Applying statistical analysis to understanding the dynamics of volcanic explosions

N. VARLEY¹, J. JOHNSON², M. RUIZ³, G. REYES⁴ & K. MARTIN⁵

¹Facultad de Ciencias, Universidad de Colima, Colima, Mexico
(e-mail: nick@ucol.mx)
²Department of Earth Sciences, University of New Hampshire, Durham, NH 03824, USA
(e-mail: jeff.johnson@unh.edu)
³Instituto Geofísico, Escuela Politécnica Nacional, Quito, Box 17-01-2759, Ecuador
(e-mail: mruiz@igepron.ec)
⁴RESCO, Universidad de Colima, Colima, Mexico (e-mail: gabrielr@cgie.ucol.mx)
⁵Department of Geology, University of South Florida, Tampa, FL 33620-5201, USA
(e-mail: kmartin@mail.usf.edu)

An erupting volcano is a complex system controlled by nonlinear dynamics and hence is difficult to model numerically. Statistical methods can be applied to explain behaviour or to aid the forecasting of future activity. The majority of previous studies have considered large-scale events: large explosive or effusive eruptions, with intervening long periods of repose. This has severely limited the size of the datasets and hence the significance of statistical results. In previous cases a simple Poisson model was applied, but often more sophisticated analysis methods are necessary to model the data. In this study, several statistical techniques are used to describe the data for smaller-scale events from four volcanoes. In each case study the events are relatively frequent explosions; this means that the datasets are large and thus allow a robust statistical analysis. First, time-series analysis is used to identify the presence of clustering or trends in the data. For stationary periods, the data are modelled in a probabilistic fashion, taking the survival function for increasing repose intervals and fitting different distributions to the data. Different types of events are identified, whose repose intervals have different distributions. This implies variation in the physics of the processes involved in the causation of the events. It is shown that activity can be divided into different periods based on the statistics, which can greatly aid in the construction of a model to explain the temporal evolution of eruptive activity. Contrasts between the volcanoes are highlighted, reflecting a variation in certain characteristics of their systems and the processes that dominate their activity.

Volcanic activity can be regarded as chaotic with no apparent form, but the use of statistical analysis can reveal hidden structure in behaviour patterns. Applied either spatially or temporally, it can be a powerful tool leading to future forecasting or improved models of the system. Denlinger & Hoblitt (1999) recognized and modelled cyclic behaviour in silicic volcanoes, which can be obvious, identified by seismicity or tilt measurements. However, longer-term patterns are often difficult to distinguish. The first stage of a probabilistic analysis is to determine if the data are stationary or if they are characterized by periodicity, clustering or a trend. The simple stream of events can be analysed as a time series. Next, a probability density function can be fitted to the repose interval data; the resulting model can provide important information on the processes that are generating the event. A dataset might be the historical record of a certain type of event, it could be a single variable, such as the occurrence of an event in time (e.g. eruptions of Volcanic Explosivity Index (VEI) > 2) or it could include another variable such as the magnitude.

When considering the probability of volcanic events, the first question to ask is whether events represent a stochastic time series, meaning that events can be represented as a random variable, indexed in time. In contrast, a deterministic model is characterized by no randomness in model parameters. Theoretically, by completely specifying a deterministic model, events may be forecast at any point in the future. Such models appear to be unachievable in natural systems.
Because volcanic processes are random in nature, time series of different types of data are stochastic and none can be entirely represented by a deterministic model.

A Poisson process will have an exponential distribution, which means that the occurrence rate, also known as the intensity function, is a constant independent of time. Early applications of statistics to volcanic activity used the concept of the point process (Reyment 1976). If the process of renewal is independent, that is each interval is independent of the previous one, and the distribution is exponential, then the system is controlled by a Poisson process. A renewal process is a generalization of the Poisson process and has independent identically distributed holding times, which can be thought of as the repair time after a previous failure of the system.

In the case of a volcano, the activity will usually increase with time, leading up to a large eruption, and then decline. This non-stationary behaviour can be observed on different scales. Therefore in the case of volcanic activity, the application of nonhomogeneous Poisson processes is generally more successful. These consider time trends in the data. Another variation that considers clustering is the generalized Poisson model, which treats the time series as a superimposition of individual sequences, each one a cluster in the data. The onsets remain distributed as a simple Poisson process.

One problem with studies of larger events considering longer time periods is that an inconsistency will usually exist within the dataset: the early data will be of much poorer quality compared with more recent activity reports, with a strong possibility of missing or badly reported events (De La Cruz-Reyna 1991). In each case, the datasets used in this study are precise and complete, giving more confidence to the relationships that are demonstrated. Further problems arise given that no definitive method exists for reporting the magnitude of an eruption. It is most logical to consider the energy released during the event; however, this is not easy to measure and there is a large discrepancy between the quality and availability of data at different volcanoes around the world. The interpretation of an event required to assign a value for the VEI can be rather subjective, when good quality data are unavailable. In this study, the relative magnitude of the event is quantified using the maximum amplitude of the associated seismic velocity trace. This provides a proxy for eruptive energy release and allows precise timing for the onset of the event.

**Brief background of each volcano**

**Volcán de Colima**

Volcán de Colima (19.51°N, 103.62°W; 3820 m) is in the western region of the Trans-Mexican Volcanic Belt (Fig. 1) and represents the currently active cone of the Colima Volcanic Complex. Activity is characterized by periods of effusive activity, often followed by explosive episodes. Since 2001 activity has increased, and up to June 2005 the volcano has been virtually in a constant state of eruption.

![Fig. 1. Location of four volcanoes. Photographs show typical explosive events from Volcán de Colima, Karymsky, Tungurahua and Erebus.](image-url)
**Tungurahua**

Tungurahua (1.45°S, 78.43°W; 5032 m) is one of the most active volcanoes of the Ecuadorean Andes (Fig. 1). Since the sizeable eruptive activity in late 1999 ended, Tungurahua has exhibited four eruptive cycles. The most recent eruptive cycle started in late May 2004, reached its climax in July, and continued until March 2005. It is characterized by minor bursts of activity (several explosions per day), which produced ash columns of altitudes no higher than 3 km above the crater.

**Karymsky**

Karymsky (54.1°N, 159.4°E; 1540 m) is an andesitic cone located in the central portion of Kamchatka’s main active arc (Fig. 1). It began its latest eruptive phase in January 1996 after 14 years of quiescence (Gordeev et al. 1997). Between 1996 and 1999, Karymsky’s behaviour consisted of discrete Strombolian explosions, with a frequency ranging from 5 to 20 events per hour.

**Erebus**

Mt. Erebus (77.55°S, 167.17°E; 3794 m) is a large stratovolcano located in Antarctica. The surface manifestation of its conduit system comprises a phonolitic lava lake with Strombolian explosions associated with the ascent and decompression of large gas slugs (Aster et al. 2003). The magma has a relatively low viscosity so the slugs can rise relatively unhindered through the magma. Activity has been variable, with a high rate observed between September and December 1984 and a low rate from December 1990 to December 1995 (Aster et al. 2003).

**Statistical studies of repose intervals between eruptions**

Various earlier studies showed that volcanic events (e.g. explosions, effusive emissions, or earthquake swarms) are largely random phenomena, but some sets present patterns, which can be related to models of their activity. The early work of Wickman (1976) examined the repose interval between events at many volcanoes and found different types of patterns within the datasets.

Medina Martinez (1983) performed the first study of the repose interval for documented historical eruptions at Volcán de Colima. A further, more detailed study considered events recorded with VEI values of 1–4 (De La Cruz-Reyna 1993). Events were subdivided into groups according to the VEI value and analysis was then performed on each group. Groups of events with VEI values of 2, 3 and 4, as well as groups with VEI ≥ 2 and VEI ≥ 3, all fitted a Poisson distribution. By examining a cumulative plot of the number of events, the period was divided into episodes each with a particular mean repose time. The distribution was also found to be Poisson within each period.

Much attention has been given to the identification of temporal or spatial clustering within the seismic record of volcano-tectonic events, particularly event clustering prior to volcanic eruptions. However, only a few studies have been carried out on the statistics of eruptions of smaller magnitude and shorter repose interval. An analysis of the seismic energy produced by explosions at Stromboli, Italy, produced a log-normal distribution with a very low variability in energy (less than two orders of magnitude; De Martino et al. 2004). An analysis of the data from a series of Vulcanian explosions at Soufrière Hills volcano, Montserrat, showed that the events fit a Weibull distribution for values of the repose interval less than the median value (Connor et al. 2003). However, if the long repose interval data in the tail were included, a much better fit was obtained with the log-logistic distribution. This is an interesting result, as this distribution implies the competition of different processes, an idea that has emerged from recent modelling of processes within the upper conduit (Melnik & Sparks 1999; Slezin 2003). The variation of repose interval in the Soufrière Hills data was from 2.77 to 33.7 h with a mean of 9.6 h (Connor et al. 2003).

**Methods**

To study the statistics of small eruptions, the timing of each event must be recorded. This can be carried out visually, perhaps using a webcam, infrasound, optical thermal sensors or, most commonly, with seismicity. The magnitude of each explosion can be quantified using parameters such as column height, rise velocity or the amplitude of the seismic signal. The analysis can consider all recorded events, or different threshold levels can be adopted. Previous studies have used the VEI to define the lower threshold of eruptive events (e.g. De La Cruz-Reyna 1993) or the volume of erupted material (Bebbington & Lai 1996).

In this study, one objective was to examine statistically the explosion process to see if any differences were revealed, which might reflect contrasting styles of activity. A statistical analysis of the repose times between explosions is
performed. In each case the events are of a relatively small magnitude: in the case of Volcán de Colima, Karymsky and Tungurahua they can be considered as discrete Vulcanian or Strombolian explosions from andesitic stratovolcanoes. Mt. Erebus, on the other hand, has an active phonolitic lava lake, with explosions resulting from bubble-burst near to the surface.

Seismicity was used to identify the start and finish of each event. The magnitude of an explosion can also be defined seismically; however, this is only approximate because of many factors that could affect the amplitude of the seismic signal, such as variation in source depth, ash content, or coupling to wall rock (Johnson et al. 2005).

Table 1 summarizes the characteristics of eruptive events at the volcanoes studied, and indicates whether the processing was performed manually or using a picking algorithm. At two of the volcanoes the events have been divided into two types based upon characteristics of the seismicity. These characteristics are described in more detail below.

In the case of Volcán de Colima three periods of activity have been chosen: May 2002, which represents a period during which low rate lava effusion was taking place; June–July 2003, which included a much larger magnitude event on 17 July, which partially destroyed the summit dome; March–September 2004, a period that led up to the onset of a new rapid effusive episode.

For Karymsky we highlight statistical explosion data from representative 2 day and 3 day periods in 1997 and 1998. The primary difference in these datasets is a relative increase in explosion frequency in 1998 and the existence of an active block andesite lava flow in 1997, which was quiet in 1998. At Tungurahua seismicity during the 45 day study period in 2004 consisted primarily of discrete short-duration (less than a few minutes) events. Finally, for Erebus, data from a long period were analysed, extending from 1984 to 2004.

At Colima and Tungurahua the events have been subdivided into two main groups, based upon the seismic waveform: explosive events demonstrate a typical impulsive onset, whereas the degassing events are more emergent in their waveform. At Colima, degassing events typically have longer codas (tails). Figure 2 shows one typical day of activity at Volcán de Colima with both types of event present.

Analysis methods

Time-series analysis

The data were first treated as a simple time series, taking the onset of each explosion as an event in time. Before the distribution of data can be considered, it needs to be established whether the data are stationary in time or not, that is whether the data have a trend or are clustered (see Nason 2006). Cumulative frequency plots can give an initial indication, with further tests possible to verify the observations.

Coefficient of variation

This simple technique can demonstrate the degree of clustering within a time series. The coefficient of variation is defined as:

$$C_v = \frac{\sigma_T}{\tau}$$  \hspace{1cm} (1)

where \(\tau\) is the mean and \(\sigma_T\) is the standard deviation of the time between events. If the controlling

<table>
<thead>
<tr>
<th>Volcano</th>
<th>Range of repose intervals</th>
<th>Types of events</th>
<th>Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volcán de Colima</td>
<td>0–2410 min</td>
<td>Exp, DG</td>
<td>Manual</td>
</tr>
<tr>
<td>Karymsky</td>
<td>3–20 min</td>
<td>Exp</td>
<td>Algorithm</td>
</tr>
<tr>
<td>Tungurahua</td>
<td>0.29–1143 min</td>
<td>Exp, DG</td>
<td>Manual</td>
</tr>
<tr>
<td>Erebus</td>
<td>7 s–2922 days</td>
<td>Exp</td>
<td>Manual</td>
</tr>
</tbody>
</table>

Exp, explosive; DG, degassing.
process is Poissonian then $C_V = 1$; if the controlling process is clustered then $C_V > 1$.

**Autocorrelation**

Autocorrelation is a relatively simple method whereby the data are compared with an identical copy that has been shifted in time by an interval known as the lag. The method can be used to detect non-randomness in data and identify a suitable time-series model. Clusters of events or periodic patterns are often highlighted. Interpretation, however, is a non-trivial matter and thus care must be taken to ensure that observed patterns are not given more significance than they deserve (Chatfield 1996). Correlation coefficients are calculated for different time lags and a plot of lag v. correlation coefficient, known as a correlogram, is used to interpret the results. If the autocorrelation function (ACF) has a cut-off at a certain lag value, a moving average (MA) process is suggested; whereas an autoregressive (AR) process produces an autocorrelation function that slowly decreases. A first-order AR process should have an exponentially decreasing ACF, whereas a higher-order process often produces a mixture of exponentially decreasing and damped sinusoidal processes.

If a series of measurements are $Y_1, Y_2, \ldots, Y_N$ at time $X_1, X_2, \ldots, X_N$, the lag $k$ of the ACF is defined as follows:

$$r_k = \frac{\sum_{i=1}^{N-k} (Y_i - \bar{Y})(Y_{i+k} - \bar{Y})}{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}. \quad (2)$$

If no time correlation exists between the intervals $X_i$ and $X_{i+k}$ the function $r_k$ will equal zero, or will fluctuate around zero. If deviations from the mean value are mainly positive during a period $\Delta$, indicating a cluster of events, then $r_k$ will receive only positive contributions for values of $k$ of the order of $\Delta$ or less. This method should be applied only to stationary datasets. If the data contain a trend, this can be removed prior to applying autocorrelation.

The autocorrelation method was used at Mt. Etna to observe time correlations between seismic stations (Vinciguerra et al. 1998). The method demonstrated that two eruptions showed a different time evolution of the movement of magma bodies. In this study, the time series is defined as the daily frequency of eruptions.

**Partial autocorrelation**

An extension of the autocorrelation method, partial autocorrelation can aid in the interpretation of a correlogram. At lag $k$, the partial autocorrelation is the autocorrelation between $X_t$ and $X_{t-k}$ that is not accounted for by lags 1 to $k - 1$. The effect of shorter lag autocorrelation is removed from the correlation estimate at longer lag times. The partial autocorrelation function (PACF) has the opposite properties to the ACF in that an AR process produces a function with a characteristic cut-off. This represents the order of the AR process.

The following gives an estimate of $\Phi_{ik}$, the partial autocorrelation function, where $r_k$ is the autocorrelation function:

$$\Phi_{ik} = \frac{r_k - \sum_{k=1}^{k-1} \Phi_{k-1,j} \Phi_{j,m-1}}{1 - \sum_{k=1}^{k-1} \Phi_{k-1,j}}. \quad (3)$$

**Fractal dimension**

The method of fractal analysis within the time domain can be used to identify self-similarity. It considers events as points on the time axis, which is then divided into smaller sections or boxes. The fraction of intervals $R$ occupied by events is then calculated before reducing the box dimension. Plotting $R$ v. box size with a log–log scale and calculating the slope then permits determination of the degree of clustering.

The general definition of a fractal set is

$$N_r = \frac{C}{r^D} \quad (4)$$

where $N_r$ is the number of objects with a characteristic linear dimension $r$, $C$ is a constant of proportionality and $D$ is the fractal dimension. The fractal distribution has the special characteristic of being scale invariant.

The box-counting method as used by Latora et al. (1998) defines $N(l)$ as the number of boxes of side length $l$ needed to include the whole object:

$$N(l) \approx \frac{1}{l^D}, \quad \text{for } l \to 0. \quad (5)$$

The total length is given by

$$L(l) = lN(l). \quad (6)$$

Hence

$$L(l) \approx l^{1-D}. \quad (7)$$

Extending the theory to probability distributions, the probability that an interval contains an event is the ratio of intervals with events divided by the total number of intervals of length $l$. Therefore

$$p(l) = \frac{N(l)}{N_{tot}(l)} = \frac{L(l)}{L_{tot}}. \quad (8)$$
By substituting equation (7) into equation (8) the relationship between the probability of occupancy and the fractal dimension is obtained:

$$P(l) = \frac{L(l)}{L_{tot}} \approx l^{1-D}.$$  \hfill (9)

Therefore,

$$D = 1 - \frac{\ln P(l)}{\ln l}.$$  \hfill (10)

Plotting $P(l)$ v. $l$ with a log–log scale will produce a straight line with gradient $1-D$ if the distribution of the data is fractal and self-similarity is observed. The fractal dimension will often be between zero and one, which indicates that events are not randomly distributed over time, but are more or less clustered. The value of $D$ defines the level of clustering of the data. The smaller its value, the more isolated is the cluster. As with any statistical analysis, this method is very sensitive to the size of the sample.

**Distribution of data**

To illustrate the distribution of a dataset, it can be plotted as a histogram, and because the data are continuous variables, the probability density function of a given distribution can be plotted for comparison. Different methods exist for estimating the parameters for a given distribution. In this study the maximum likelihood method was used. Tests such as Anderson–Darling, Kolmogorov–Smirnov or $\chi^2$ can then be used to test the fit of the distribution. Another useful plot in survival analysis is the survival function, which gives the probability of a repose time with duration greater than $\tau$ and is defined in general as

$$S(\tau) = P(T > \tau)$$  \hfill (11)

where $T$ is a positive continuous variable and $P(T)$ represents the probability. The area under the plot represents the expected lifetime of the event or process. A straight regression line through the step survival function suggests a random Poisson process.

In the general formulae of several distributions given below, the **location** parameter has not been included. This parameter is merely a translation of the distribution along the $x$-axis.

**Poisson distribution**

If the distribution of $N$ events in time is random then the number of events in an interval of size $\tau$ follows a Poisson distribution, with the probability of $k$ events in the interval given by

$$P(k) = \exp(-\lambda \tau) \frac{(\lambda \tau)^k}{k!}; \quad k = 0, 1, 2, \ldots$$  \hfill (12)

$\lambda$ represents the intensity parameter related to the number of events per unit time.

**Exponential distribution**

The exponential distribution is related to a Poisson process, as the time between events will have an exponential distribution. The probability density function is given by

$$P(\tau) = \lambda e^{-\lambda \tau}; \quad \tau > 0.$$  \hfill (13)

This function has just the one scale parameter, $\lambda$. The survival function is given by

$$S(\tau) = e^{-\lambda \tau}$$  \hfill (14)

which is the probability that the next event will take place after a period of time $\tau$.

**Weibull distribution**

The Weibull probability represents a generalization of the exponential distribution. Its density function is given by

$$P(\tau) = \lambda \gamma (\lambda \tau)^{\gamma-1} \exp(-\lambda \tau)^\gamma$$  \hfill (15)

where $\tau$ is the repose interval, $\gamma$ represents a shape parameter and $\lambda$ is the scale parameter (Lee & Wang 2003). The survival function is given by

$$S(\tau) = \exp(-\lambda \tau)^\gamma.$$  \hfill (16)

The Weibull model was applied to repose interval data for volcanic eruptions in New Zealand (Bebbington & Lai 1996). A graphical approach was used to determine the variables, by plotting the function with the appropriate axis then using regression to calculate the variables.

For the Weibull distribution the following can be derived from the cumulative frequency function $F(\tau)$:

$$\ln \left[ \ln \left( \frac{1}{R(\tau)} \right) \right] = \lambda \ln \tau + \lambda \ln \left( \frac{1}{\gamma} \right)$$  \hfill (17)

where $R(\tau) = 1 - F(\tau)$. Therefore, plotting $\ln[\ln(1/R(\tau))]$ against $\ln \tau$ allows $\lambda$ to be calculated from the slope and $\gamma$ from the intercept.

For Taveuni, Fiji, the Weibull distribution fitted the data for 101 events (Cronin et al. 2001). It was noted that a value of $\gamma < 1$ indicates a decrease in the survival rate from the time of the previous
eruption onset. This results in a clustering of events. If \( \gamma > 1 \), the survival function increases after the eruptions, giving a more regular spacing of events. A value of \( \gamma = 1 \) reduced the distribution to a homogeneous Poisson process. The scale factor \( \lambda \) reflects underlying activity being equal to the mean repose interval for the Poisson distribution.

**Log-logistic distribution**

The following define the probability density and survivor functions, respectively:

\[
P(\tau) = \frac{\alpha \gamma \tau^{\gamma-1}}{(1 + \alpha \tau)^{\gamma+1}} \quad (18)
\]

\[
S(\tau) = \frac{1}{1 + \alpha \tau^\gamma}. \quad (19)
\]

The two-parameter log-logistic probability density function includes parameters for scale (\( \alpha \)) and shape (\( \gamma \)). To date, the applicability of the log-logistic distribution has been limited in Earth Science, although its use is increasing; for example, it was recently found that the heat fluxes for different regions of the Earth’s surface fitted a log-logistic distribution (Shapiro & Ritzwoller 2004). The use of this distribution in volcanology is limited to the study of the repose intervals between Vulcanian explosions of Soufrière Hills volcano, Montserrat (Connor et al. 2003).

**Gamma distribution**

The general formula for the probability density function of the gamma distribution is

\[
p(\tau) = \frac{\lambda}{\Gamma(\gamma)} (\lambda \tau)^{\gamma-1} e^{-\lambda \tau}; \tau > 0, \gamma > 0, \lambda > 0 \quad (20)
\]

where \( \gamma \) is the shape parameter, \( \lambda \) is the scale parameter and \( \Gamma \) is the gamma function, which has the formula

\[
\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt \quad (21)
\]

where \( a \) is a positive integer. The survival function has the formula

\[
S(\tau) = \frac{\lambda}{\Gamma(\gamma)} (\lambda \tau)^{\gamma-1} e^{-\lambda \tau} \quad (22)
\]

The gamma distribution reduces to an exponential distribution when \( \gamma = 1 \). Like the Weibull distribution, it is one of the so-called lifetime distributions, which are often used to describe failure within a system or of a product. This function is particularly flexible and arises naturally in processes for which the waiting times between Poisson distributed events are relevant. It occurs when a system is failing with exponentially distributed backups, as part of a standby model (NIST 2005). Other uses are in Bayesian reliability analysis. With a standby model the system includes an identical component that remains in an ‘off’ state until the primary component fails; then it switched on and the system continues to operate.

The gamma distribution has been used as an extension of the Gutenberg–Richter Law to describe the magnitude–frequency relationship of various earthquake catalogues, with the suggestion that they were produced by a uniform mechanism extending across the magnitude range, but culminating in a taper at the largest magnitudes (Main 1996; Kagan 1997). In volcanology the gamma distribution has been used in a Bayesian model of eruptive events (Ho 1990; Solow 2001) and has been applied to the spatial distribution of volcanoes within active plate margins (de Bremond d’Ars et al. 1995).

**Results**

Each dataset was analysed using the methods discussed above. Here some of the results are presented.

**Time-series analysis**

Figures 3 and 4 show the frequency (i.e. the number of events per unit time) and cumulative frequency of explosive events for two of the datasets. A clear increase in the number of explosive events per day can be observed during the month of May 2002 at Volcán de Colima. This is also highlighted by the gradient change in the cumulative number of events plot. This indicates that the data are non-stationary during this period, with a possible transition between regimes occurring after 7 May. During the second part of the month, the frequency of explosions rapidly increases to a peak then steadily decreases. The frequency of degassing events remains approximately constant. A clear increase can be seen in the ratio of explosions to degassing events.

The coefficient of variation, defined by equation (1), is given in Table 2. In most cases, lower coefficients of variation are obtained if the events are divided into the two families.

The frequency of events and cumulative frequency of the Erebus data are plotted in Fig. 4. It can be clearly seen that the data for the full period from September 1984 to July 2002 are clustered,
Fig. 3. Volcán de Colima: May 2002, general plots of events per day for each dataset (Exp, explosions; DG, degassing events): (a) frequency of events in events per day; (b) cumulative frequency; (c) ratio between explosive and degassing events; (d) and (e) fraction of events with repose interval greater than $t$. 

N. VARLEY ET AL.
with many periods of quiescence separated by enhanced activity. The period between November 1999 and February 2001 was selected for subsequent analysis, as it offers a large period of relatively continuous activity. The frequency plots for this period show increasing eruptive activity towards the middle of the period, followed by a decline during the second half. This type of waxing and waning activity is typical of eruptive episodes at many volcanoes.
Table 2. The mean, standard deviation (SD) and coefficient of variation (CV) of data acquired from Volcán de Colima, Karymsky, Tungurahua and Erebus

<table>
<thead>
<tr>
<th>Period</th>
<th>Type</th>
<th>Events</th>
<th>Mean (minutes)</th>
<th>SD</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volcán de Colima</td>
<td>All</td>
<td>5754</td>
<td>6.767</td>
<td>8.729</td>
<td>1.29</td>
</tr>
<tr>
<td>May 2002</td>
<td>Exp</td>
<td>4728</td>
<td>8.612</td>
<td>12.75</td>
<td>1.48</td>
</tr>
<tr>
<td></td>
<td>DG</td>
<td>1007</td>
<td>42.2</td>
<td>53.59</td>
<td>1.27</td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>826</td>
<td>106.2</td>
<td>156.1</td>
<td>1.47</td>
</tr>
<tr>
<td>June–July 2003</td>
<td>Exp</td>
<td>519</td>
<td>170.1</td>
<td>219.4</td>
<td>1.29</td>
</tr>
<tr>
<td></td>
<td>DG</td>
<td>306</td>
<td>283.5</td>
<td>385.6</td>
<td>1.36</td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>1629</td>
<td>187.1</td>
<td>228.3</td>
<td>1.22</td>
</tr>
<tr>
<td>Mar–Sep 2004</td>
<td>Exp</td>
<td>862</td>
<td>353</td>
<td>370.7</td>
<td>1.05</td>
</tr>
<tr>
<td></td>
<td>DG</td>
<td>766</td>
<td>399</td>
<td>562.6</td>
<td>1.41</td>
</tr>
<tr>
<td>Karymsky</td>
<td>1997</td>
<td>–</td>
<td>3.87</td>
<td>6.96</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>1998</td>
<td>–</td>
<td>1064</td>
<td>3.78</td>
<td>2.003</td>
</tr>
<tr>
<td>Tungurahua</td>
<td>July–Aug. 2004</td>
<td>All</td>
<td>1675</td>
<td>17.27</td>
<td>48.87</td>
</tr>
<tr>
<td></td>
<td>Exp</td>
<td>1388</td>
<td>45.85</td>
<td>133.0</td>
<td>2.90</td>
</tr>
<tr>
<td></td>
<td>DG</td>
<td>709</td>
<td>87.88</td>
<td>304.1</td>
<td>3.46</td>
</tr>
<tr>
<td></td>
<td>Nov. 1999–Mar. 2001</td>
<td>–</td>
<td>1262</td>
<td>560</td>
<td>761.6</td>
</tr>
</tbody>
</table>

**Autocorrelation and fractal dimension analysis**

The autocorrelation correlogram is given in Fig. 5 for each period of analysis for Volcán de Colima, with the data divided into the two event types. The curved line marks approximate 95% confidence levels for the significance of each correlation. This is equivalent to $\pm 2/\sqrt{N}$, where $N$ is the number of days in each dataset. Some striking differences can be observed between the plots. For the May 2002 data the degassing events show much less correlation than the explosion events. This suggests that there is more clustering of the latter whereas the degassing events are more randomly distributed in time. The explosion data show evidence of some periodicity in the ACF, suggesting that an AR model is appropriate. The PACF in Fig. 5b indicates that the order of the model is one.

Of the three periods, June–July 2003 clearly shows the lowest correlations and least clustering. For the March–September 2004 period, the degassing events show short-term clustering, and the ACF possibly suggests a second-order MA model. The explosive events appear to fit a mixed AR–MA model.

A similar analysis was carried out for the seismic amplitudes of the events for the period June–July 2003. In this study, maximum seismic amplitude is used as a proxy for eruption magnitude; however, it cannot be assumed that the relationship is linear. The summing of amplitudes for one day gives an approximation of total eruption energy release for that day, and can be justified for the purposes of trend and cluster analysis. Figure 6 shows the variation in amplitude with time over the 2 month period. Figure 6a shows the average daily amplitude, which shows signs of a downward trend. Plotting the total amplitude for each day shows the trend more clearly in Fig. 6b. Finally, Fig. 6c shows the data adjusted to remove the trend by calculating a simple difference between consecutive values. With the trend removed, the autocorrelation analysis was performed, with the correlogram shown in Fig. 7. The difference can be seen between the correlograms before and after the trend was removed. The trend is clearly visible within the first plot; the ACF does not decay to zero, which suggests that the series is not stationary. The second plot shows that there may be some clustering of short duration, but the autocorrelation is negative. Here the differencing and subsequent ARIMA (0,1,1) model seem appropriate.

To investigate the applicability of fractal analysis of the data, the frequency–magnitude relationship was investigated. Plotting the frequency of events with a magnitude greater than $m$ on a log–log scale should produce a straight line if a power law is obeyed, as with the Gutenberg–Richer Law in seismology. Because the relationship between magnitude and maximum seismic amplitude cannot be assumed to be linear, caution must be exercised in reaching any conclusions.
from this part of the study. The frequency–magnitude relationship for the June–July 2003 Volcán de Colima data is shown in Fig. 8. Removing events with amplitude of ≤ 3 units produced the best fit to a power law. Including all the data the regression gave a correlation coefficient of 0.91, but removing the smaller events improved the correlation to give a coefficient of 0.99. This demonstrates that the data can be modelled using a power law. Events with this range of amplitudes were then analysed to obtain the fractal dimension. The smallest box length depended upon the resolution of the data, and in this case it was chosen to be twice the length of the smallest repose.
interval. The length was then increased by a factor of $2^n$ until the total time period was covered. The value for $D$ was calculated at 0.082 from the gradient of the plot shown in Fig. 9. This value shows that some clustering may be present but only on a small scale.

In the case of Tungurahua, the clustering that is clear in the frequency plots is confirmed by the correlogram in Fig. 10. The initial two autocorrelation values lie outside the confidence intervals, indicating a short-term correlation between values.

Fig. 6. Variation of magnitude of explosive events during June–July 2003 at Volcán de Colima: (a) average amplitude; (b) total daily amplitude; (c) adjusted daily amplitude (trend removed).

Fig. 7. Correlograms of amplitude of explosive events during June–July 2003 at Volcán de Colima: (a) daily cumulative amplitude; (b) amplitude adjusted for trend.

Fig. 8. Frequency–magnitude plot (log–log) of explosion data during June–July 2003 at Volcán de Colima: (a) full dataset; (b) magnitude >3.
Subsequent values then decay and show sinusoidal variation, indicating an AR model. The PACF confirms the model to be AR(2).

**Distribution of repose intervals**

Table 3 gives the results of the distribution fitting. In each case the exponential distribution was rejected by tests that attempted to fit the distribution to a Poisson model.

For the Volcán de Colima May 2002 data none of the distributions tested fitted the dataset if all the events were included. The data were then divided into two families of events: explosions and degassing. The Kruskal–Wallis test was applied to the two datasets of repose intervals, and showed them to be inhomogeneous. The closest fit for explosive events was obtained with the log-logistic distribution, although it failed both the Anderson–Darling and Kolmogorov–Smirnov tests. For degassing events, a good fit was obtained with the Weibull distribution.

The data were then divided into two time intervals corresponding to a dramatic increase in the frequency of explosive events per day after 7 May (Fig. 3). The degassing events did not show the same increase. It can be seen that the ratio between the two types of events increased significantly from this date onwards. For the period 1–7 May, very good fits were obtained for the log-logistic distribution for both event types, with the same scale and shape parameters. For the period 8–31 May the distribution changed for both event types to the Weibull distribution, again producing a good fit, with similar shape parameters for both event types. The probability density function of the best-fitting distribution is shown with the data in Fig. 11. In Fig. 12 the survivor functions are plotted against the data.

For Tungurahua, the only distribution that fitted the data and passed the goodness-of-fit tests was the log-logistic distribution, whether the data were taken as a whole or divided into the two classes of events. Figure 13 shows the fit.

In the case of Erebus, the frequency of events plot (Fig. 4) clearly shows that the events are clustered around certain periods. To investigate the distribution one period was selected: November 1999–March 2001. During this period the activity was more continuous. The best-fitting log-logistic distribution is shown in Fig. 14.

**Concluding remarks**

In the case of two of the volcanoes (Volcán de Colima and Tungurahua), the events clearly could be classified into at least two groups. The seismic event associated with a volcanic explosion has been modelled as the result of a single force (Nishimura & Hamaguchi 1993). However, the variety of seismic waveforms that can be observed is indicative of the complexity of the eruptive processes involved. An active volcanic system represents complex interactions dependent upon several variables associated with the physical and geochemical characteristics of the magma, pressure transients and the inhomogeneities of the volcanic edifice itself. Studies of the hypocentres of volcanic seismic events have shown that they can be produced at a variety of depths within or beneath the edifice. Differences in the seismic waveforms
produced by groups of events indicate a difference in the process of energy release and migration over time. The difference could be a function of the location of the source of the event or variation in time of one or more of the controlling parameters.

In this study, it has been shown that the two subsets of events (termed explosive and degassing) are statistically independent. Their distribution is clearly different within the datasets and is seen to vary with time. During 1–7 May 2002 at Volcán de Colima the repose interval between explosive events produced a Weibull distribution, whereas for the degassing events it was log-logistic. During the latter part of the month, however, both types gave a Weibull distribution. The division between the two periods was marked by a significant increase in the frequency of both event types. This possibly indicates a batch of more gas-rich magma arriving near the surface and hence an increased rate of degassing. Alternatively, the increase could have been a reflection of a deeper process, such as an increased pressure differential between the magma chamber and the upper level of magma within the system, or the ascent of this magma towards the surface. By considering further parameters, such as the gas composition, or a detailed seismic analysis, a model could be produced to explain this behaviour. The change in distribution of the repose intervals between degassing events before and after 7 May suggests some change within the system affecting the production of this event type.

The presence of a larger magnitude event on 17 July 2003 had a noticeable effect on the Volcán de Colima data. Interestingly, the frequency of explosive events remained constant; however, the degassing events increased in frequency before the larger event and maintained a higher level afterwards. Again, a significant tendency has been highlighted that can be introduced into a model of the degassing processes and the causation of the larger magnitude event.

Autocorrelation analysis proved an effective method for identifying patterns within the data. Combined with partial autocorrelation, it can show if an AR or MA model is appropriate for the data. Further analysis could determine the parameters for these models. Defining a statistical model for different periods of activity can assist in

### Table 3. Distributions fitted to event data

<table>
<thead>
<tr>
<th>Period</th>
<th>Event type</th>
<th>Distribution</th>
<th>Parameters</th>
<th>Goodness-of-fit test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Location</td>
<td>Scale</td>
</tr>
<tr>
<td>Volcán de Colima</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>May 2002</td>
<td>Exp</td>
<td>L</td>
<td>0.25</td>
<td>5.138</td>
</tr>
<tr>
<td></td>
<td>DG</td>
<td>W</td>
<td>0.5</td>
<td>38.93</td>
</tr>
<tr>
<td>1–7 May 2002</td>
<td>Exp</td>
<td>W</td>
<td>0.25</td>
<td>21.02</td>
</tr>
<tr>
<td></td>
<td>DG</td>
<td>L</td>
<td>0.25</td>
<td>16.00</td>
</tr>
<tr>
<td>8–31 May 2002</td>
<td>Exp</td>
<td>W</td>
<td>0.249</td>
<td>09.81</td>
</tr>
<tr>
<td></td>
<td>DG</td>
<td>W</td>
<td>0.5</td>
<td>43.17</td>
</tr>
<tr>
<td>June–July 2003</td>
<td>All</td>
<td>G</td>
<td>0.25</td>
<td>189.5</td>
</tr>
<tr>
<td></td>
<td>Exp</td>
<td>G</td>
<td>0</td>
<td>250.7</td>
</tr>
<tr>
<td></td>
<td>DG</td>
<td>G</td>
<td>0.75</td>
<td>605.8</td>
</tr>
<tr>
<td>Mar.–Sep. 2004</td>
<td>All</td>
<td>G</td>
<td>0.25</td>
<td>304.7</td>
</tr>
<tr>
<td></td>
<td>Exp</td>
<td>G</td>
<td>0</td>
<td>417.3</td>
</tr>
<tr>
<td></td>
<td>DG</td>
<td>G</td>
<td>0.25</td>
<td>792</td>
</tr>
<tr>
<td></td>
<td>Av./day</td>
<td>L</td>
<td>0</td>
<td>204.8</td>
</tr>
<tr>
<td>Erebus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nov. 1999–Mar. 2001</td>
<td>–</td>
<td>L</td>
<td>2.983</td>
<td>351.9</td>
</tr>
<tr>
<td>Karymsky</td>
<td>1997</td>
<td>Exp</td>
<td>W</td>
<td>1.009</td>
</tr>
<tr>
<td></td>
<td>1998</td>
<td>Exp</td>
<td>W</td>
<td>1.0085</td>
</tr>
<tr>
<td>Tungurahua</td>
<td>2004</td>
<td>All</td>
<td>L</td>
<td>0.288</td>
</tr>
<tr>
<td></td>
<td>Exp</td>
<td>L</td>
<td>0</td>
<td>12.74</td>
</tr>
<tr>
<td></td>
<td>DG</td>
<td>L</td>
<td>0</td>
<td>21.29</td>
</tr>
</tbody>
</table>

The distributions considered were log-logistic (L), Weibull (W) and gamma (G); the event types considered were explosions (Exp), degassing (DG) or both (All). The parameters of the distributions have been calculated using the maximum likelihood method. Two goodness-of-fit tests were performed on the data from each event. The level of significance (alpha level) that the model is not rejected is given in the table as test statistic(significance level), respectively. R signifies rejection.
the conception of a physical model of the volcanic processes. Fractal dimension analysis produced values of $D < 1$ for the three datasets examined. $D$ was calculated at 0.082 for the June–July 2003 Colima dataset and 0.149 for the March–September period for explosive events. This suggests that the earlier data were less clustered than the later. An even higher value was obtained for the Erebus data at 0.776. In each case there was a strong power-law relationship between event magnitude and frequency.

Vulcanian explosions have been modelled based on the plugging of the conduit with a rock cap (Woods 1995). During the interval between explosions, a foam-like two-phase layer may form beneath the cap, composed of exsolved volatiles with possibly a groundwater component. The velocity and duration of the eruption is dependent upon gas mass fraction, temperature, diameter of solid fragments and the pressure beneath the cap (Woods 1995). This type of model indicates the parameters that might show variation and thus produce a change in the frequency of events.

Explosive volcanism can be treated as a survival process: its analysis implies the statistical modelling of time-to-outcome data. This type of analysis is most common in medicine, social science or engineering. The explosion is the result of the failure of a cap after the build-up of volatile pressure beneath. The repose interval represents the time during which the pressure is accumulating prior to the failure of the system. It can be regarded as the survival or failure time. Distributions applied in survival analysis include the Weibull,
The fitting of distributions to the data has shown that none of the datasets are the result of a random or simple renewal process. Instead, fits were obtained by the log-logistic, Weibull and gamma distributions, which implies that the system can indeed be treated as a survival process, in which the system goes through cycles of ‘survival’ and subsequent failure.

The theory of material failure has previously been applied to the seismic activity of Volcán de Colima (De la Cruz-Reyna & Reyes-Dávila 2001). Data taken from the accumulated seismic energy were applied to a viscoelastic model, which was used to predict the point of failure of the system. Large accelerations of energy release were treated as precursors to a larger explosive and the 1998–1999 effusive events.

In this study, the Weibull distribution offered the best fit for Volcán de Colima during 8–31 May 2002 for both types of events. If the whole month was considered as one dataset, the degassing events also produced a distribution that could be approximated as Weibull. In each case the shape parameter was less than one. This signifies that there is a decrease in the survival rate from the time of the previous eruption onset. For the other two periods studied, the Gamma distribution
produced the best fit for the complete datasets and after dividing the data into the two event types. For the March–September extended period of data, a daily mean for the repose period was calculated as a way of smoothing. Interestingly, this then produced a log-logistic distribution with a shape parameter of 3.79.

Although the two Karymsky datasets showed variation in the explosion frequency, both can be modelled with a Weibull distribution with a shape parameter greater than one. This indicates that the survival rate is increasing. The Tungaruhua data fitted best a log-logistic distribution, whether for the combined event types or for explosive and degassing event types separately. The shape parameter varied slightly between the two event types at 1.27 for explosive events and 1.16 for degassing.

Fitting the log-logistic distribution implies that the model can be represented by competing processes. Connor et al. (2003) first applied this model to volcanic conduit processes. Vulcanian explosions are controlled by pressure build-up and then release once a mechanical threshold has been exceeded. The degassing of the magma is controlled by its viscosity, volatile contents and the permeability of both the magma and wall-rock. Pressure is increased by increased exsolution or an increase in the viscosity as a result of crystallization. In contrast, the opening of fractures, or an increase in the permeability of the magma, acts to reduce pressure build-up. In the study of Soufrière Hills volcano, Montserrat, the shape parameter was determined to be four for the log-logistic distribution (Connor et al. 2003). The two values obtained for the Karymsky 1998 data and the smoothed March–September Colima data are close to this value, indicating some similarity in behaviour. In contrast, the distribution of the
Tungurahua data had much lower shape parameters, which is a reflection of a heavier tail to the distribution; that is, a greater number of events separated by a longer repose interval.

A key factor in the explosions at Volcán de Colima is the ash content. It has varied greatly during the explosive activity of 2003–2005 and may reflect the depth of the source. Evidence from measurements of SO₂ flux indicate that after the 2001–2003 effusive eruption the magma column receded to a superficial reservoir, where it remained until a further injection into the deep magma chamber and resulting over-pressurization once again forced the magma to fill the fracture systems and rise towards the crater. During the 2003–September 2004 period the magma was degassing in the upper reservoir as a result of convective overturn. Ash samples have been found to contain juvenile particles, indicating that the source of ash-rich explosion must be close to the surface of the magma reservoir. Ash-poor eruptions result from accumulation of gas higher in the upper edifice. Both classes of event, explosive and degassing, have been observed to produce ash. It would therefore appear that the distinguishing factor is not depth, but some complex function of energy release and transfer. The statistical analyses will be combined with other datasets to produce a model to explain these processes.

Different eruptive regimes can be identified within the data, based upon the frequency of the events. A similar approach was taken by De la Cruz-Reyna (1993) when considering historical data with VEI > 2. He identified four regimes during the previous 450 years, with Volcán de Colima switching between states of high and low magma production. As shown, different regimes can also be identified within datasets of the smaller magnitude events studied here.

At Erebus the source process is very different from a brittle failure type mechanism that might be associated with earthquake activity. Rowe et al. (2000) analysed the magnitude of explosions during the period of October 1997 and July 1998. A log cumulative frequency distribution curve produced two distinct gradients and three proposals were presented: the existence of two separate populations of events, a favoured median bubble size, or power-law behaviour with underestimates of the smaller magnitude events. Within the selected period of this study, no distribution fitted the repose intervals well; however, the best fit was obtained again with the log-logistic distribution. The shape parameter was low, at 1.61, suggesting a heavy tail. However, because of the non-stationary character of the data, it is probably inappropriate to fit a continuous distribution model.

This paper presents a statistical analysis of explosion data for four volcanoes. The results show contrasts between the systems, which can be interpreted and included in models reinforced by other data. Using techniques of time-series analysis, stationarity and clustering within the data were investigated. The autocorrelation, partial autocorrelation and fractal dimension methods can be used to demonstrate the level of clustering. For the data examined, a power-law relationship was observed between explosive event magnitude and frequency, analogous to the Gutenberg–Richter relationship in tectonic seismology, although over a smaller magnitude range.

It is clear that the process that creates the explosivity is not a random Poisson process, but is better modelled as a process of survival. The nine datasets investigated all demonstrated that one of the distributions used in survival analysis provided the best model. Following the analysis by Soufrière Hills volcano, Montserrat (Connor et al. 2003), the log-logistic function appears to provide a useful probabilistic model for the repose intervals between...
explosive events. Further analysis is necessary, combined with data from other sources, to better understand the significance of the shape parameter. Because of competition between processes within the upper volcanic edifice, this distribution is highly appropriate for modelling data of this type. In cases where the Weibull or gamma distribution offers the best model, it is likely that one group of processes is dominating the system and hiding the effect of competition. The gamma distribution may suggest that the eruptions can be modelled as a backup process, where the degassing produces a pressure build-up in more than one fracture system. After one system fails, producing an explosive event from that fracture system, a second system becomes active, before it too fails. Observation has shown that there are several sources for the explosive events distributed around the summit crater.

Significant changes related to the frequency of events were reflected by changes in the distribution of repose intervals. Two event types were identified at two of the volcanoes (Colima and Tungurahua) and variation of the response interval was seen to be independent, implying that a different process is producing each event type. Further work is necessary to understand the significance of the shape parameters of the distributions, but it is clear that this form of probabilistic modelling provides a useful tool for the forecasting of future activity.

Further reading

A good general text on time-series analysis is that by Chatfield (1996), and on fractal analysis, with many useful examples in Earth Science, that by Turcotte (1997). General information on explosive eruption models has been given by Freundt & Rosi (1998).

The authors wish to thank the many students who have assisted in the analysis of the seismic data: A. Carter, K. Diller, M. Hoffman, S. Jenkins, R. Lowe, R. Nicholson, S. Pearson, M. Rogers, G. Royer, J. Sussams, M. Zack and B. Z. Espinoza. Extra appreciation is due to C. McClaren and L. Edwards. Many thanks go to K. Jones, R. Aster and P. Kyle of the Department of Earth and Environmental Science and Bureau of Geology and Mineral Resources, New Mexico Institute of Mining and Technology, for providing the Erebus data. N. V. received financial support from UCol FRABA 288/04 and CONACYT 38228-T. M. R. received financial support through NSF EAR No. 0440225, and J. J. through NSF EAR No. 147292. Thanks go to C. M. Hernandez-Suárez for spotting some errors, the editors for their assistance, and the anonymous reviewer whose efforts greatly improved this paper.

References


GORDIEV, E. I., KASAHARA, M., LEVINA, V. I., MIYAMACHI, H. & CHEBROV, V. N. 1997. Magma activity at Karymsky volcano and Academy Nauk Caldera (Kamchatka, Russia) triggers large tectonic (M7.0) event. EOS Transactions, American Geophysical Union, 78(46), F442.


