Available online at www.sciencedirect.com
ScienceDirect
Physica A 378 (2007) 498–504
PHYSICA A
www.elsevier.com/locate/physa

Modelling train delays with q-exponential functions

Keith Briggs[a,*], Christian Beck[b]

[a]Complexity Research Group, BT, Adastral Park, IP5 3RE, UK
[b]School of Mathematical Sciences, Queen Mary, University of London, Mile End Road, London E1 4NS, UK

Received 9 November 2006; received in revised form 29 November 2006
Available online 28 December 2006

Abstract

We demonstrate that the distribution of train delays on the British railway network is accurately described by q-exponential functions. We explain this by constructing an underlying superstatistical model.

Keywords: Superstatistics; q-exponentials; Fat-tail distributions; Delay distributions; Railway traffic

1. Introduction

Complex systems in physics, engineering, biology, economics, and finance, are often characterized by the occurrence of fat-tailed probability distributions. In many cases there is an asymptotic decay with a power law. For these types of systems more general versions of statistical mechanics have been developed, in which power laws are effectively derived from maximization principles of more general entropy functions, subject to suitable constraints [1–4]. Typical distributions that occur in this context are of the q-exponential form. The q-exponential is defined as $e_q(x):=(1+(1-q)x)^{1/(1-q)}$, where q is a real parameter, the entropic index. It has become common to call the corresponding statistics 'q-statistics'.

A possible dynamical reason for q-statistics is a so-called superstatistics [5]. For superstatistical complex systems one has a superposition of ordinary local equilibrium statistical mechanics in local spatial cells, but there is a suitable intensive parameter β of the complex system that fluctuates on a relatively large spatio-temporal scale. This intensive parameter may be the inverse temperature, or the amplitude of noise in the system, or the energy dissipation in turbulent flows, or an environmental parameter, or simply a local variance parameter extracted from a suitable time series generated by the complex system [6]. The superstatistics approach has been the subject of various recent papers [7–12] and it has been applied to a variety of complex driven systems, such as Lagrangian [13,14] and Eulerian turbulence [15,6], defect turbulence [16], cosmic ray statistics [17], solar flares [18], environmental turbulence [19], hydroclimatic fluctuations [20], random networks [21], random matrix theory [22] and econophysics [23].

*Corresponding author. Tel.: +44 1473 641 911; fax: +44 1473 642 161.
E-mail address: Keith.Briggs@bt.com (K. Briggs).

0378-4371/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2006.11.084

If the parameter β is distributed according to a particular probability distribution, the χ^2-distribution, then the corresponding superstatistics, obtained by integrating over all β, is given by q-statistics [1–4], which means that there are q-exponentials and asymptotic power laws. For other distributions of the intensive parameter β, one ends up with more general asymptotic decays [8].

In this paper we intend to analyse yet another complex system where q-statistics seem to play an important role, and where a superstatistical model makes sense. We have analysed in detail the probability distributions of delays occurring on the British rail network. The advent of real-time train information on the internet for the British network (http://www.nationalrail.co.uk/ldb/livedepartures.asp) has made it possible to gather a large amount of data and therefore to study the distribution of delays. Information on such delays is very valuable to the traveller. Published information is limited to a single point of the distribution—for example, the fraction of trains that arrive with 5 m of their scheduled time. Travellers thus have no information about whether the distribution has a long tail, or even about the mean delay. We find that the delays are well modelled by a q-exponential function, allowing a characterization of the distribution by two parameters, q and b. We will relate our observations to a superstatistical model of train delays.

This paper is organized as follows: first, we describe our data and the methods used for the analysis. We then present our fitting results. In particular, we will demonstrate that q-exponentials provide a good fit of the train delay distributions, and we will show which parameters (q, b) are relevant for the various British rail network lines. In the final section, we will discuss a superstatistical model for train delays.

2. The data

We collected data on departure times for 23 major stations for the period September 2005–October 2006, by software which downloads the real-time information webpage every minute for each station. As each train actually departs, the most recent delay value is saved to a database. The database now contains over two million train departures; for a busy station such as Manchester Piccadilly over 200,000 departures are recorded.

3. The model and parameter estimation

Preliminary investigation led us to believe that the model

$$e_{q,b,c}(t) = c(1 + b(q-1)t)^{1/(1-q)} \tag{1}$$

would fit well; here t is the delay, $0<q<2$ and $b>0$ are shape parameters, and c is a normalization parameter. We have $e_{q,b,c}(t) = c(1-bt) + O(t^2)$ as $t \to 0$ and $\log(e_{q,b,c}(t))/\log(t) \to 1/(1-q)$ as $t \to \infty$. These limiting forms allow an initial estimate of the parameters; an accurate estimate is then obtained by nonlinear least squares. We also have

$$\lim_{q\to 1} e_{q,b,c}(t) = c\,\exp(-bt), \tag{2}$$

so that q measures the deviation from an exponential distribution. An estimated q larger than unity indicates a long-tailed distribution.

We did not include the zero-delay value in the fitted models. Typically 80% of trains record $t = 0$, indicating a delay of 1 min or less (the resolution of the data). Thus, our model represents the conditional probability distribution of the delay, given that the train is delayed 1 min or more.

In order to provide meaningful parameter confidence intervals, we weighted the data as follows. Since our data is in the form of a histogram, the distribution of the height c_i of the bar representing the count of trains with delay i will be binomial. In fact, it is of course very close to Gaussian whenever c_i is large enough, which is the case nearly always. The normalized height $f_i = c_i/n$ (where n is the total number of trains) will therefore have standard deviation $\sigma_i = (nf_i(1-f_i))^{1/2}/n \approx c_i^{1/2}/n$. We used these values as weights in the nonlinear least squares procedure, and hence computed parameter confidence intervals by standard methods, namely from the estimated parameter covariance matrix. We find that typically q and b have a correlation coefficient of