Risk analysis: examples and discussion (Plenary Talk)

David R. Brillinger
Statistics Department, University of California, Berkeley, CA

ABSTRACT: Risk analysis, that is the problem of estimating the probabilities of rare and damaging events, is important in many fields. One can mention the risks from: floods, earthquakes, forest fires, and space debris. Computing probabilities is basic to risk analysis. The estimated probabilities may be fed into the computaion of insurance premiums for example. In the article the examples of seismic risk analysis and forest fire chances are considered in some detail.

Keywords: catastrophe modelling, earthquake risk, insurance, Mercalli Intensity, mutual information, natural hazards, risk assessment, risk theory, wildfire risk.

1 INTRODUCTION

The fields of civil engineering and statistics have much in common, including the origin of their names

civil = *civilis* = *relating to state* and *statistical* = *status* = *state*

Both fields have much to offer society.

Risk may be defined as the probability of some hazardous event or catastrophe, the chance something bad will happen. In many cases huge amounts of money are involved (National Academy of Science 1998). The principal concern is low probability-high consequence events, events that lead to damage, loss, injury, death, environmental impairment for example. Often the work is done as an aid to decision making. In consequence risk models and risk management pervade modern technical life. The events are in space and time, hence the theory of stochastic processes plays an important role.

The field of risk analysis cuts across the environmental sciences including things like: landslides, avalanches, earthquakes, floods, huricanes, tornadoes, forest fires, space debris, sea storms, hail storms, ...

A common tool in the work of risk analysis is a *catastrophe model*. This may be defined as: a set of databases and computer programs designed to analyze the impact of different scenarios on hazard-prone areas (National Academy of Science 1998). In practice these models combine scientific risk assessments of hazard with historical records to estimate the probabilties of disasters of different magnitudes and the resulting damage to affected structures. The information may be presented in the form of expected annual losses and/or the probability that in a given year the claims will exceed a certain amount.

Risk analyses may be required officially. To cite a specific example: a Core Damage Frequency (CDF) value of 10^{-4} per reactor year is the value endorsed by the Nuclear Regulatory Commission in a Staff Requirements Memorandum as a benchmark objective for accident prevention (Nuclear Regulatory Commission 1997). This rate is the probability of damage to a reactor core within a year.

A formal risk analysis often includes: i) estimation of probabilities, ii) determination of the distribution of damage and iii) preparation of products like formulas, graphics, hazard risk maps. There is extensive use of computing science, substantive subject matter and statistical methods.

The sections of the paper are: Introduction, Civil Engineering and Statistics, Insurance, Two Examples, Risks of Risk Analyses, and Summary and Conclusions.

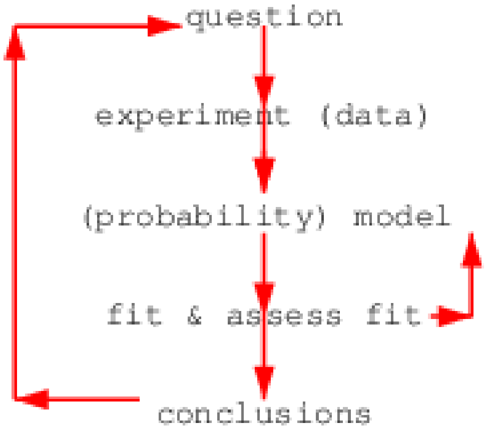


Figure 1: Diagram showing the flow of a statistical investigation.

2 CIVIL ENGINEERING AND STATISTICS

The subjects of Civil Engineering and Statistics have much in common and they have a fair amount of joint history. To focus on early work, Gauss contributed to both surveying and least squares. He made basic contributions to each field during the triangulation of Hannover in the years 1820-1844.

Gauss's work is commemorated on the German Mark note pictured in (Reid 2000). Gauss estimated that he handled more than a million figures in his surveying work.

In any case few of the talks at this meeting do not contain important elements of both civil engineering and statistics.

2.1 Statistics

Statistics may be defined as the science of using data wisely. It involves data collection, data analysis, data reduction, data modelling, data-based inferences. Its *paradigm* is that a datum is a realization of a random variable. Random variables come in many types including images, videos, scatter of points, counts and ... .

Statistics relates directly to the Scientific Method. See the box and arrow diagram of Figure 1.

Persons who collaborate with statisticians expect: techniques (e.g. MOM, OLS, MLE, Bayesian, ...), uncertainties, efficiency, and goodness of fit amongst other things.

Among the methods to be used here are: robust/resistant least squares, the generalized linear model, nonparametric estimation, statistical packages (Chambers and Hastie 1990), and the jackknife.

3 INSURANCE

Insurance measures are often taken to "manage" catastrophic phenomena. Further, consideration of insurance aspects can motivate research work and risk analysis.

Insurance against loss has been around for many years. One can mention the code of Hamurabi (1950 BC) which involved so-called bottomry, a form of marine insurance (National Academy of Science 1998). The catastrophes such of the Great Fire of London 1666 AD, and the U.S. floods of the late 1800's led to the development of insurance as a major business (Clark 1999), (Raynes 1964).

The insurance premium paid is meant to reflect the risk potential. In one method to compute a premium one needs the likelihood and potential damage from possible events. Modern formulae for premiums involve probabilities and distributions. Historically they involved only expected values (Clark 1999).

The *pure risk premium* for damage L is given by

$$P \;=\; E\{L\} \;=\; Prob\{L \neq 0\}E\{L \,|\, L \neq 0\}$$

It is an expected value and seen to involve both a probability and an expected value.. In practice premiums are loaded to cover costs and solvency. Denoting the reserve available at time t by R_t, the insurer wishes

$$Prob\{\sum_j L_{jt} > R_t + \sum_j P_{jt},\ t = 1,2,...\} \le \epsilon \quad (1)$$

for some small $\epsilon > 0$ with j summing over the changing number of risks. Formulae for a *loaded premium* include

$$P \;=\; (1+\alpha)\mu_L,\ \mu + \beta\sigma_L,\ \mu_l + \gamma\sigma_L^2,\ \alpha\mu_L + \beta\sigma_L + \gamma\sigma_L^2$$

for some α, β, γ where

$$\mu_L \;=\; E\{L_t\}, \quad \sigma_L^2 \;=\; var\{L_t\}$$

(Beard, Pentikainen, and Pesonen 1969) or they may be based on the expression (1) and involve the full distributions of the quantities involved. There are various practical details to be dealt with including: taxes, reinsurance, exposure, inflation, investment return, lags, interest rates and there are other approaches. For example an extreme value approach is taken in (Embrechts, Kluppelberg, and Mikosch 2000) and there is a market driven approach (Bohman 1979). This last is adaptive and evolutionary and involves using time series data on income and expenses to compute a premium from predicted future expenses. It is interesting to read in (Clark 1999) of the empirical efforts of companies in the 18th century to find effective premium rates on the basis of their gains and loses.

In summary, in determining insurance premiums probabilities need to be estimated and so too do distributions of losses. Some related results may be found in (Brillinger 1993).

4 TWO EXAMPLES

By way of illustration consider risk analyses of earthquakes and of wildfires. In both thinplate splines are used below to approximate smooth functions of two variables. These splines have the form

$$f(x,y) = \alpha + \beta x + \gamma y + \sum_{k=1}^{K} \delta_k r_k^2 \, log \, r_k \quad (2)$$

with the (x_k, y_k) nodes and

$$r_k = \sqrt{(x-x_k)^2 + (y-y_k)^2}$$

the distance from the k-th node to (x,y), (Powell 1992).

4.1 Seismic Risk Assessment (SRA)

SRA will be defined as the process of estimating the probability that certain performance variates at a site of interest exceed relevant critical levels within a specified time period, as a result of nearby seismic events.

In such a circumstance it is convenient to break down the problem conceptually as in Figure 2. The figure supposes two seismic sources are of concern.

The discussion below works backwards from a structure at a site of concern to the locations, times and sizes of earthquakes.

a) *Damage.* There are a variety of ways to describe and estimate damage. An important method uses the Modified Mercalli Intensity (MMI). One reason for its importance is that values may be derived from historic accounts. Another is that it refers to damage directly.

MMI values are given by roman numerals I to XII (and sometimes 0 referring to no impact.) The scale is ordinal, increasing with increasing severity of damage. For example the definition of MMI $VIII$ includes: "Damage slight in specially designed structures; considerable in ordinary substantial buildings; ... Fall of chimneys; ..." (Bullen and Bolt 1985).

There are values that have been proposed to convert MMI values into damage percentages for different types of structures. The following table is an example of a so-called *damageability matrix*. It was

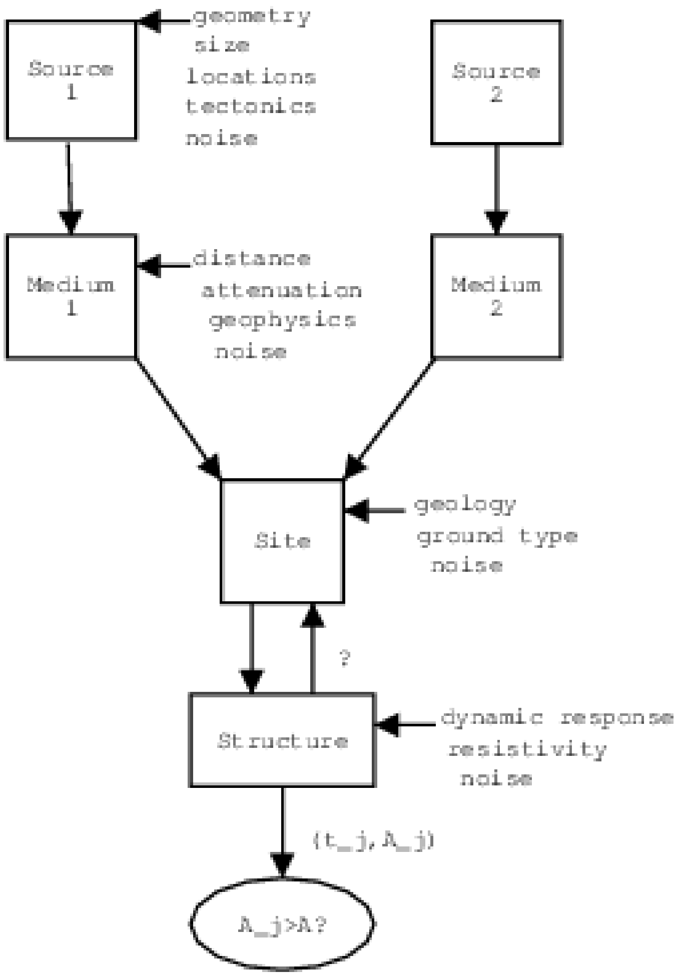


Figure 2: A box and arrow diagram corresponding to the components of an SRA. The t_j are the times of events and the A_j the corresponding performance variable values. A is a threshold level.

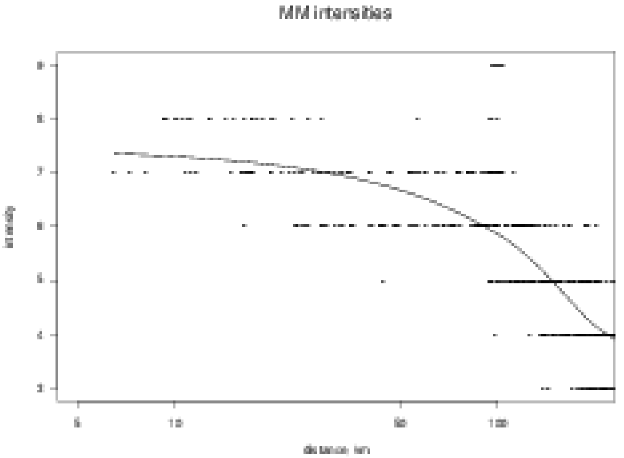


Figure 3: Loma Prieta MMIs versus distance from epicenter.

given in (Munich Re 1991). The entries are loss ratios per risk category in %.

MMI	VI	VII	VIII	IX	X
residential	.3	1.5	5	16	40
commercial	.6	2.5	9	25	60
industrial	.1	.7	3	11	30

b) *Attenuation.* To illustrate the general decay of seismic energy as the signal passes through the earth consider the Loma Prieta event of 17 October 1989. The epicenter of the event was near Santa Cruz, California. The October 1991 issue of the *Bulletin of the Seismological Society of America* was devoted to the Loma Prieta Earthquake and its effects. There were 62 deaths, 1300 buildings destroyed, and $7 billion in damage. The observed MMI values and a hand-sketched isoseismal map may be found in (Stover, Reagor, Baldwin, and Brewer 1990).

Figure 3 plots observed MMI values versus their distance from the epicenter of the event. One sees a general falling off of level as the distance from the epicenter increases. One also sees a cluster of high levels at distances around $100km$ from the epicenter. A smooth robust/resistant line has been added. The computation of this line assumed that the MMI values might be treated as numerical. Support for this assumption is given below.

To process the data one seeks a statistical distribution for ordinal-valued variates. One way to construct such is to postulate the existence of a latent process ζ and ordered cut points a_i such that the MMI value at the location with coordinates (x, y) is given by

$$I_{x,y} = i \; if \; a_i < \zeta_{x,y} \leq a_{i+1}$$

Consider the model

$$\zeta_{x,y} = f_{x,y} + \epsilon_{x,y} \qquad (3)$$

with $f_{x,y}$ deterministic and smooth and with $\epsilon_{x,y}$ having an extreme value distribution, i.e. $Prob\{\epsilon \leq u\} = 1 - exp\{-e^u\}$. The use of the extreme value distribution is plausible given the nature of destruction. It and the corresponding use of the cloglog link mean that the function glm() of Splus ((Chambers and Hastie 1990)) may be employed for the computations, (McCullagh and Nelder 1989), Chapter 5. In the analysis below the function $f_{x,y}$ is expressed in terms of thin plate splines In the computations the data (x_j, y_j) were standardized.

The results are given in Figures 4 and 5. The first figure provides the estimate of $f_{x,y}$ obtained by fitting the model just described. One sees a general dying off of the function values as one moves away from the epicenter, except for a rise near San Francisco. This phenomenon appeared in Figure 3 and has been associated with reclaimed land.

Figure 5 provides the estimated cut points and approximate marginal 95% bounds. One sees that for the MMI values II and above the cut points fall close to a straight line. This finding lends support to the use of MMI values as if they were numerical as was done in preparing Figure 3.

See (Brillinger, Chiann, Irizarry, and Morettin 2001) for further details concerning this type of analysis including the use of shrinkage estimates.

A formal relationship describing the fall-off in energy with distance as it passes through the medium is needed. Following (Joyner and Boore 1981), consider an attenuation form

$$log(-log(1 - Prob\{I = i\})) =$$
$$\alpha_i + \beta d + \gamma log(d) + \delta M \qquad (4)$$

where d is the distance of a point of concern to the epicenter of the event and M the event's magnitude. This was fit to the Loma Prieta data using a robust resistant algorithm and the results are provided in Figure 6 for $i = 0, V, VIII$. (As only one event was involved δ could not be estimated.) In the case of MMI $VIII$ one sees a rapid fall-off with distance.

c) *Event locations and times.* One imagines a marked spatial-temporal point process of earthquake

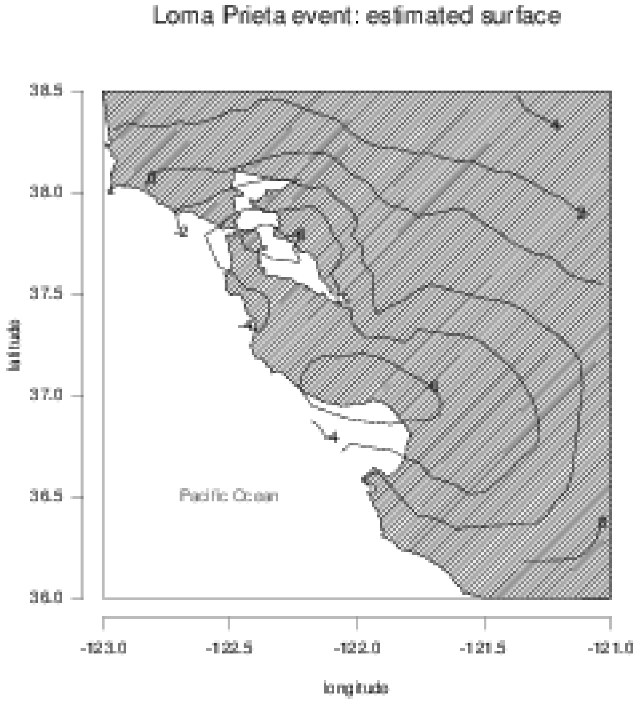


Figure 4: The estimate of $f_{x,y}$ of the model (3). The circle represents the epicenter of the event.

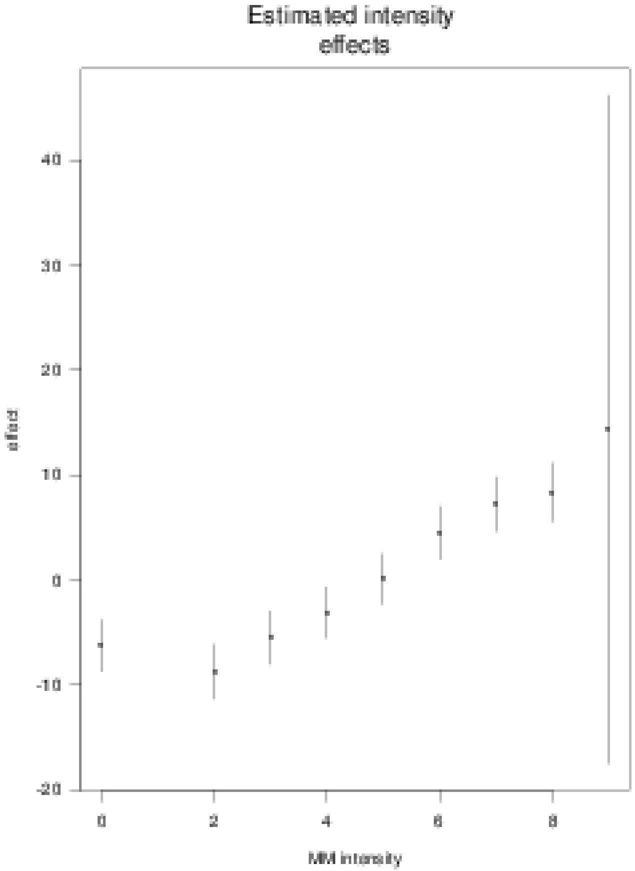


Figure 5: Estimated cut points, $\hat{a}_i$, and 2 s.e.'s. There were no MMI I's in the data set.

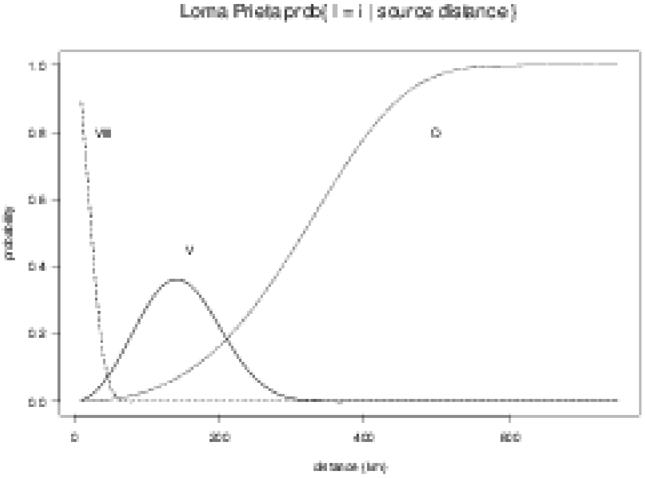


Figure 6: Estimated probabilities of the indicated MMIs as a function of distance from the epicenter.

locations, times and sizes. In California many faults have been located. In Figure 2 just two sources have been hypothesized, but there could be many. In an expression like (4) one might take d to be the distance to the nearest point on the fault from the site. Faults have been modelled as a line segments and plane segments with event magnitude related to their size. There are many geological fault maps to work with.

Commonly renewal processes are employed to model the sequence of times. The intervals between events might be assumed exponential, Weibull or lognormal.

As an example of a fair premium computation, consider a commercial building 25km from an epicenter and an event like Loma Prieta. For this case, using the factors in Table 1, the estimated expected loss is

$$.6 * .103 + 2.5 * .389 + 9 * .475 + 25 * 0 \ = \ 5.31\%$$

The computations are set up using conditional probabilities (Pregibon 1980). Assuming the damage percents are constant the standard error is .71

4.2 Wildfires

The second example of the paper concerns the problem of predicting the occurrence of forest fires as a function of place and time. Let occurrences be denoted by $(x_j, y_j, t_j),\ j = 1, 2, 3, ..., n$ with (x, y) location and t time. One has a point process in space and time.

To illustrate the circumstance consider Figure 7. The bottom panel shows the locations of forest fires in Oregon during the period 1989-1996, specifically those that occurred in Federal lands. These lands are indicated in the top panel of the figure.

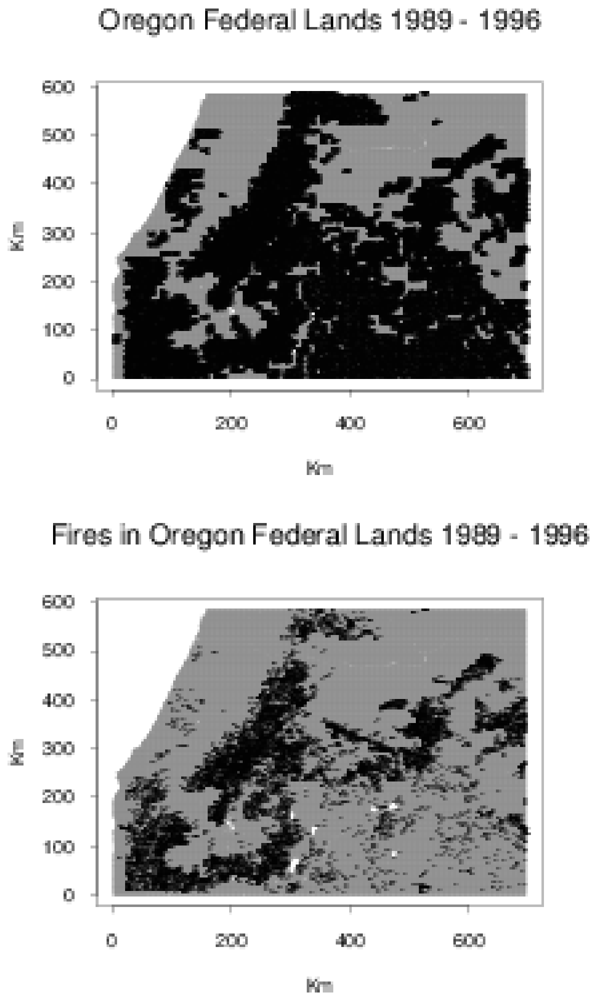
The data set for Oregon was large, 578,192,400 voxels and 15,786 fires. To be able to carry out exploratory data analyses a sample of the data were used. All the voxels with fires were employed, but only a sample of those where no fires occurred. The sampling fraction was $\pi \ = \ .00012$. This lead to 58094 cases.

To formalize the problem consider voxels $(x, x + dx] \times (y, y + dy] \times (t, t + dt]$ and let

$$N_{x,y,t} \ = \ 1 \quad if\ a\ fire\ in\ the\ (x, y, t) - voxel$$

$$= \ 0 \quad otherwise$$

For convenience suppose that the voxel sides $dx,\ dy,\ dt \ = \ 1$. (In the data and computations $dx,\ dy \ = \ 1\ km$ and dt is 1 day.) Let H_t denote the



Oregon Federal Lands 1989 - 1996

Fires in Oregon Federal Lands 1989 - 1996

Figure 7: The top panel shows the Federal lands in Oregon. The bottom provides fire locations.

history of the process up to and including time t, and consider the probability

$$Prob\{N_{x,y,t} = 1 \mid H_t\} = p_{x,y,t}$$

In the work presented next a logit model is assumed, specifically

$$\begin{aligned} logit\ p_{x,y,t} &= log\ p_{x,y,t}/(1 - p_{x,y,t} \\ &= g_1(x,y) + g_2(d) + \zeta \qquad (5) \end{aligned}$$

with (x, y) location, d day of the - year, and ζ a year effect. Logit models have been used previously in estimating fire risk, see for example (Martell, Otukol, and Stocks 1993). In the computations the g functions are assumed to be smooth and are represented by spline functions. The spatial term, g_1, involved a thin plate spline and the day term, g_2, was a periodic spline. The Splus function make.rb() of Funfits, (Funfits 2002), was employed. There were 60 nodes used and these were taken to be on a 10km grid throughout the region.

With the logit link, conditional on the sample, one had a generalized linear model, (McCullagh and Nelder 1989), with an offset of $log\ 1/\pi$. This meant that standard glm computer programs could be used for the analysis. (The new logit was $logit\ p' = logit\ p + log(1/\pi)$.)

The results are provided in Figure 8 giving the estimates of the functions g_1, g_2 and the effects ζ. (The ζ of (5) are assumed fixed here, but in work in progress here they are being assumed random.) Examining the top panel one sees fewer fires in SE Oregon, as could have been anticipated from the bottom panel of Figure 8. In the middle panel of the figure one notes a substantial day of the effect - many more fires in the summer. The bottom panel shows a definite year effect. The year effect values are relative to 1996 as 0 and the horizontal line is at level 0. This analysis provides base values to refer in computations directed at forecasting using predictive explanatories.

Consider the problem of prediction using some of the indices that have been proposed, specifically the problem of predicting whether a fire becomes large once it has started. Mutual information (MI) will be used to find which indices are most highly associated with large fires. For a contunous variable, X, and a discrete variable, Y, MI is defined as:

$$E\{log\frac{p(x,y)}{p_X(x)p_Y(y)}\} \qquad (6)$$

where

$$p(x,y)dx = Prob\{x < X < x + dx\ and\ Y = y\}$$

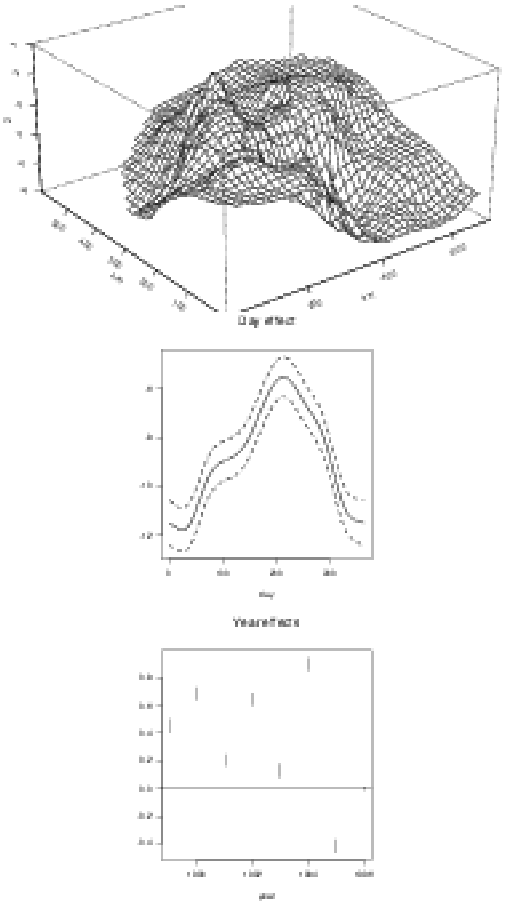


Figure 8: The top panel provides the estimated location effect for g_1 of the model (5). The middle panel provides the estimated day of the year effect. The bottom panel shows the estimated year effect. Approximate 95% error bounds are indicated.

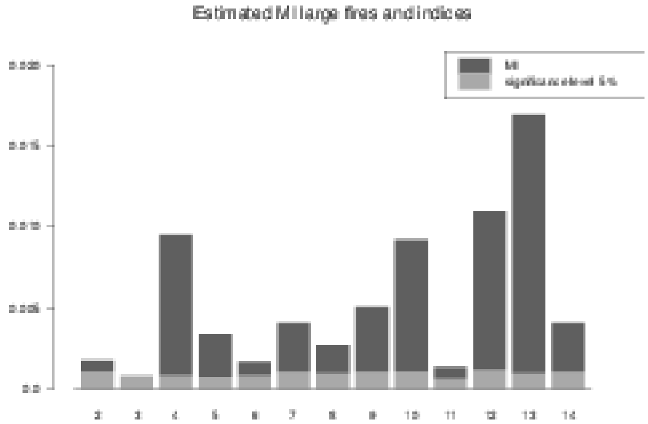


Figure 9: The estimated coefficient of MI between a large (as opposed to small) fire and a selection of predictor variables.

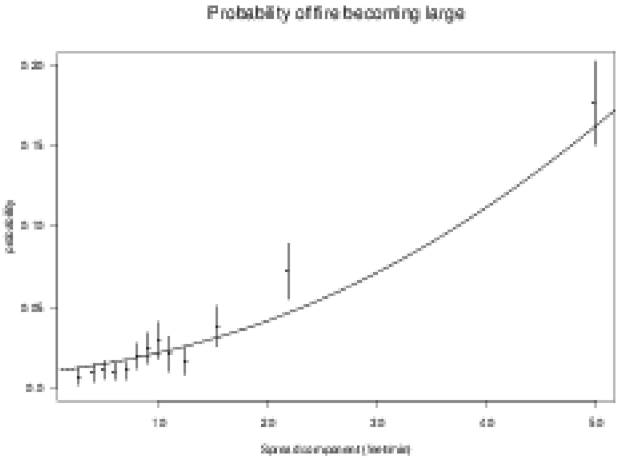


Figure 10: Probability of a fire becoming a large fire as a function of Spread Component. The arrows give 2 s.e. limits.

Mutual information may be estimated using a discrete version of (6). Amongst the properties of MI are that MI = 0 implies that the variates involved are statistically independent.

Figure 9 provides estimates of the MI between the binary variable, $Y = 0$ or 1, referring to whether a fire becomes large. In the figure there are 13 explanatories, numbered 2 through 14. The largest, number 13, is the so-called Spread Component (SC) index. This is a numerical value derived from a mathematical model that integrates the effects of wind and slope with fuel bed particle properties to compute the forward rate of spread at the head of the fire, (Deeming, Burgan, and Cohen 1977). The SC was chosen for the next analysis because it had the highest estimated MI with the variate a fire becoming large.

Consider the model

$$log(-log(prob\{large\ fire|fire\ and\ SC\})) =$$

$$\alpha + \beta SC$$

The use of $log(-log))$ comes from extreme value considerations. Figure 10 gives the results with the dots and vertical lines corresponding to naive estimates of the probability, (i.e. based on the experience in an interval of SC values), and ± 2 standard error limits. One sees the steady increase in probability with the level of SC. In the data set studied the the SC level ranged from 1 to 169 with a median of 8 and quartiles of 5 and 12.

More details of the data set and analyses may be found in (Brillinger, R., Preisler, and Benoit 2003).

5 RISKS OF RISK ANALYSES

At this point in time many risk models have been constructed. These have been based on varying amounts of data and have involved differing methods of validation. Many make use of simulation/Monte Carlo. As in the case of those techniques risk analyses have important limitations. Briefley one can refer to: overdependence on simulation, lack of data, lack of empirical validation, identifiability issue (different inputs can lead to similar outputs), the sample used for fitting may be used to validate the model, assumption of independence, and the limitations of sensitivity analysis.

6 SUMMARY AND CONCLUSIONS

Risk analysis has been considered in general and for two specific examples. The examples were motivated by the problem by indicating the type of information insurers would like.

The demand for risk analyses is growing steadily, in part because the costs of replacing destroyed structures are growing and in part because of the steady increase in the population living in hazardous areas.

In the examples the scientific questions included: i) what is a fair insurance premium to cover the damage that might be experienced by a structure at a particular location given an earthquake of a given size and location? and ii) what is the probability of a fire in the Federal Lands in Oregon becoming large? The examples have in common that they are seeking probabilities and distributions. The solutions have in common that data and subject matter are basic.

Statistical methods are basic to risk assessments. This is obvious because probabilities and data are involved. It is also the case because statistics adds important things to what the engineers and scientists tend to know and do on their own. Statisticians add things like efficiency results, extensions to different data types, uncertainty measures. The statistical elements that appear in the examples include: the use of stochastic models, probabilities or rates as products, nonparametric estimation, generalized linear model data types, robust regression, biased sampling, mutual information, statistical packages, ...)

The examples presented have shown that the: stochastic approach is highly effective, that there are difficulties and opportunities, that there are solutions and that there are lots of open problems.

The use of engineering judgement needs to be referred to. It is sometimes incorporated via Bayesian methods. There is also statistical judgement. It arises in the picking of a statistical model, in deciding when a fit is adequate for continuing to the next stage and in considering what conclusions may be drawn reasonably from a completed analysis.

An interesting practical problem arises, namely how are government regulators to assess proprietary huricane and earthquake models used by insurance companies? (National Academy of Science 1998).

ACKNOWLEDGEMENTS

For collaborative work on MMI analyses I thank C. Chiang, R. A. Irizarry, and P. A. Morettin. For collaborative work on the wildfire example I thank H. K. Preisler and J. Benoit of the US Forest Service. I thank A. Der Kiureghian for the invitation to speak.

The research was supported in part by NSF Grants DMS-0102869 and DMS-0203921 and by a U.S. Forest Service Contract 02-JV-11272165-020.

REFERENCES

Beard, R. E., T. Pentikainen, and E. Pesonen (1969). *Risk Theory*. London: Chapman and Hall.

Bohman, H. (1979). Insurance protection for large risks with a low claim frequency. *Proc. 42nd Session of the ISI 48*, 125–135.

Brillinger, D. R., H. K. Preisler, and J. Benoit (2003). Risk assessment: a forest fire example. In *Festschrift for Terry Speed*, Lecture Notes in Statistics. IMS.

Brillinger, D. R. (1993). Earthquake risk and insurance. *Environmetrics 4*, 1–21.

Brillinger, D. R., C. Chiann, R. A. Irizarry, and P. A. Morettin (2001). Automatic methods for generating seismic intensity maps. *J. Applied Probability 38A*, 189–202.

Bullen, K. E. and B. A. Bolt (1985). *An Introduction to the Theory of Seismology, Fourth Edition*. Cambridge: Cambridge Press.

Chambers, J. M. and T. J. Hastie (1990). *Statistical Models in S*. Pacific Grove: Wadsworth.

Clark, G. (1999). *Betting on Lives*. Manchester: Manchester Press.

Deeming, J. E., R. E. Burgan, and J. Cohen (1977). The national fire-danger rating system - 1978. Technical Report INT-333, USDA Fores Service, Intermountain Forest and Range Experiment Station, Ogden.

Embrechts, P., C. Kluppelberg, and T. Mikosch (2000). *Modelling Extremal Events*. Berlin: Springer.

Funfits (2002). www.cgd.ucar.edu/stats/Software/Funfits.

Joyner, W. B. and D. M. Boore (1981). Peak horizontal acceleration and velocity from strong-motion records including records from 1979 Imperial Valley, California. *Bull. Seism. Soc. Amer. 71*, 2011–2038.

Martell, D., S. Otukol, and B. J. Stocks (1993). A logistic model for predicting daily people-caused forest fire occurrence. *Canadian J. Forest Research 19*, 1555–1563.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models, Second Edition*. London: Chapman and Hall.

Munich Re, . (1991). *Insurance and Reinsurance of Earthquake Risk*. Munich Re: Munich.

National Academy of Science, . (1998). *Paying the Price*. Washington, DC: J. Henry Press.

Nuclear Regulatory Commission, . (1997). *Notices 34321-34326*, Volume 62, Number 122. Washington, DC: NRC.

Powell, M. J. D. (1992). The theory of radial basis function approximation in 1990. In *Advances in Numerical Analysis*, Volume 2, Oxford. Oxford Press.

Pregibon, D. (1980). Discussion of 'Regression models for ordinal data'. *J. Roy. Statist. Soc. Ser. B 42*, 138–139.

Raynes, H. E. (1964). *A History of British Insurance*. London: Pitman.

Reid, F. (2000). The mathematician and the banknote: Carl Friedrich Gauss. *Parabola 36*(2), 2–9.

Stover, C. W., B. G. Reagor, F. Baldwin, and L. R. Brewer (1990). Preliminary isoseismal map for the Santa Cruz (Loma Prieta), California earthquake of October 18, 1989 UTC. Technical Report 90-18, National Earthquake Information Center, Denver.