water level at Windsor which is accurate to less than 3 centimeters. Dr. Durrant's model is so accurate, it can detect when industrial users withdraw and discharge water up and down the Thames Valley. This was done without building a complex simulation, in fact, without any programming beyond what was necessary to get the data into the *winGamma* software.

In this case, the advantages of using the Gamma Test were shown to be:

- the model is easier and cheaper to build
- the accuracy is unprecedented
- the calculations are fast enough that it is possible to perform continuous, real-time prediction using a standard PC





# Gamma Test Technical Description

The Gamma Test is a statistical technique which estimates the noise, or error variance, in a data sample. It is analogous to the sum of squared error in linear regression, but it works for any smooth function on continuous data.

Let  $x$  be the vector of input variables, and  $y$  be the output. Assume that the data is described by an underlying model of the form

$$y = f(x) + r$$

where  $f$  is some unknown smooth function, and  $r$  is a stochastic variable with mean zero and bounded variance. The function  $f$  is smooth if its first and second derivatives are bounded (finite at all points). The Gamma Test estimates what proportion of the variance of  $y$  is caused by the unknown function  $f$ , and what proportion is caused by the random variable  $r$ . Put another way,  $\Gamma$  is an estimate for the noise variance relative to the best possible smooth function  $f$ . Other moments of  $r$  can be estimated as well (e.g. kurtosis).

If  $\Gamma$  is small, then we know that the function  $f$  exists, and the output value  $y$  is largely determined by the input variables (vector  $x$ ). In that case, the *winGamma* software will automatically produce a predictive model. By supplying additional data, we can make this model arbitrarily close to the true function  $f$ .

On the other hand, if  $\Gamma$  is large, then  $y$  is primarily the result of random variation. This means that your ability to predict is limited by one of four conditions:

- 1) you are not measuring some important input variable,
- 2) noise is being introduced due to measurement error,
- 3) you don't have enough data to model a complex curve, or
- 4) there are discontinuities in the underlying causal function.

If the Gamma Test cannot produce a predictive model, you will know that one or more of these conditions holds.

For further information, send an email to [info@gammatest.com](mailto:info@gammatest.com), or visit our website at [www.gammatest.com](http://www.gammatest.com)



*Why stumble when you can see?*

# Anglo-American Chaos

Anglo-American Chaos is a corporation formed to exploit the power of the Gamma Test. We are looking for customers and partners who

- 1) are working with continuous processes, and
- 2) have data in which a number of inputs should predict an output.

If you fit this profile, we offer you ease and accuracy of prediction unlike anything you have experienced before.

Anglo-American Chaos provides consulting, data analysis, and software sales. If you are interested in the analysis of a nonlinear process, and you have data, take the next step and contact us.

Email:

[info@gammatest.com](mailto:info@gammatest.com)

