

GUIDE TO NOTATION

This is not a complete guide to notation in the thesis, but a summary of symbols used with consistent meanings throughout. Other symbols have different meanings in different sections, particularly the more mathematical.

α	shape parameter of Weibull distribution - dimensionless
d	source-receptor distance in units of hundreds of kilometres in Chapter 2 and in metres otherwise
l	loglikelihood of data
λ	decay parameter of radionuclide in units of s^{-1}
λ_w	washout coefficient of radionuclide in units of s^{-1}
M	mean of exposure distribution in same units as μ
μ	scale parameter of Weibull distribution in μCism^{-3} for exposures to time-integrated air contamination and μCim^{-2} for those to dry or wet deposition
P_s	probability of exposure to a release of duration s hours
$p(\theta)$	probability that source geostrophic wind is directed into the 30° sector centred on θ° , restricted to winds $>5 \text{ ms}^{-1}$
$p_w(\theta)$	probability that in wet conditions source geostrophic wind is directed into the 30° sector centred on θ° , restricted to winds $>5 \text{ ms}^{-1}$
S	standard deviation of exposure distribution in same units as μ
σ	scale parameter of generalized Pareto distribution in Chapter 8, in units of pCism^{-3} for exposures to time-integrated air concentration and pCim^{-2} for those to dry or wet deposition.
t	duration of release of radionuclide to the atmosphere in units of hours
v_d	deposition velocity of radionuclide in units of ms^{-1}

APPENDIX: THE MODEL IN USE

A.1 Introduction; the probability of exposure

Suppose that it is proposed to site a nuclear installation at the city of Luxembourg. In this appendix the methods developed in the thesis are used to assess the consequences to an inhabitant of the city of Brussels of a release of one Curie of Cs_{137} from the installation at an arbitrary time, for various release durations.

The cities are situated 180 kilometres apart, and the angle from north subtended by Brussels at Luxembourg is approximately $\theta = 315^\circ$. Luxembourg has latitude 49.37° north and longitude 6.08° east. The estimation of the probability of exposure to air contamination or wet deposition is performed using the results of Chapter 2. Recall that

$$\hat{P}_3 = \exp\{ \hat{\beta}_0 + \hat{\beta}_1 \log\{p(\theta)\} + \hat{\beta}_2 d \} \quad \dots A.1.1$$

and

$$\hat{P}_t = 1 - \exp\{ -\hat{P}_3 (t/3)^{\hat{\delta}} \} \quad \dots A.1.2.$$

Here \hat{P}_s is the estimated probability of exposure due a release of duration $s > 3$ hours, $p(\theta)$ is the long-run proportion of source geostrophic winds of speed 5 m/s or more directed into the sector of arc 30° centred on θ° , d is the source-receptor distance in hundreds of kilometres, and the values of the parameters $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\delta}$ may be found in Tables 2.1 and 2.3. To calculate the probability of exposure to wet deposition, $p(\theta)$ should be replaced in A.1.1 by $p_w(\theta)$ - the long-run proportion of source geostrophic winds directed into the sector of arc 30° centred on θ° restricted to occasions when it is raining at the source - and the values of $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\delta}$ in Tables 2.2 and 2.4 used.

The appropriate values of $p(\theta)$ and $p_w(\theta)$ are read off from Figure 2.12. Luxembourg is very close to the edge of its grid

element (8E, 49N); it may be wise to bear this in mind. For the element (8E,49N), $p(315) \approx p_w(315) \approx 0.06$. There is only a very small difference between these values and the values $p(\theta) \approx p_w(\theta) \approx 0.05$ for the element (4E, 49N), so the choice of value 0.06 is not crucial.

There is yet the choice of which set of parameter values uniform or Mediterranean - to use. Consultation with the naive classifications in Figures 2.14 and 2.16 shows that the uniform values of $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\delta}$ are appropriate, a conclusion confirmed by the more detailed classifications in Figures 2.15 and 2.17 and knowledge of the climatology of the site.

Use of equations A.1.1 and A.1.2 then gives the estimates of P_t shown in Table A.1. That is, for exposures to air contamination due to a release of duration three hours,

$$\log\{\hat{P}_3\} = -1.12 + 0.284\log\{p(315)\} - 0.110d,$$

and with $p(315)=0.05$ and $d=1.8$, $\log\{\hat{P}_3\} = -2.12$ and $\hat{P}_3 = 0.12$. To evaluate the probability of exposure due to a release of duration 6 hours, equation A.1.2 is used with $t=6$ and $\delta=0.6209$. Thus

$$\hat{P}_6 = 1 - \exp\{-0.12(6/3)^{0.62}\} = 0.157.$$

The effect of using $p(315)=0.05$ is to reduce the estimates of P_t by between 0.01 and 0.02 - an insignificant amount.

The estimates are of limited value without some idea of their accuracy. This is provided for releases of duration 3 hours by use of equation 2.3.6 and the variances and correlations from Tables 2.1 for exposures to air contamination and 2.3 for exposures to wet deposition. In this latter case, recalling that

$$\text{Cov}(\hat{\beta}_j, \hat{\beta}_k) = \text{s.e.}(\hat{\beta}_j) \times \text{s.e.}(\hat{\beta}_k) \times \text{Correl}(\hat{\beta}_j, \hat{\beta}_k),$$

we find upon using this identity and substituting values from the

Release duration t	Exposures to air contamination			Exposures to wet deposition		
	est. prob. \hat{P}_t	95% confidence limits		est. prob. \hat{P}_t	95% confidence limits	
		lower	upper		lower	upper
3 hours	.120	.084	.173	.033	.016	.068
6 hours	.169	.144	.194	.055	.039	.071
12 hours	.248	.208	.288	.092	.064	.119
1 day	.355	.293	.416	.152	.104	.199
3 days	.579	.472	.687	.318	.214	.423
1 week	.769	.630	.908	.521	.353	.690

Table A.1: Estimated probabilities of exposure at Brussels due to releases from Luxembourg.

left-hand-side of Table 2.2 in to equation 2.3.7, that

$$\tau^2 = \text{Var}\{ \log(\hat{P}_3) \} = 3.865 \times 10^{-3}.$$

Since in this case the estimate of σ is $\sigma = 0.3622$ from Table 2.2, $\omega = \sqrt{(\tau^2 + \sigma^2)} = 0.368$, and the 95% confidence interval for the true value of P_3 is

$$\left(\hat{P}_3 \exp(-1.96\omega), \hat{P}_3 \exp(1.96\omega) \right) = (0.016, 0.068).$$

For releases of longer duration use must be made of equations 2.4.2, 2.4.3, and 2.4.4. Consider, for example, deriving 95% confidence limits for the probability of exposure to wet deposition for releases of duration 6 hours. In this case $\mu' = \log\{\hat{P}_3\} + \delta \log\{t/3\} = -2.87$ and $\tau = 6.217 \times 10^{-2}$, and substituting these into 2.4.4 and solving numerically gives $v^* = 3.5 \times 10^{-3}$ for $k = 1$ and $v^* = 7.0 \times 10^{-3}$ for $k = 2$, so that $J_1(\mu', \tau)$ is 0.945 and $J_2(\mu', \tau)$ is 0.93. Use of equation 2.4.2 with $n_t = 6072/t$ then gives 95% confidence interval $P_6 \pm 1.96/\sqrt{\text{Var}\{P_6\}}$: that is, the interval (0.039, 0.071) contains the true value of P_6 with probability 0.95. Such confidence intervals for P_t for several values of release duration t are shown in Table A.1.

Equation 2.4.4 may easily be solved graphically by plotting the curves $\log\{v\}$ and $\mu' + \log\{k\tau\} - \tau v$ on the same graph and finding for given τ , μ' and k the unique value v^* of v at which they intersect. Typically v^* is close to zero. Alternatively v^* may be found by a simple bisection or other line search using a programmable pocket calculator.

There is no reason to suspect unusual divergence of trajectories close to Brussels, although the general north-westerly passage of air-masses over the area of concern suggests that the equations may slightly overestimate the real long-run values of P_t . However the effect is unlikely to be big. In cases such as this the estimates

and their confidence intervals are generally very accurate.

A.2 The distribution of exposure levels

The distribution of imaginary exposure levels experienced by a Brussellois due to the hypothetical release of a single Curie of Cs₁₃₇ from the installation at Luxembourg may be estimated using the results of Chapter 3. Caesium 137 has half-life 28 years, corresponding to a decay constant $\lambda = 4.0446 \times 10^{-11} \text{ s}^{-1}$ negligible compared with puff travel times; its deposition velocity v_d is assumed to be $1.0 \times 10^{-3} \text{ ms}^{-1}$; and its washout coefficient λ_w is assumed for the purpose of the calculation to be $5 \times 10^{-5} J^{0.8} \text{ s}^{-1}$, where J is the rainfall rate in mm/hour. The source receptor distance d is $1.8 \times 10^5 \text{ m}$. Note that in this and the next section d is measured not in hundreds of kilometres but in metres.

For a release of duration three hours, the estimated mean \hat{M}_3 and standard deviation \hat{S}_3 of the time-integrated air concentration distribution are found by use of equations 3.3.1 and 3.3.2 respectively. Thus

$$\log(\hat{M}_3) = 5.155 - 0.1583d\lambda - 2.869 \times 10^{-4}dv_d - 9.263 \times 10^{-3}d\lambda_w - 0.9399\log(d)$$

and

$$\log(\hat{S}_3) = 10.41 - 0.1153d\lambda - 2.892 \times 10^{-4}dv_d + 4.209 \times 10^{-4}d\lambda_w - 1.328\log(d),$$

so that substitution for d, λ , v_d and λ_w gives for exposures to time-integrated air concentrations due to three-hour releases of Cs₁₃₇ $\hat{M}_3 = 1.74 \times 10^{-3} \text{ } \mu\text{Cism}^{-3}$ and $\hat{S}_3 = 3.29 \times 10^{-3} \text{ } \mu\text{Cism}^{-3}$. The effect of longer release duration is found using the equations

$$\hat{M}_t = \hat{M}_3(t/3)^{\hat{q}} \quad \dots\text{A.2.1,}$$

Release duration t	Estimated mean exposure \hat{M}_t (μCism^{-3})	Estimated exposure s.d. \hat{S}_t (μCism^{-3})	Coefficient of variation \hat{S}_t/\hat{M}_t	Weibull shape parameter α	Weibull scale parameter μ (μCism^{-3})
3 hours	1.741×10^{-3}	3.294×10^{-3}	1.89	0.56	1.161×10^{-3}
6 hours	1.252×10^{-3}	2.392×10^{-3}	1.91	0.56	8.347×10^{-4}
12 hours	9.111×10^{-4}	1.737×10^{-3}	1.91	0.56	6.074×10^{-4}
1 day	6.469×10^{-4}	1.261×10^{-3}	1.95	0.54	4.043×10^{-4}
3 days	3.834×10^{-4}	7.594×10^{-4}	1.98	0.54	2.396×10^{-4}
1 week	2.561×10^{-4}	5.136×10^{-4}	2.01	0.53	1.552×10^{-4}

Table A.2: Details of estimated time-integrated air concentration exposure distributions

at Bussels due to releases of Cs_{137} from Luxembourg.

α	$\Gamma(1 + 1/\alpha)$	$\left\{ \frac{\Gamma(1 + 2/\alpha)}{\Gamma(1 + 1/\alpha)^2} - 1 \right\}^{\frac{1}{2}}$	α	$\Gamma(1 + 1/\alpha)$	$\left\{ \frac{\Gamma(1 + 2/\alpha)}{\Gamma(1 + 1/\alpha)^2} - 1 \right\}^{\frac{1}{2}}$
0.4	3.323	3.141	1.35	0.917	0.749
0.45	2.479	2.606	1.4	0.911	0.724
0.5	2.0	2.236	1.45	0.907	0.701
0.55	1.702	1.965	1.5	0.903	0.679
0.6	1.505	1.758	1.55	0.899	0.659
0.65	1.366	1.595	1.6	0.897	0.640
0.7	1.266	1.462	1.65	0.894	0.622
0.75	1.191	1.353	1.7	0.892	0.605
0.8	1.133	1.261	1.75	0.891	0.590
0.85	1.088	1.181	1.8	0.889	0.575
0.9	1.052	1.113	1.85	0.888	0.561
0.95	1.023	1.053	1.9	0.887	0.547
1.0	1.0	1.0	1.95	0.887	0.535
1.05	0.981	0.953	2.0	0.886	0.523
1.1	0.965	0.910			
1.15	0.952	0.872			
1.2	0.941	0.837			
1.25	0.931	0.805			
1.3	0.924	0.776			

Table A.3: Coefficient of variation for Weibull distribution as a function of shape parameter α .

and

$$\hat{S}_t = \hat{S}_3 (t/3)^{\hat{\delta}_2} \quad \dots \text{A.2.2.}$$

Appropriate values of $\hat{\delta}_1$ and $\hat{\delta}_2$ are found in Table 3.8 and are $\hat{\delta}_1 = -0.4761$ and $\hat{\delta}_2 = -0.4617$ for exposures to air contamination. The means M_t and standard deviation S_t of estimated exposure distributions for releases of duration t hours are shown in Table A.2.

The shape and scale parameters α and μ of the estimated Weibull distribution of exposures may be found by solving the equations

$$\hat{S}_t / \hat{M}_t = \sqrt{\left\{ \frac{\Gamma(1+2/\alpha)}{\Gamma(1+1/\alpha)^2} - 1 \right\}},$$

and

$$\hat{M}_t = \mu \Gamma(1+1/\alpha)$$

for α and μ . Here $\Gamma(y)$ is the gamma function $\int_0^{\infty} u^{y-1} e^{-u} du$. The first of these equations involves α alone and may be solved by reading off the appropriate values of α and $\Gamma(1+1/\alpha)$ from Figure A.1, or by interpolation in Table A.3 if need be, or numerically by a bisection or other straightforward root search on a programmable pocket calculator. Values of α are usually in the range 0.5-1.0. Values obtained graphically for the present example are given in Table A.2.

The mean, standard deviation, and value of the Weibull scale parameter μ - all in microCuries - for the corresponding dry deposition distributions are found by multiplying \hat{M}_t , \hat{S}_t and μ_t for the air concentration distribution by the value of v_d . The value of the Weibull shape parameter α remains the same since it is scale-invariant.

The corresponding calculations for the wet deposition distribution are carried out using equations 3.3.3 and 3.3.4 to give

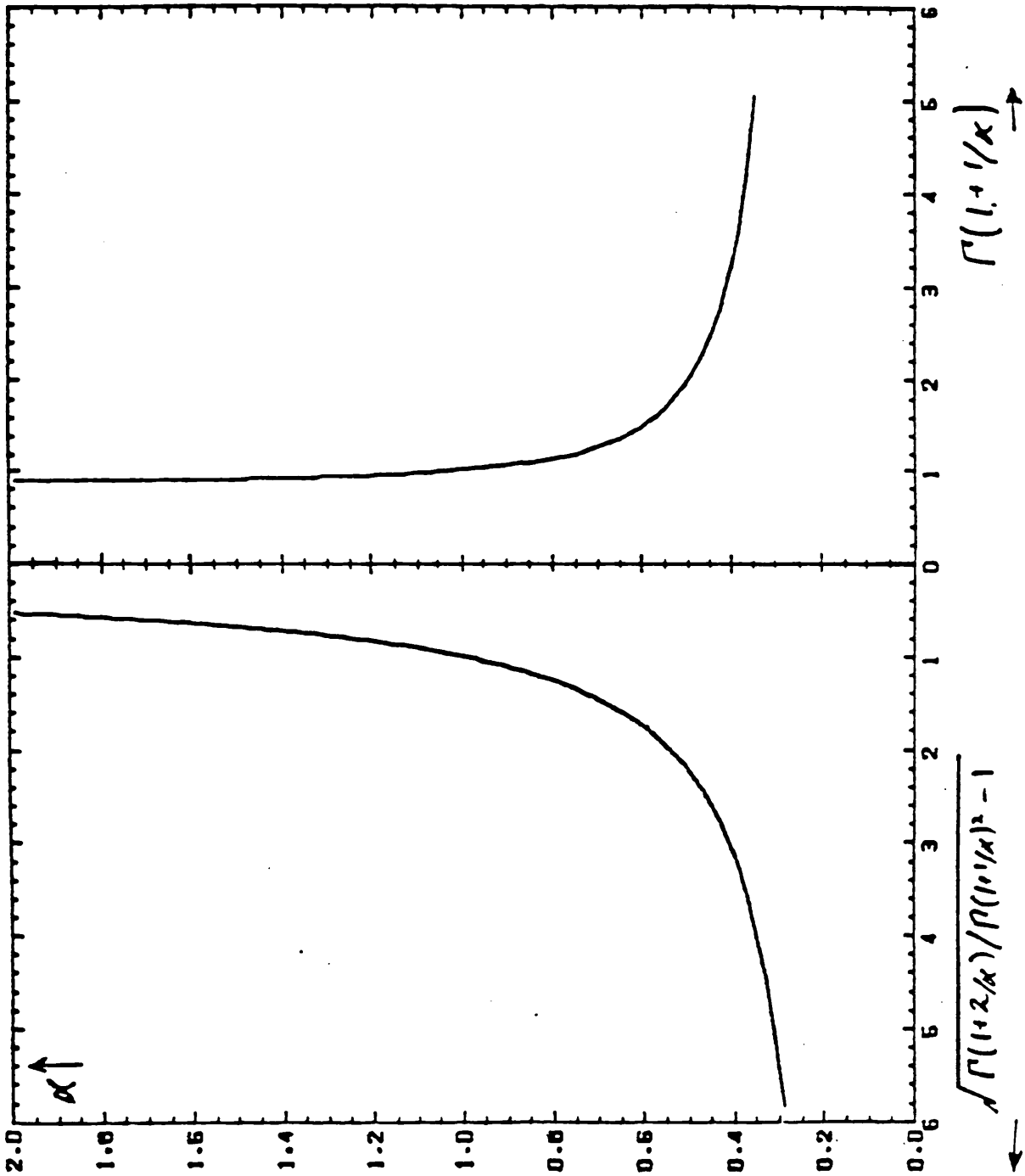


Figure A.1: Nomogram to find value of α and $\Gamma(1 + 1/\alpha)$ for given

$$\hat{S}_t / \hat{M}_t = \left\{ \frac{\Gamma(1 + 2/\alpha)}{\Gamma(1 + 1/\alpha)^2} - 1 \right\}^{\frac{1}{2}}$$

the values of \hat{M}_t and \hat{S}_t in microCuries, and then using equations A.2.1 and A.2.2 above with appropriate values of $\hat{\delta}_1$ and $\hat{\delta}_2$ from Table 3.8 to see the effect of release duration on the exposure distributions.

Once the estimated values of α_t and μ_t are established, any required percentile of the Weibull distribution of exposures

$$F_t(y) = 1 - \exp[-(y/\mu_t)^\alpha]$$

may be found. The $p \times 100\%$ point of the distribution is y_p , where $y_p = \mu_t [-\log\{1-p\}]^{1/\alpha}$. The estimated expected risk to some activity with risk function $r(y)$,

$$\int r(y) \alpha_t \mu_t^{-1} (y/\mu_t)^{\alpha-1} \exp[-(y/\mu_t)^\alpha] dy,$$

is easily found analytically or numerically. For a linear risk function with $r(y) = a + by$, the estimated expected risk is $a + b\hat{M}_t$, with estimated standard deviation $b\hat{S}_t$. The estimates will only very rarely be more than a factor of three different from the values which would have been observed had MESOS been used to compute them, provided the statistical model is used within the ranges of the parameters for which it was derived, as in the current example.

Less stress has been laid on finding confidence intervals for levels of exposure than for probabilities of exposure. They are to an extent made spurious by the degree of approximation introduced by the dispersion calculations in MESOS, and by the degree of error introduced by replacing calculated MESOS exposure distributions with estimated Weibull exposure distributions. The results of Section 3.4 indicate that the upper 70% of the statistically estimated exposure distribution is likely to lie within a factor of two to three of the distribution which would have been calculated using MESOS, that the lower 30% of the distribution will make a tiny contribution to the

total exposure, and that the lower 30% of the MESOS exposures will typically be greater than their statistically estimated counterparts. If explicit confidence intervals for M_t and S_t are needed, they can be found using Tables 3.5 and 3.7 and Normal approximations, as discussed on page 176 of the thesis.

A.3 High levels of exposure

The likely distribution of high levels of exposure may be predicted using the methods of Chapter 8. For a release of duration t hours, setting $\xi_0=0$ in equation 8.1.1 and using equation 8.2.3 with the appropriate parameter values in Table 8.1 yields for time-integrated air concentration distributions the thresholds and values of the Poisson parameter ρ_t in Table A.4. Values of the geometric clustering probability q_t - which is zero for releases of duration longer than one day - are found in Table 8.2, and use of equation 8.3.1 with parameter estimates for $\xi_0=0$ enables calculation of the appropriate values of the generalized Pareto scale parameter σ_t . For $\xi_0=0$ the estimated value of the generalized Pareto shape parameter is $k = -0.233$. In the present case, for example, with $\lambda = 4.0446 \times 10^{-11} \text{ s}^{-1}$, $v_d = 1.0 \times 10^{-3} \text{ ms}^{-1}$, $\lambda_w = 5.0 \times 10^{-5} \text{ J}^{0.8} \text{ s}^{-1}$, and $d = 1.8 \times 10^5 \text{ m}$, we find from equation 8.1.1 that with $\xi_0=0$ the threshold is $7.069 \times 10^{-3} \text{ } \mu\text{Cism}^{-3}$ for exposures to air contamination due to releases of duration $t=3$ hours. From Table 8.2(a) the appropriate value of q_t is 0.5, and equation 8.2.3 gives

$$\begin{aligned} \rho_t &= \exp\{ \alpha + 0.6 \log\{p(\theta)\} + \log\{t/3\} \} \\ &= \exp\{ -4.18 + 0.6 \log\{0.05\} + \log\{3/3\} \} \\ &= 2.8 \times 10^{-3}, \end{aligned}$$

using the appropriate value of α from Table 8.4(a) and setting $p(315)=0.05$ and $t=3$. Finally σ_t is found using the equation

$$\begin{aligned}\sigma_t &= \exp\{ 23.441 - 0.135d\lambda - 2.304 \times 10^{-4}dv_d - 9.893 \times 10^{-3}d\lambda_w \\ &\quad - 1.227\log(d) - 0.720\log(t/3) \} \\ &= 2.538 \times 10^{-3} \mu\text{Cism}^{-3}\end{aligned}$$

when $t=3$ hours. Values of q_t , ρ_t , and σ_t for releases of longer duration appear in Table A.4. The estimated probability of a single release of duration t hours exposing the receptor is $\rho_t/(1-q_t)$; its estimated variance is a little less than $t\rho_t/6072(1-q_t)^2$, using the fact that estimation of ρ_t was based on either 2024 releases of duration 3 hours in 1973 or 2880 releases of duration 3 hours in 1976. Thus if $t=3$ hours, the estimated probability of an exposure over the threshold at Brussels is $\rho_t/(1-q_t) \doteq 0.006$ or so.

Exposures which exceed the threshold level for a given release duration follow the generalized Pareto distribution with form

$$F_t(y) = 1 - (1 - ky/\sigma_t)^{1/k}$$

since $k \neq 0$. The $p \times 100\%$ point y_p of the distribution is easily found by setting $F_t(y_p) = p$ and solving for y_p in the appropriate equation above, so $y_p = \sigma_t \{ 1 - (1-p)^k \} / k$, since $k \neq 0$. These formulae may be used in the same way as the Weibull formulae in the previous section.

Suppose now that T consecutive releases each of duration t hours escape from the installation, and that it is required to find the distribution of the maximum exposure Y_{\max} which is experienced in Brussels, there exceeding the threshold. Under the clustering model described in Chapter 8 this is

$$\text{Prob}(Y_{\max} < y) = \exp\left\{ \rho_t T \left[\frac{\{1-q_t\}F_t(y)}{\{1-q_t\}F_t(y)} - 1 \right] \right\}.$$

The percentile y_p corresponding to a given probability $p \exp\{-\rho_t T\}$ may be found by solving the equation $\text{Prob}(Y_{\max} < y_p) = p$, and is

$$y_p = \sigma_t \{ 1 - (1-p^*)^{1/k} \} / k$$

Release duration t	Threshold (μCism^{-3})	ρ_t	q_t	σ_t (μCism^{-3})
3 hours	7.069×10^{-3}	2.8×10^{-3}	0.5	2.538×10^{-3}
6 hours	4.372×10^{-3}	5.6×10^{-3}	0.35	1.541×10^{-3}
12 hours	2.707×10^{-3}	1.13×10^{-2}	0.25	9.353×10^{-4}
1 day	1.677×10^{-3}	2.26×10^{-2}	0.1	5.678×10^{-4}
3 days	7.845×10^{-4}	6.79×10^{-2}	0.	2.574×10^{-4}
1 week	4.367×10^{-4}	0.158	0.	1.399×10^{-4}

Table A.4: Calculated parameter values for exposure to high levels of Cs_{137} air contamination.

since $k \neq 0$, where

$$p^* = \frac{\rho_t T + \log(p)}{\rho_t T + q_t \log(p)}$$

If $p < \exp\{-\rho_t T\}$, then $y_p = 0$ since there is an atom of probability of size $\exp\{-\rho_t T\}$ at $y=0$, corresponding to the event of no exceedances of the threshold in an interval of length T . For given values of p and T , p^* and y_p are easily obtained from these formulae. For example, if $t=6$, $T=4$ and $p=0.9$, the value of y_p obtained may be regarded as the value which would be exceeded at Brussels by the maximum of four consecutive 6-hour releases each of size one Curie only in one out of ten such release incidents, on average.

The results of Section 8.4 show that levels of exposure predicted using the generalized Pareto distribution will lie within a factor 1.3-1.6 of their MESOS counterparts, for all types of contamination. Provided that the model is used only for nuclides whose depletion parameters lie in the range of those used to construct it, as here, prediction of probabilities of high exposures will also be very accurate. Wet deposition probabilities may be underestimated by a factor of up to four at receptors just downwind of the source or where there is orographic enhancement of rainfall, but are generally predicted more accurately. Neither of these considerations applies here, so there is no reason to disbelieve the predictions. If required, variances of predicted return values can be found using equations 8.3.5 and 8.3.8 of the thesis, with values from Tables 8.6 and 8.8. Use of confidence intervals based on these variances, however, seems likely to give the corresponding values an air of spurious accuracy, bearing in mind the likely range of error in the MESOS calculations.

A STATISTICAL MODEL FOR CONTAMINATION

DUE TO

LONG-RANGE ATMOSPHERIC TRANSPORT OF

RADIONUCLIDES

Anthony Christopher Davison

A thesis submitted for the degree of Doctor of Philosophy
in the University of London.

Environmental Safety Group, and Statistics Section,
Nuclear Power Section, Department of Mathematics
Department of Mechanical Engineering

Imperial College of Science and Technology

August 1984

ABSTRACT

The study of contamination due to long-range atmospheric transport of radionuclides after a release from a nuclear installation is relevant to siting policy for power stations, especially for the assessment of its implications for the peoples of other states. MESOS is a computer model which uses real weather data to calculate exposures due to atmospheric transport of radioisotopes, based on approximations to the dispersion and depletion of notional puffs of nuclides in the atmosphere.

A statistical analysis of the MESOS data is performed to enable the estimation of long-run distributions of exposure - to air contamination and dry and wet deposition - at points remote from any source in Western Europe from which a release of duration between three hours and one week is deemed to have taken place. This statistical model works quite well compared with MESOS and provides simple cheap predictions of exposure distributions.

Particular attention must be paid to rare episodes leading to big exposures, so statistical methods which extend the threshold models used by hydrologists are developed. The generalized Pareto distribution - the natural parametric family for excesses of continuous random variables over high thresholds - is studied in detail, and characterized as 'threshold-stable'. Estimation problems tackled include: fitting the distribution to complex data by maximum likelihood; sensitivity of maximum likelihood estimators to sample extremes; small-sample behaviour of maximum likelihood estimators; and the efficiency of least squares and moment estimators. Diagnostic procedures developed are based on residuals, a test of fit, and the study of influence.

These techniques are applied to the MESOS data with the aim of modelling high exposure episodes as accurately as possible.

For my mother and father

ACKNOWLEDGEMENTS

I am very grateful to my supervisors, Drs H.M. ApSimon and R.L. Smith, for their invaluable guidance and help - not to mention their patience - throughout the course of this work. Their suggestions did much to shape it.

Others too numerous to mention - or even remember - made useful comments at different stages; I thank particularly Professor D.R. Cox. I thank also Philippa Manning, who produced and interpreted the MESOS windroses.

I am obligated to Dr A.J.H. Goddard, who gave administrative support for the work, which was funded by CEC contracts 1191-81-11 L/V and 1231-83-12 L/V.

I am grateful to the Statistics Section research students - especially Patty, Charles, and Bianca - for their encouragement; and to my fellows at Maresfield Lodge for suffering my peninsularity during the past few months.

Finally, I thank Annie Macpherson and Elizabeth Hancock - who by their efficient typing transformed a sometimes illegible manuscript - and Haroldo Cantanhede, who drew some of the figures.

TABLE OF CONTENTS

Abstract		1
Dedication		2
Acknowledgements		3
Table of Contents		4
List of Tables		6
List of Figures		9
Chapter 1	Introduction	12
1.1	Background, motivation, and discussion	12
1.2	The MESOS exposure database	24
1.3	The structure of this thesis	29
Chapter 2	Probabilities of Exposure	35
2.1	Introduction	35
2.2	Regression analysis: tools, diagnostics, and transformations	42
2.3	Exposure probabilities for releases of duration three hours	47
2.4	Exposure probabilities for releases of longer duration	66
2.5	A taxonomy of windroses	74
2.6	A verification study	84
Chapter 3	The Distribution of Levels of Exposure	92
3.1	Introduction, and exploratory analysis	92
3.2	Confirmatory analysis	107
3.3	The effect of covariates	120
3.4	Verification of the fitted equations	146
Chapter 4	Excesses over High Thresholds	178
Chapter 5	The Generalized Pareto Distribution	184

Chapter 6	Estimation of the Distribution	196
6.1	Maximum likelihood estimation: generalities	198
6.2	Maximum likelihood estimation: simple random samples	203
6.2.1	Existence of estimators	203
6.2.2	Their small-sample properties	204
6.2.3	Influence and censoring	212
6.3	Some other estimators	216
6.3.1	Least squares estimators	216
6.3.2	Moment estimators	218
6.4	Two tests for tail weight	220
Chapter 7	Diagnostics	226
7.1	Residuals	226
7.2	A score test	233
7.3	Influence	238
Chapter 8	High Exposure Episodes	243
8.1	Introduction	243
8.2	A clustering model for extremes	246
8.3	High exposure levels	255
8.4	A verification study	272
Chapter 9	Summary, Conclusions, and Discussion	285
References		291

LIST OF TABLES

1.1	Parameters of nuclides used in data analysis	28
1.2	Sources and receptors used in data analysis	30
1.3	Sources and receptors used for model verification only	32
1.4	Parameters of nuclides used for model verification only	33
2.1	Details of regression equations for probabilities of exposure in all conditions due to releases of duration three hours	55
2.2	Details of regression equations for probabilities of exposure in wet conditions due to releases of duration three hours	61
2.3	Parameter estimates for dependence of probabilities of exposure in all conditions on release duration	69
2.4	Parameter estimates for dependence of probabilities of exposure in wet conditions on release duration	71
2.5(a,b)	Comparison of MESOS and statistically predicted %-probability of exposure at 16 receptor points for several release durations: dry deposition	85
2.6(a,b)	Comparison of MESOS and statistically predicted %-probability of exposure at 16 receptor points for several release durations: wet deposition	88
2.7	Numbers out of 16 receptors at which MESOS exposure probabilities lie outside their 95% confidence limits	91
3.1	Estimated Weibull parameters for MESOS exposure distributions	106
3.2(a-c)	Anderson-Darling statistics for exposure data	108
3.3	Loglikelihood ratio statistics for Weibull distribution within generalized gamma family	118
3.4	Weibull and generalized gamma parameter estimates: Cadarache Xe ₁₃₅ three-hourly time-integrated air concentration data	119
3.5	Details of regression equations for Weibull air concentration distributions	126
3.6(a,b)	Comparison of Weibull and actual means and standard deviations of three-hourly exposure distributions	133
3.7	Details of regression equations for Weibull wet deposition distributions	138
3.8	Details of regression equations for dependence of Weibull means and standard errors on release duration	145

3.9	Percentage contributions to mean exposure by excesses over thresholds: Kr_{85} distributions	175
6.1	Numbers out of 100 simulated random samples in which likelihood equation roots could be found	204
6.2	Standardized bias of maximum likelihood estimates for generalized Pareto distribution	206
6.3	Multivariate skewness and kurtosis of maximum likelihood estimates in 1000 simulated samples from the generalized Pareto distribution	208
6.4	Proportion of likelihood confidence region statistics less than the 95% and 99% points of their asymptotic distribution in 1000 simulated samples of size n , for different underlying values of k	210
6.5	Observed 95% and 99% points of statistics for likelihood-based confidence regions, in 1000 samples of size n , for different underlying values of k	211
6.6	Asymptotic relative efficiency(%) of upper-truncated maximum likelihood estimation of the generalized Pareto distribution, for different truncation probabilities α	213
6.7(a-c)	Comparison of influence curves	215
6.8	Asymptotic relative efficiency(%) of least squares estimates of generalized Pareto distribution	217
6.9	Asymptotic relative efficiency(%) of moment estimates of generalized Pareto distribution	219
6.10	Standardized variances and covariances of moment estimates k^* and σ^* of generalized Pareto distribution	220
6.11	Power of S and T for different values of k : proportion of 1000 simulations significant at one-sided 5% level	224
7.1	Residual comparison function $G(\cdot)$ and its first few derivatives for several values of k	232
7.2	Observed significance points of score test of fit W_u	237
7.3	MESOS exposure dataset and influence diagnostics: time-integrated air contamination through 1976 800 km north of Mol due to unit releases of Kr_{85} during every three-hour period; exceedances of $1.1 \times 10^{-10} \text{ Cism}^{-3}$	240
8.1	Comparison of observed and expected cluster sizes for exposure to air contamination due to three-hour releases of Kr_{85} from Mol in 1976: threshold $\xi_0=0$	250
8.2(a,b)	Values of clustering probability q for different thresholds and release durations	251
8.3	Comparison of observed and expected cluster sizes for exposure to wet deposition due to three-hour releases	

	of $I_{131}(p)$ from Heysham during 1973: threshold $\xi_0=0$	253
8.4(a,b)	Dependence of Poisson rate parameter ρ_t on threshold parameter ξ_0	255
8.5(a,b)	Reductions in deviance due to introduction of successive covariates for different threshold levels ξ_0	257
8.6	Parameter estimates for different thresholds for exposures to time-integrated air concentrations	258
8.7	Parameter estimates for different thresholds for exposures to wet deposition	264

LIST OF FIGURES

1.1(a,b)	Features of long-range dispersion modelled in MESOS	14
1.2	Histograms for time-integrated air concentrations due to three-hourly releases of $I_{131}(p)$ from Mol during 1973	18
1.3	Time-series of exposures to time-integrated air concentration at receptor 800 km north of Mol due to unit releases of Kr_{85} every three hours during 1976	23
1.4	Area covered by 1973 meteorological database	25
1.5	Area covered by 1976 meteorological database	27
1.6	Radial grid system for Mol	31
2.1(a-c)	Probability of air contamination and wet deposition due to daily releases from Mol during 1973	36
2.2(a,b)	Proportion of MESOS geostrophic windroses greater than 5 m/s directed from 30° sectors at Mol during 1973	37
2.3(a,b)	Trajectory roses at various distances from the source, for releases of duration three hours	38
2.4(a,b)	Proportions of MESOS geostrophic windroses greater than 5 m/s directed from 30° sectors	40
2.5	Partially maximized loglikelihood $\ell_{\max}(\lambda)$ for Box-Cox power transformation of probabilities of exposure to air contamination	51
2.6	Nesting structure of regression models fitted to log-probabilities of exposure to air contamination	53
2.7(a-d)	Residual plots for regression of log-probabilities of exposure to air contamination	57
2.8	Nesting structure of regression models fitted to log-probabilities of exposure to wet deposition	60
2.9(a-d)	Residual plots for regression of log-probabilities of exposure to wet deposition	64
2.10	Nesting structure for dependence of probabilities of exposure to air contamination on release duration	69
2.11	Nesting structure for dependence of probabilities of exposure to wet deposition on release duration	71
2.12	A map of European geostrophic windroses: proportion of winds of speed 5 m/s or more compared with all windspeeds, for winds in all conditions and winds in wet conditions	75

2.13	Map area and grid elements for windrose classification	76
2.14	A simple classification of geostrophic winds of speed 5 m/s or more in all conditions	78
2.15	A detailed classification of geostrophic winds of speed 5 m/s or more in all conditions	79
2.16	A simple classification of geostrophic winds of speed 5 m/s or more in wet conditions	81
2.17	A detailed classification of geostrophic winds of speed 5 m/s or more in wet conditions	82
3.1	Comparison of moments for different distributions	96
3.2(a-e)	Standardized moment plots for exposure distributions due to three-hour releases	97
3.3(a-d)	Histograms of exposure distributions	100
3.4(a,b)	Cumulative frequency distributions of exposures for unit releases over three hours from Mol during 1973	105
3.5(a-h)	Comparison of MESOS and fitted Weibull cumulative exposure distributions	112
3.6(a-d)	Residual plots for regression of Weibull log-means; air contamination data	123
3.7(a-d)	Residual plots for regression of Weibull log-standard deviations; air contamination data	127
3.8	Scatter plots for nuclide exposures north of Mol through 1973	131
3.9(a-d)	Residual plots for regression of Weibull log-means; wet deposition data	139
3.10(a-d)	Residual plots for regression of Weibull log-standard deviations; wet deposition data	141
3.11(a-h)	Comparison of MESOS and statistically predicted cumulative exposure distributions; Case 1 air contamination due to releases from Hannover	148
3.12(a-d)	Comparison of MESOS and statistically predicted cumulative exposure distributions; Case 2 air contamination due to releases from Hannover	152
3.13(a-d)	Comparison of MESOS and statistically predicted cumulative exposure distributions; Case 4 air contamination due to releases from Hannover	155
3.14(a-f)	Comparison of MESOS and statistically predicted cumulative exposure distributions; Kr ₈₅ air contamination due to releases from Heysham	157

3.15(a-f)	Comparison of MESOS and statistically predicted cumulative exposure distributions; Xe ₁₃₃ air contamination due to releases from Cadarache	160
3.16(a-f)	Comparison of MESOS and statistically predicted cumulative exposure distributions; Xe ₁₃₅ air contamination due to releases from Mol in 1973	163
3.17(a-f)	Comparison of MESOS and statistically predicted cumulative exposure distributions; Case 1 wet deposition due to releases from Stuttgart	167
3.18(a-f)	Comparison of MESOS and statistically predicted cumulative exposure distributions; Case 4 wet deposition for releases from Stuttgart	170
5.1	Mean excesses over thresholds: Kr ₈₅ time-integrated air concentrations 800 km north of Mol due to unit releases over three-hourly periods through 1976	187
6.1(a,b)	Power of S and T in small samples	225
7.1	Comparison of deviance and Cox-Snell residuals for generalized Pareto distribution	231
7.2	Jackknifed maximum likelihood estimates for data in Table 7.3	242
8.1	Partially maximized loglikelihood $\ell_{\max}(k)$ for high exposures to air contamination; threshold $\xi_0=0$	260
8.2(a,b)	Residual plots for model for high exposures to air contamination; threshold $\xi_0=0$	262
8.3	Partially maximized loglikelihood $\ell_{\max}(k)$ for high exposures to wet deposition; threshold $\xi_0=0$	266
8.4(a,b)	Residual plots for model for high exposures to wet deposition; threshold $\xi_0=0$	268
8.5(a-f)	Plots of differences of log-observed and log-expected order statistics for high levels of air contamination due to releases from Hannover	275
8.6(a-f)	Plots of differences of log-observed and log-expected order statistics for high levels of wet deposition	281

1. INTRODUCTION

1.1 Background, motivation, and discussion

Article 37 of the Euratom Treaty among the nations of the European Economic Community reads:

Each member state shall provide the Commission with such general data relating to any plan for the disposal of radioactive waste in whatever form as will make it possible to determine whether the implementation of such plan is liable to result in the radioactive contamination of the water, soil, or air space of another member state.

Although direct contamination of the water or soil of another country could happen as a result of a nuclear discharge, by far the most serious risk to the majority of its population would be due to transport of radioactive material in the air and consequent exposures; either directly, from air contamination and irradiation from deposition on the ground, or indirectly - from consuming produce from polluted farmland, for example. For this reason work in the Environmental Safety Group at Imperial College has been directed over the past decade towards the study of long-range atmospheric transport of radionuclides, to provide methods whereby the implications of the Article 37 requirements may be assessed.

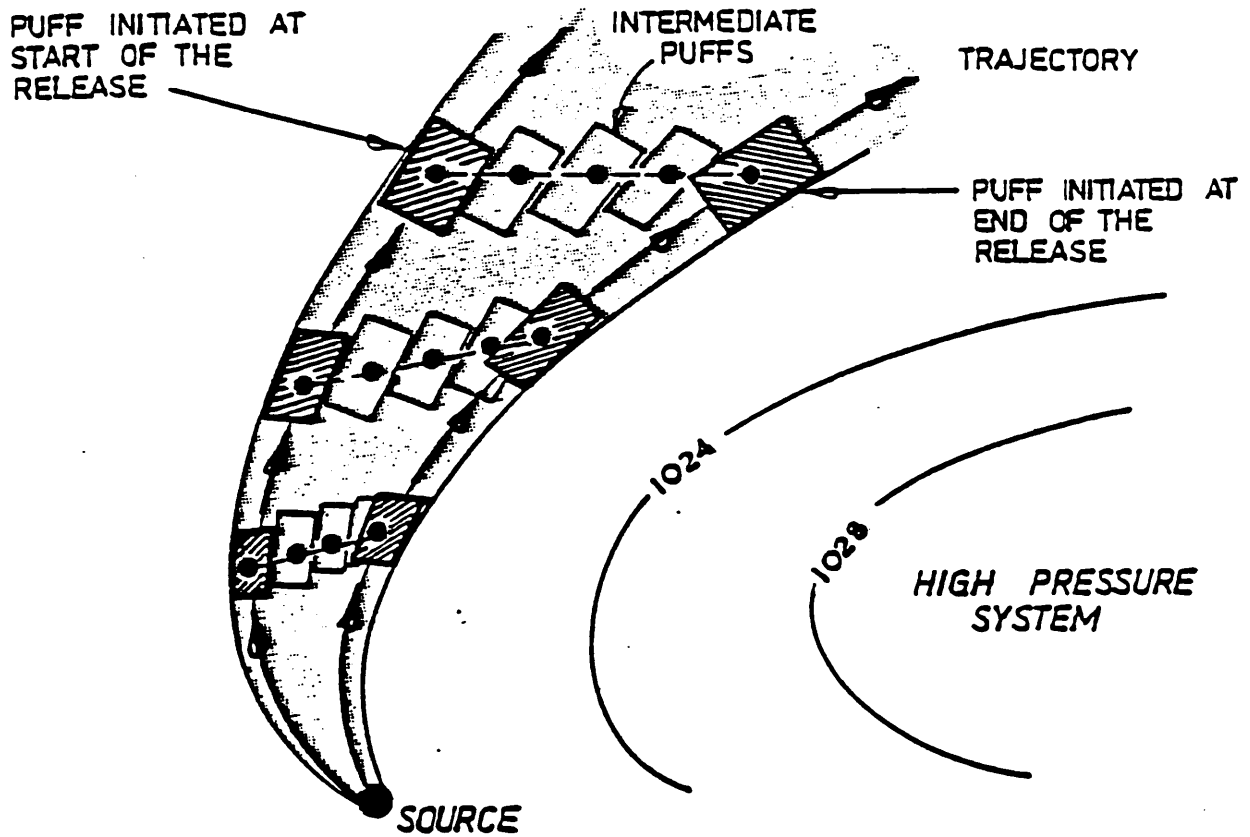
The MESOS computer model for long-range transport, dispersion, and deposition of radionuclides, described by ApSimon et al.(1983), has formed the basis for much of the work of the Group.

The idea fundamental to MESOS is this: a continuous constant release of a nuclide from a source at a known location in Western Europe is modelled as a plume interpolated between an imaginary series of tracked puffs released every three hours. The positions of the puffs, concentrations of material near the ground, and dry and wet deposition, are calculated every ten minutes for the first three hours, and hourly thereafter. The puff trajectories develop in as

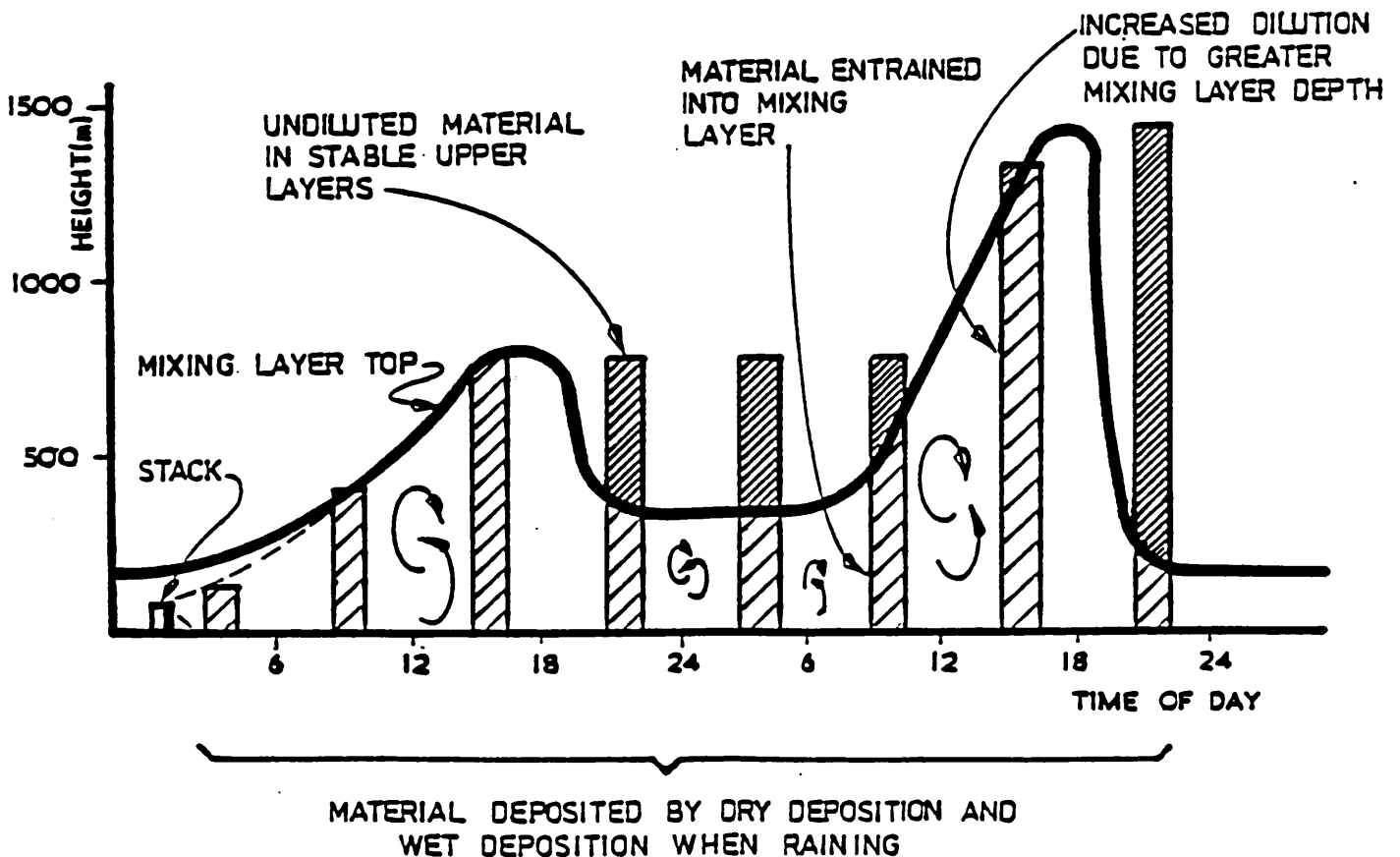
realistic a way as possible based on 'present weather' observations from meteorological stations and ships throughout Western Europe. Geostrophic winds which advect the puffs are deduced from a series of pressure fields interpolated between three-hourly measurements. The pressure fields, and hence the trajectories, reflect the influence of major orographic features.

Each puff is treated as a vertical column of pollutant which expands laterally and evolves vertically in a way which depends on the local behaviour of the boundary layer at the time of passage. The mean velocity of the puff within the boundary layer is derived from a wind profile - determined by the roughness of the underlying surface, the mixing layer depth, and current atmospheric stability - which backs in direction and decreases in strength from the boundary layer to the ground. Convective or mechanically induced turbulence may carry material aloft, where it may be isolated if the mixing layer subsides; later it may be re-entrained if the mixing layer grows again. Mixing layer depth follows a diurnal cycle, influenced by the roughness of the underlying terrain and wind strength. As travel time increases, the lateral dispersion of the puffs is gradually dominated by synoptic divergence between their paths, an important effect neglected in simpler models. Vertical and lateral dispersion are illustrated in Figure 1.1, taken from Wrigley(1982).

Each combination of nuclide, exposure mode, and source has two bodies of associated results. Firstly, the puff histories: lateral and vertical dispersion of each puff, the vertical distribution of material within it, and deposition; and rainfall and other meteorological variables. These are recorded in the Lagrangian framework of the puff as functions of time since its release, until the puff leaves the map area or four days have elapsed. The effect on overall exposure levels of ignoring puffs under these



(a) Lateral Dispersion



(b) Vertical Dispersion

Figure (1.1) Features of Long-Range Dispersion Modelled in MESOS

circumstances is small: most puffs cross the map area in between one and three days, and although they may pass over the grid again after leaving it they are generally relatively very dispersed due to the cumulative impact of several cycles of diurnal mixing. Consequently little is lost by neglecting them. As the plume passes over a large grid of points centred at the source, the notional pollution at each point due to different modes of exposure - air contamination, and dry and wet deposition - is recorded. These receptor histories are the second set of data: time-series of exposures due to successive puffs, observed at the fixed Eulerian grid of points.

One of the features of the MESOS exposure database is that given a nuclide, release duration, source, and receptor, a complete probability distribution of exposures is available. Its elements correspond to all the variety of different weather conditions that a puff may experience along its path from source to receptor and which may prevail while it crosses the receptor. This is a big improvement on simple models which only allow the long-run exposures for a fairly crude breakdown of different types of source meteorology - or in some cases only long-run mean exposures - to be calculated. It has the additional virtue of making possible the isolation and study of types of weather patterns leading to particular pollution episodes; for example those generating especially high exposures at a receptor or set of receptors. The distributions of exposure levels can be integrated with respect to the distribution over the map area of anything for which data are available - people, milk production, or anything else of interest. The risk to that population or activity associated with notional releases from the installation can then be evaluated.

MESOS has advantages over many other long-range dispersion models, which are often based on straight-line trajectories and

source weather behaviour, perhaps modified in various ways to make the model more realistic. The grid of real weather data used in MESOS enables the trajectories to be much more complex: they may be diverted around mountain ranges; they may loop so as to exposure grid-points more than once; and they may turn systematically at long ranges due to prevailing weather patterns. Another important point is that the spatio-temporal evolution of each puff is treated in some detail. Like the pressure field, the occurrence of rainfall is based on three-hourly 'present weather' data smoothed over grid squares: it varies in space and time in a fairly realistic way. In particular, features incorporated in the 'present weather' data such as enhancement of rain over mountains are reflected in the model.

However there are many other weather effects which MESOS does not take into account, such as thermal winds, nocturnal jets, and small-scale topographic winds. Nor does the puff model allow material to be lost to the upper troposphere, although this may occur in frontal regions. Moreover the resolution of the rain data is too coarse for accurate representation of the effect of wet deposition at short distances from the source, where a puff may be comparable in size with a convective rain cell: wet deposition at such distances may be more patchy than the MESOS data - which are smoothed over grid cells of area $\sim 10^4$ square kilometres - suggest.

The verification of a model both by internal calibration and comparison with other models designed for the same purpose, and if possible its validation against suitable observational or - much better - experimental data, are obviously important if the potential user is to have confidence in its results.

As far as possible MESOS has been so calibrated that its derived marginal distributions of meteorological variables correspond to their observed distributions; but the difficulty of measuring many

atmospheric quantities simultaneously makes it difficult to be sure that it always mimics their joint behaviour well.

Various other simpler models have been compared with MESOS. ApSimon and Goddard(1983, Appendix A) contrast MESOS exposures with those from two other models; one a straight-line plume model with a single boundary layer stability category, and the other a trajectory model with puffs released every twelve hours but vertical and lateral puff development independent of changing weather conditions. Exposure distributions for the other two models are more homogeneous than for MESOS because of the greater variability of possible puff histories it can allow; and high exposures calculated with MESOS tend to be greater because the other models allow fewer meteorological parameters to vary. Figure 1.2, taken from their report, shows histograms of the logarithms of time-integrated air concentrations at sixteen receptors, for notional unit daily releases of Iodine 131(p) from Mol in Belgium through 1976. The extended lower tail of the histograms is typical of the MESOS results but not of the others, which do not cater for the possibility of exposures from indirect trajectories.

It is much easier to verify a long-range dispersion model by calibrating it against weather data and comparing it with other models than it is to validate it by checking its exposures with real ones, because suitable data are hard to find. Experimental data adequate to check models against external reality would be invaluable, and inert tracer studies are at present being performed in the United States (Ferber and Heffter, 1983). Exposure data from properly designed experiments to test mesoscale dispersion models over Europe do not exist - and possibly never will. There are few detailed data from observational studies at long distances from a source with a known release pattern.

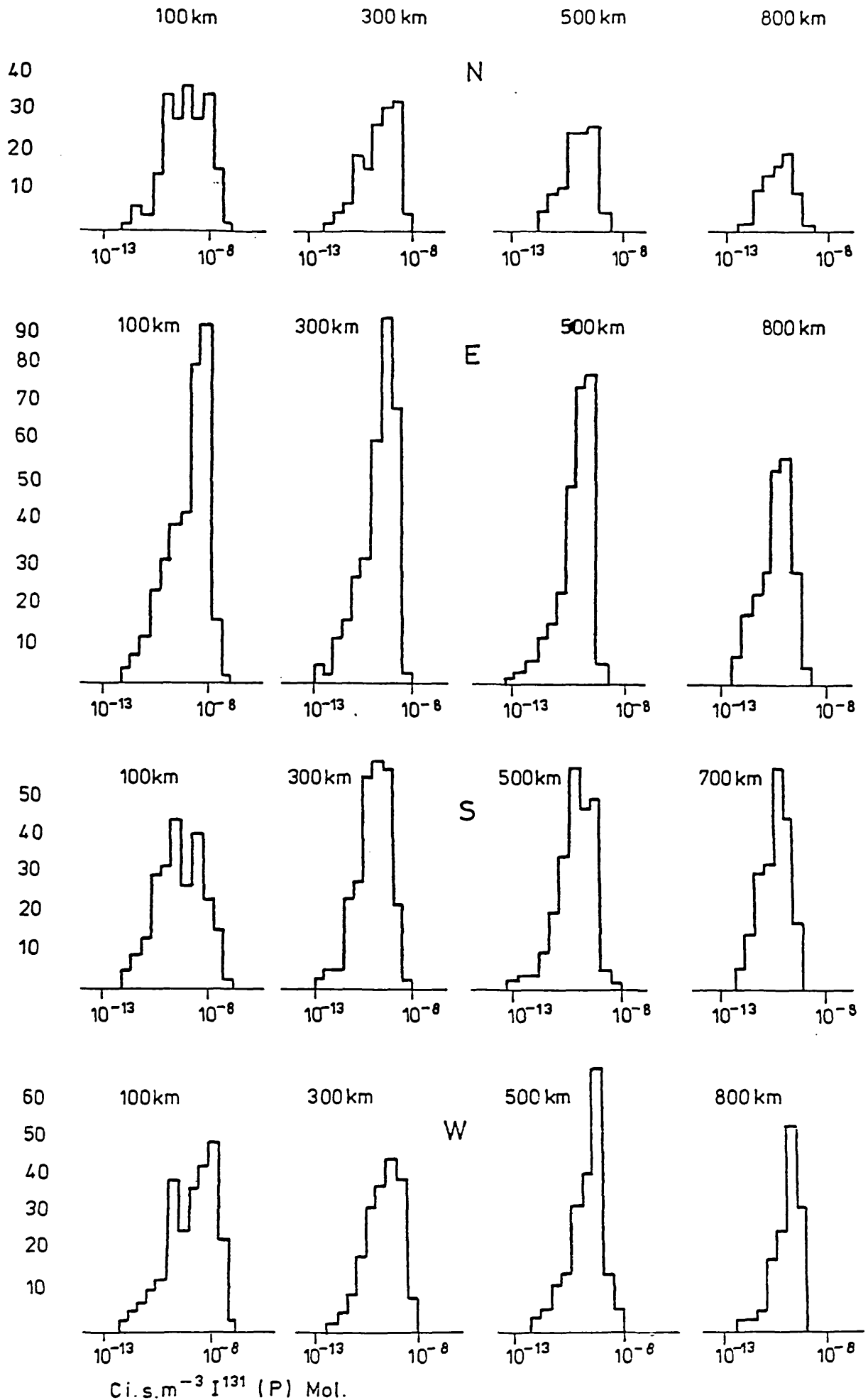


Figure 1.2: Histograms for Time Integrated Air Concentrations due to Three-Hourly Releases of $I^{131}(P)$ from Mol during 1973

Since appropriate data with known releases at the source could not be found, Wrigley(1982) tried to check MESOS against data gathered in Northern Europe shortly after the radioactive discharge of estimated size between 20,000 and 30,000 Curies which accompanied the fire in a reactor at Windscale on the 10th-11th October, 1957. He concluded that:

The application of the MESOS model to the dispersion of I_{131} releases during the Windscale accident has given reasonable agreement with measurements, in view of the uncertainties in the source term and the rather complicated meteorological conditions. MESOS calculations are generally within a factor of three of the measurements and this degree of accuracy is comparable with other long-range dispersion models (Maul,1980).

The uncertainty in trajectories leads to the most significant errors in the dispersion calculations. Trajectories have been shown to be rather unstable because of the movement of the frontal system and the high pressure system during and after the period of I_{131} emissions at Windscale. Under such conditions, trajectories are sensitive to relatively small variations in the backing angle of the advecting boundary layer winds.

The problems associated with predicting boundary layer trajectories could limit the ability of dispersion models to forecast, with confidence, the consequences of an unplanned release incident. However the specification of accurate trajectories is of lesser importance in the development of statistical forecasts for the impact of unplanned releases at remote receptors. In the statistical approach, predicting the frequency of exposure at remote receptors is the main interest, rather than giving accurate forecasts of individual trajectory paths.

Much is known about dynamical meteorology, and hence the dispersion and trajectory calculations for the puffs are not thought to be the main cause of any systematic error in the MESOS database. More uncertainty is introduced because less is known about how nuclides wash out of the air. The physics of droplet formation is complicated and not well understood, yet washout has to be simply parametrized so that calculations can be performed efficiently. Dry deposition is better understood, but uncertainty about the deposition

velocities to be ascribed to nuclides remains. The values of deposition velocity v_d and washout coefficient λ_w used in MESOS are thought to cover their likely ranges, but are not more accurate than that. However the statistical methods described in Chapter 3 allow the effect of varying v_d and λ_w to be estimated.

Despite the uncertainties surrounding individual MESOS trajectories, it remains true that the database is the result of the most comprehensive attempt yet made to study long-range atmospheric dispersion of radioisotopes in Europe. A vast body of useful information is available. The central point of the preceding discussion is this: the attitude users of MESOS results should have is one of bold scepticism. The data are the best yet, and every attempt has been made to check the model; but the results are still only generally right to within a factor of three or so. This thesis has feet of clay - not sand.

Others have used MESOS to investigate various aspects of pollutant transport through the air: Maul(1980) describes a study of sulphur compound transport which uses the MESOS database and integrates the diffusion equations governing the vertical distribution of pollutant in the puff; apart from the Windscale study, Wrigley(1982) used MESOS to obtain the large body of data partly summarized in this thesis; Crompton(1982) compared MESOS and simpler models based only on source meteorology for computing collective doses of radioactivity to the population due to planned and unplanned releases; and Alecio(1984) used it to assess United Kingdom siting policy for nuclear power stations, paying special attention to episodes leading to the highest few exposures.

The central goal of this thesis is to develop a simple probabilistic model of atmospheric dispersion over long distances, based on the MESOS receptor histories; a model easily used by the

non-expert but retaining as far as possible the full scope of possible uses of the database. The model can be used to estimate the long-run expected exposure distribution, arising from a release of duration between three hours and one week, of a nuclide with given half-life, deposition velocity and washout coefficient, at a point between 100 and 1000 kilometres away from a source anywhere in Western Europe. It is embodied in a system of mathematical equations - the result of a statistical analysis of the MESOS data - and is suitable for use in satisfying the Article 37 requirements.

Careful efforts have been made to verify the statistical model by comparing it with the MESOS data. They are, first, assessment of how well the statistical equations fit the data upon which they are based and the importance and physical significance of discrepancies between them; and second, comparison of statistically predicted and MESOS exposure distributions for releases not used in building the model. Together they provide a good idea both of the nature of differences which may arise, and of their probable size and direction. However it is important to realise that such verification is essentially internal: it consists of comparisons between a computer model and a set of equations derived from it.

Throughout this thesis the degree of accuracy of the results obtained and methods used is summarized in the fashion characteristic of statistical work: probability statements, confidence intervals, and significance. However the statistical significance of an apparent effect does not necessarily imply that it is practically important or that it has a physical meaning; and conversely an important physical effect may not be detectable in given data with a high degree of certainty. It is the practical and physical aspects of an analysis that are generally of more interest to the user of statistical results: this is kept in mind throughout. Thus effects

found to be statistically significant have been ignored in the eventual model as practically unimportant or physically meaningless, but others motivated by the physics of pollutant transport but not statistically noticeable have been retained. In particular, one implication of Wrigley's comments quoted above about the likely accuracy of MESOS exposure levels is that statistically highly significant differences between the original and the predicted distributions based on them are not important in practical terms. This is dealt with in Chapter 3. Cox(1982) gives a fuller discussion of these distinctions between statistical and scientific - in his case biological - and practical significance.

Rather strong independence assumptions have been made for most of the statistical fitting, in order both that well-tried methods could be used for much of the work and that a set of equations easily understood by the non-statistician would be its eventual outcome. In particular it has been assumed that time-series of exposures at different receptors are independent, although this is clearly untrue in general. Further it is assumed in Chapter 3 that individual exposures at single receptors are independent and identically distributed. Independence must be regarded as dubious for exposures arising from the same set of weather conditions, and the variation in the weather from month to month casts doubt on the assumption that all the observations have the same marginal distribution. On the other hand exposure levels may vary during a single episode by factors of 10^3 - 10^4 , but seasonal variations are hard to detect, suggesting that such seasonal differences as do occur are uninteresting compared with possible diurnal changes. Figure 1.3 shows the time-series of exposures to Krypton 85 time-integrated air concentration observed at three-hourly intervals 800 kilometres north of Mol, due to unit releases every three hours. Contamination is

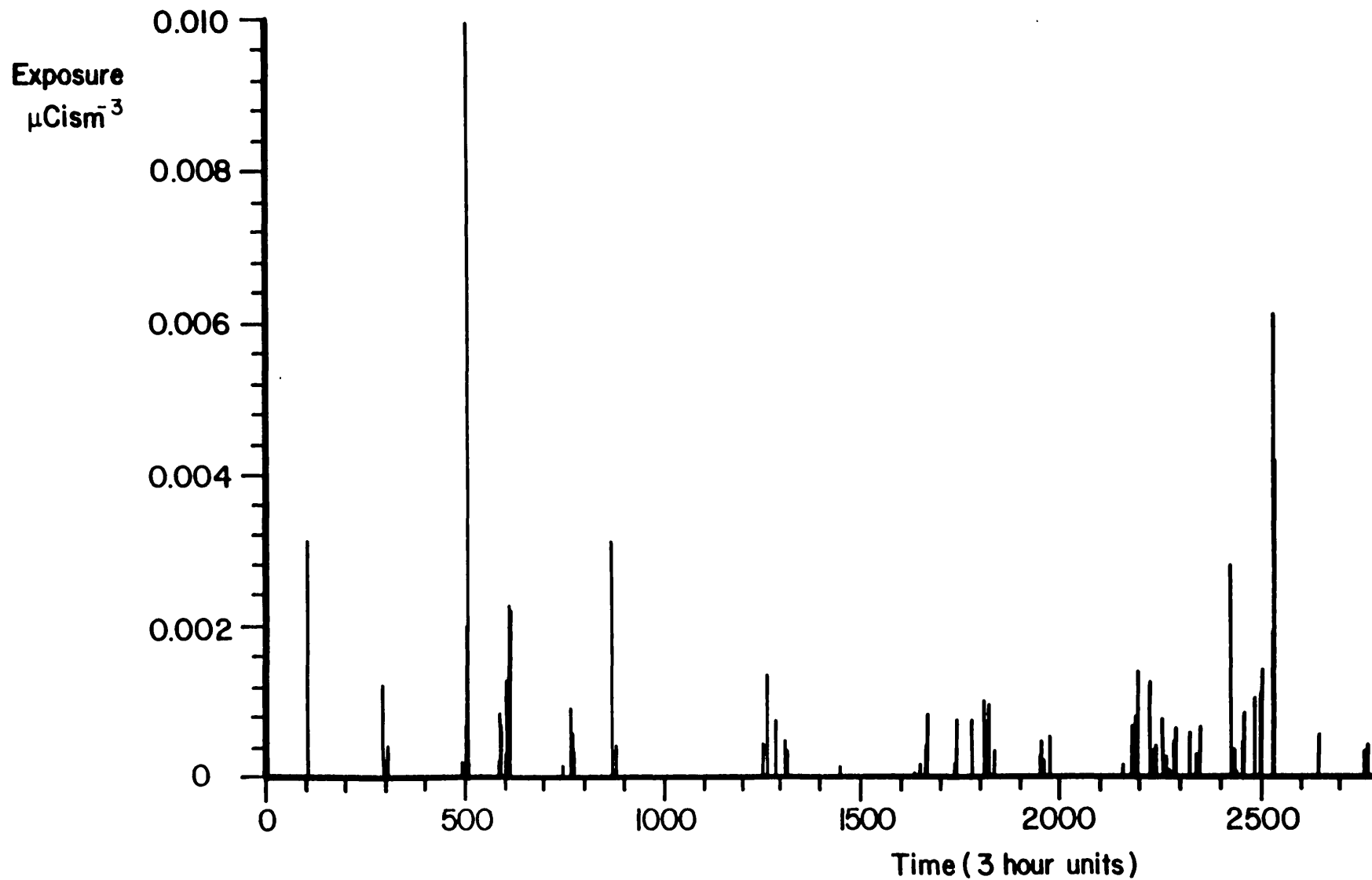


Figure 1.3 : Time-series of exposures to time-integrated air concentration at receptor 800 km north of Mol due to unit releases of Kr_{85} every three hours during 1976.

clearly episodic and there are no obviously seasonal changes in either the levels of pollution or when it occurs.

The overall effect of assuming independence of results for different receptors and of different exposures at single receptors is the over-estimation of the precision of the fitted equations. For this reason the verification of the statistical model against MESOS, limited as it is, is an essential element of this thesis. It gives a clear idea of the likely range of probable errors of the model as compared with MESOS.

Had the aim of the work been to produce a stochastic model for the dynamics of pollution episodes a much more complicated approach would have been necessary. As will be seen, the approach chosen must be regarded as remarkably successful when the complexity of the situation it models is taken into account.

1.2 The MESOS exposure database

Two bodies of weather data exist for input to MESOS. The first covers most of middle and late 1973 and early 1974, a period regarded as 'average' in meteorological terms. This database has a number of fairly short gaps when information adequate for the construction of pressure fields over the whole of Northern Europe was not available, but has been treated as a continuous record for the purpose of analysis. The area covered by the database, from 10° East to 20° West and 62° to 44° North, is shown in Figure 1.4. The sources from which notional releases are deemed to have taken place are Heysham, Karlsruhe, and Mol.

The second weather database covers almost all 1976 except for a single missing day, henceforth ignored. The year had an unusually long hot summer and wet autumn and winter. During the summer there were a number of almost stationary blocking anticyclones for long



Figure 1.4 : Area covered by 1973 meteorological database.

periods, which would have led to high levels of exposure had there been a nuclear discharge to the atmosphere. The 1976 database, which extends southwards to 36° North and hence covers all the mainland of Western Europe, is shown in Figure 1.5. The notional sources for which data through 1976 exist are Mol, Cadarache, and Ispra.

Exposure data for unit discharges of radioactivity over periods of three hours are stored on magnetic tape for imaginary releases from five sources: Mol in both 1973 and 1976, and two others each year. The years chosen are thought adequately to represent the likely range of variation of European weather as it affects long-range dispersion: in particular, levels of contamination in 1976 are thought to be rather higher than those in an 'average' year.

Data exist for each source for twelve combinations of nuclide and exposure mode: time-integrated air concentrations for Iodine 131(p), Iodine 131(g), Caesium 137, Krypton 85, Xenon 133, and Xenon 135; and wet and dry deposition data for the depositing Iodine and Caesium isotopes. The deposition velocities and washout parameters these radionuclides are supposed to have for the MESOS calculations, and their half-lives, are given in Table 1.1. The inert noble gases Kr₈₅, Xe₁₃₃, and Xe₁₃₅ do not deposit, so only air contamination as material passes overhead is important. I₁₃₁ is considered in two forms - particulate and gaseous - with different washout coefficients. When the results are considered together with those for Cs₁₃₇ this allows the effects of decay constant, deposition velocity, and washout coefficient to be estimated.

Dry deposition, wet deposition, and time-integrated air concentrations are calculated separately for the depositing radioisotopes, whereas the air contamination for inert ones essentially depends only upon travel time from the source to the receptor and the vertical cross-section of the puff orthogonal to its

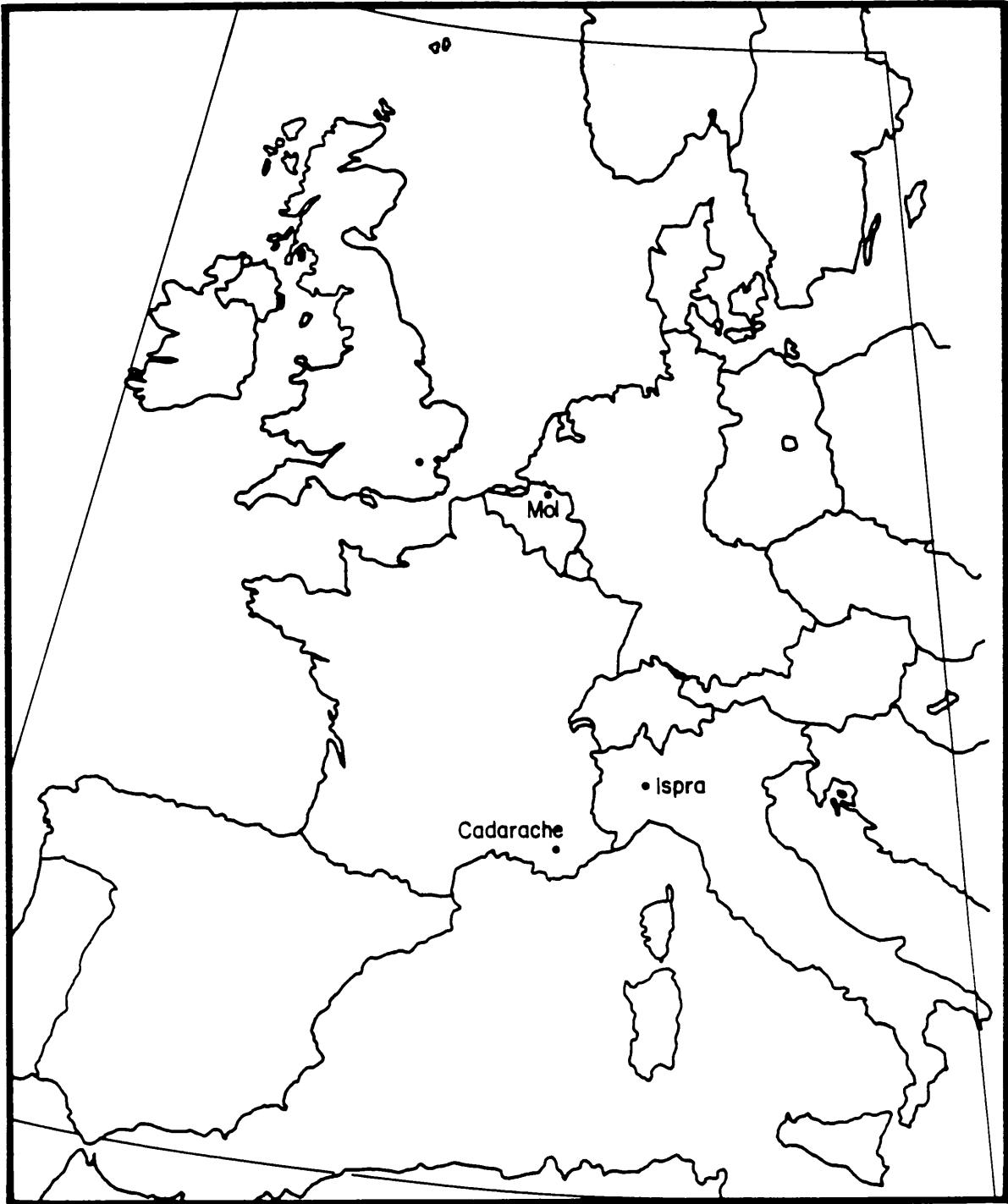


Figure 1.5 : Area covered by 1976 meteorological database.

Nuclide	Half-life	Deposition velocity V_d (ms^{-1})	Washout coefficient* λ_w (s^{-1})
$I_{131}(\text{p})$	8.1 days	3×10^{-3}	$5 \times 10^{-5} J^{0.8}$
$I_{131}(\text{g})$	8.1 days	3×10^{-3}	$1.5 \times 10^{-5} J^{0.8}$
Cs_{137}	28 years	1×10^{-3}	$5 \times 10^{-5} J^{0.8}$
Kr_{85}	10 years	—	—
Xe_{133}	5.3 days	—	—
Xe_{135}	9.1 hours	—	—

Table 1.1 Parameters of nuclides used in data analysis

* J is rainfall rate in mm/hour

trajectory. Thus contamination can be evaluated for any inert nuclide with known decay constant once travel time and cross-section are known. This is later used to create exposure data for 'pseudo-nuclides' - notional isotopes whose decay constants do not correspond to those of actual substances, but which help with estimation problems encountered with wet deposition.

Exposure data are available every three hours only at sixteen receptors for each source. Their locations are shown in Table 1.2. Daily exposure data exist for a large grid of receptors centred at each source - whose arrangement for Mol is displayed in Figure 1.6 - but they have not been used in the analysis.

In addition to the data described above, results for unit exposures for three-hourly releases from Hannover and Stuttgart through 1973 exist at the receptors shown in Table 1.3, for all three exposure modes of the nuclides whose parameters are shown in Table 1.4. These were not used to fit the statistical equations, but retained instead for their verification.

Exposures for unit releases of duration longer than three hours - one day, for example - may be obtained from the three-hourly ones by aggregating them in blocks of eight, and dividing the resulting exposures by eight to get daily contamination levels. This artifice has been used where necessary to get data for release durations of up to one week.

1.3 The structure of this thesis

The level of technical background needed to read different parts of this thesis varies. I hope that much of it is intelligible to anyone with basic knowledge about the boundary layer and a familiarity with statistical notions. However the technical content of Chapters 5-7 is greater; they take for granted some understanding

SOURCE	Heysham	Karlsruhe	Mol	Cadarache	Ispra
COUNTRY	Britain	Germany	Belgium	France	Italy
LATITUDE	54.03°N	49.70°N	51.18°N	43.71°N	45.81°N
LONGITUDE	2.91°W	8.43°E	5.12°E	5.77°E	8.63°E
1973 CALC.	✓	✓	✓	—	—
1976 CALC.	—	—	✓	✓	✓
1	N 100km	N 100km	N 100km	N 100km	N 100km
2	N 300	N 300	N 300	N 300	N 300
3	NE 800	N 500	N 500	N 800	N 800
4	E 100	N 800	N 800	NNE 1100	N 1500
5	E 300	N 1300	E 100	NE 300	E 100
6	E 800	E 100	E 300	NE 700	E 300
7	E 1300	E 300	E 500	E 100	SE 300
8	ESE 1100	E 500	E 800	E 300	SE 500
9	SE 800	S 100	S 100	ESE 800	SE 800
10	SE 1300	S 300	S 300	S 100	S 100
11	SSE 1100	W 100	S 500	WSW 800	S 700
12	S 100	W 300	S 700	W 100	W 100
13	S 300	W 500	W 100	W 300	W 500
14	S 800	W 800	W 300	NW 300	NW 300
15	W 100	NW 800	W 500	NW 700	NW 500
16	W 300	NW 1300	W 800	NNW 1100	NW 1100

Table 1.2 Sources and receptors used in data analysis

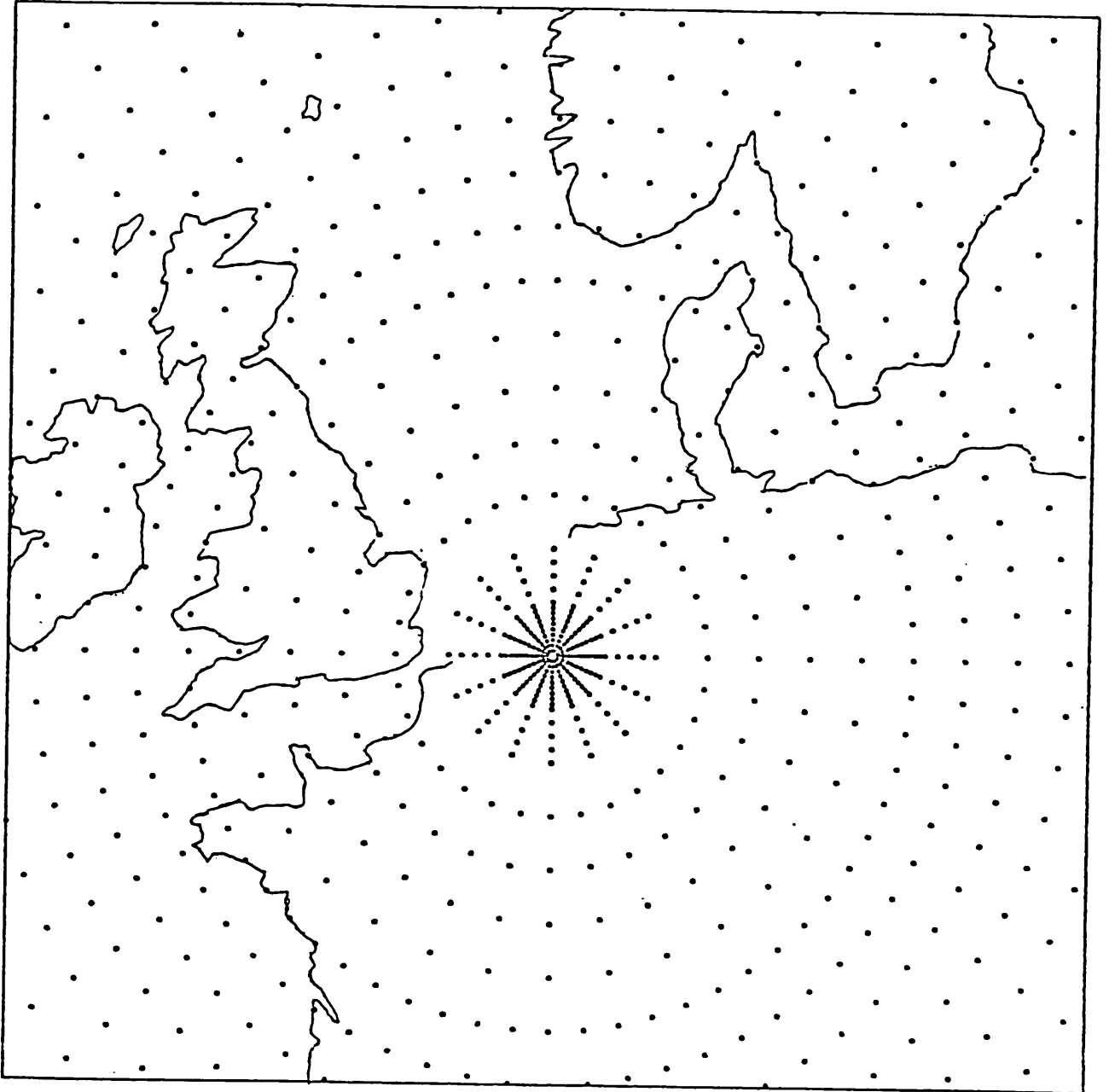


Figure 1.6 : Radial Grid System for Mol

	Stuttgart	Hannover	
SOURCE	Stuttgart	Hannover	
COUNTRY	W. Germany	W. Germany	
LATITUDE	48.47°N	52.23°N	
LONGITUDE	9.12°E	9.44°E	
1973 calc.	✓	✓	
1976 calc.	—	—	
<hr/>			
Relative positions of receptor points	1	NE 100 km	N 100 km
	2	NE 200	N 200
	3	NE 400	N 400
	4	NE 800	N 800
	5	SE 100	E 100
	6	SE 200	E 200
	7	SE 400	E 400
	8	SE 600	E 600
	9	SW 100	S 100
	10	SW 200	S 200
	11	SW 400	S 400
	12	SW 600	S 800
	13	NW 100	W 100
	14	NW 200	W 200
	15	NW 400	W 400
	16	NW 800	W 800

Table 1.3 Sources and receptors used for model verification only

Nuclide	Decay constant λ (s^{-1})	Deposition velocity V_d (ms^{-1})	Washout coefficient* λ_w (s^{-1})
Case 1	0	1×10^{-3}	$2.6 \times 10^{-5} J$
Case 2	0	1×10^{-2}	$2.6 \times 10^{-5} J$
Case 3	2.8×10^{-6}	1×10^{-2}	$1.3 \times 10^{-4} J$
Case 4	0	1×10^{-2}	$1.3 \times 10^{-4} J$

Table 1.4 Parameters of nuclides used for model verification only

* J is rainfall rate in mm/hour.

of the methods of modern statistics. This is unavoidable.

The probability distribution of exposure levels at a remote receptor may be thought of as a combination of the long-run exposure probability for the receptor, and the distribution of contamination levels when it is exposed.

In Chapter 2 the modelling of probabilities of exposure at distant receptors is discussed; and a model for both wet and dry exposure probabilities is proposed, fitted, commented upon, and verified.

In Chapter 3 the complementary problem of finding the distributions of exposure levels at remote receptors is tackled. The Weibull distribution is shown to give a reasonable fit to the data overall; its variation with nuclide characteristics and source-receptor distance is expressed in terms of a simple set of equations motivated by basic physical considerations; and the resulting model is discussed and verified.

Then attention is focussed on the provision of a statistical model for extreme events - very high levels of exposure. Chapter 4 contains a brief review of modelling exceedances over high thresholds, then the next three chapters give the thesis' main contribution to statistical methodology.

Chapter 5 gives an account of and proves some results for the generalized Pareto distribution, an essential ingredient in modelling excesses over high levels; Chapter 6 considers in detail statistical aspects of the distribution - especially how to use it to model complex data; and Chapter 7 goes into diagnostic methods for checking that a fitted model describes the data well.

In Chapter 8 these ideas are applied to the MESOS data. Chapter 9 contains a summary, conclusions, and comments on various aspects of the work. The appendix gives an example of the use of the model.

2. PROBABILITIES OF EXPOSURE

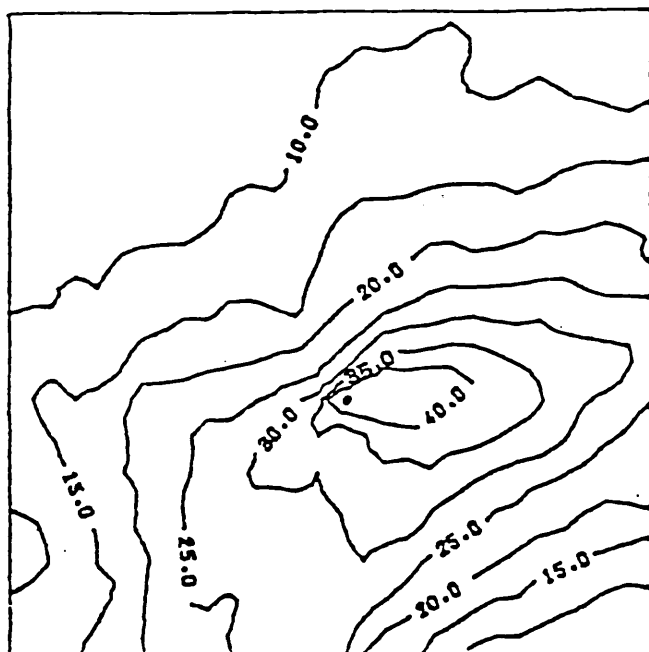
2.1 Introduction

The aim of this chapter is to find a simple method of predicting the probability of exposure to pollution for releases over periods of from three hours to one week at points between 100 and 1000 kilometres from the source of pollutant.

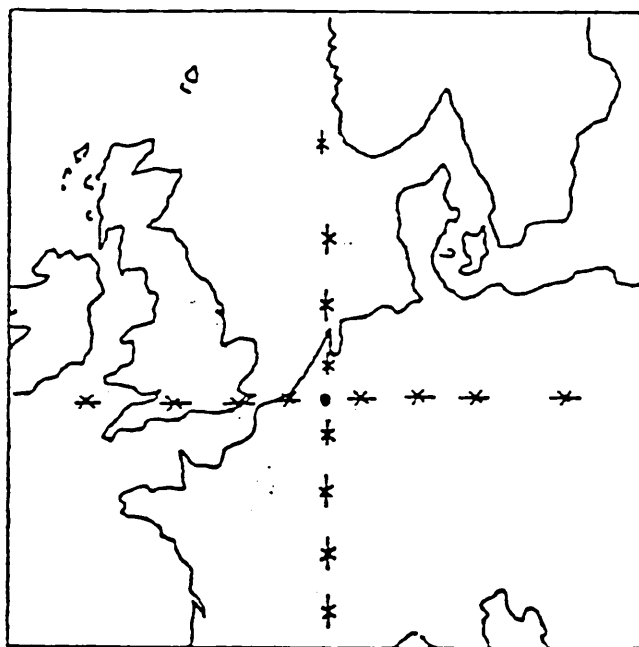
Consider Figure 2.1, from Wrigley(1982), which shows the probability of air contamination and wet deposition due to daily releases from Mol during 1973. The probabilities drop with distance from Mol, and show also a fairly marked pattern apparently depending on the source windrose. Probabilities of exposure to wet deposition seem to depend more on the incidence of rain-bearing air masses at the source. Figure 2.2 shows the proportion of MESOS geostrophic winds greater than 5 m/s directed from 30° sectors at Mol during 1973. The obvious close connection between the figures suggests that it should be possible to relate exposure probabilities to windrose proportions and source-receptor distance.

The trajectory roses in Figure 2.3 show that at short and medium ranges exposure probabilities are likely to be highly correlated with source windroses, but however this need not be true further away. The plot shows the numbers of puffs - out of totals of about 2024 in 1973 and 2880 in 1976 - crossing 10° sectors at various distances from their sources, and rather accentuates anisotropy of the trajectories. At distances of 750 km the effect of trajectory swinging due to obstacles such as the Alps is evident: it is less so at shorter distances. Such effects - to an extent site-specific - are not uniform throughout Europe and it is difficult to allow for them in building a general model for exposure probabilities.

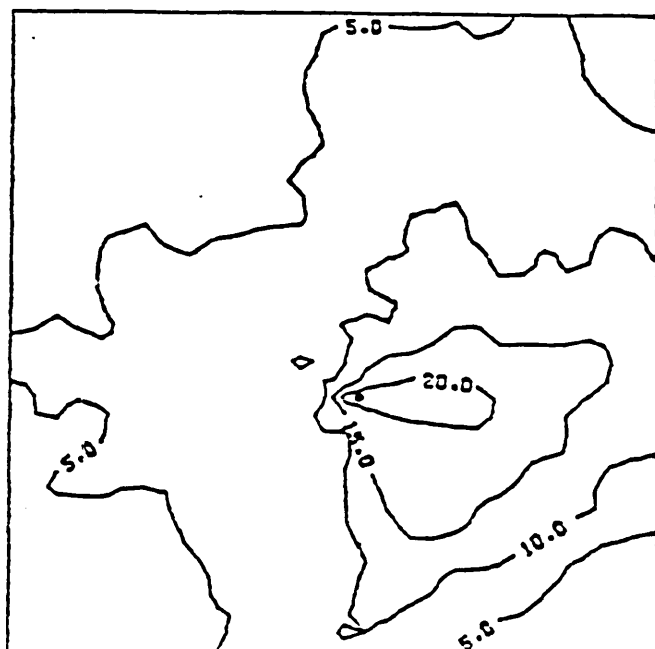
The proportions of geostrophic winds greater than 5 m/s are used



(a) Probability (%) of air contamination

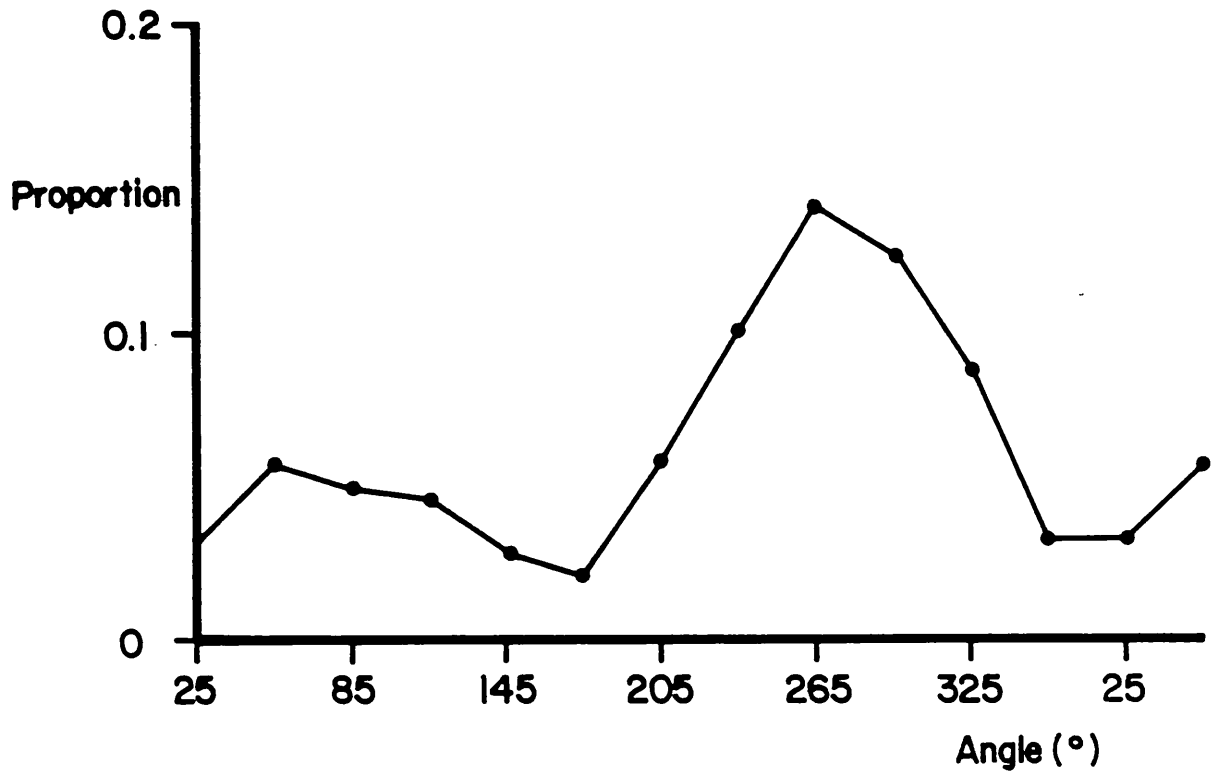


(c) Database map area and Mol receptors

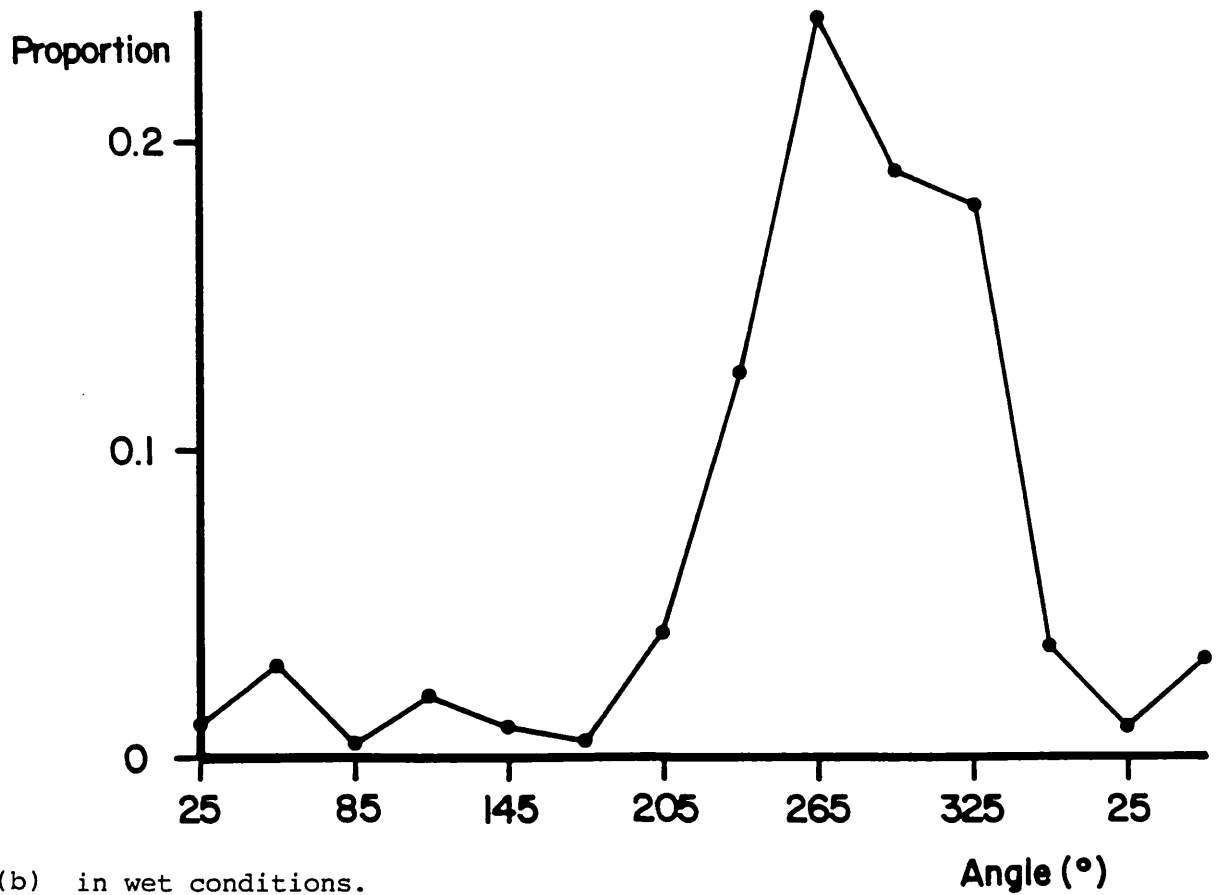


(b) Probability (%) of wet deposition

Figure 2.1 : Probability of air contamination and wet deposition due to daily releases from Mol during 1973.

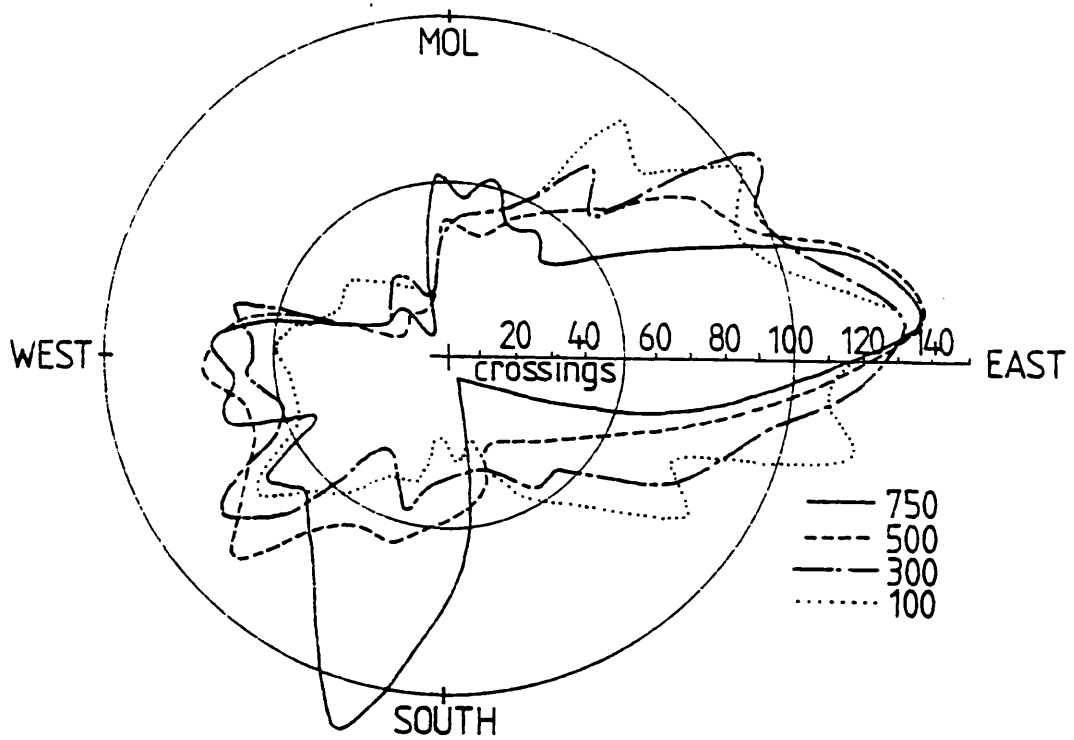


(a) in all conditions.

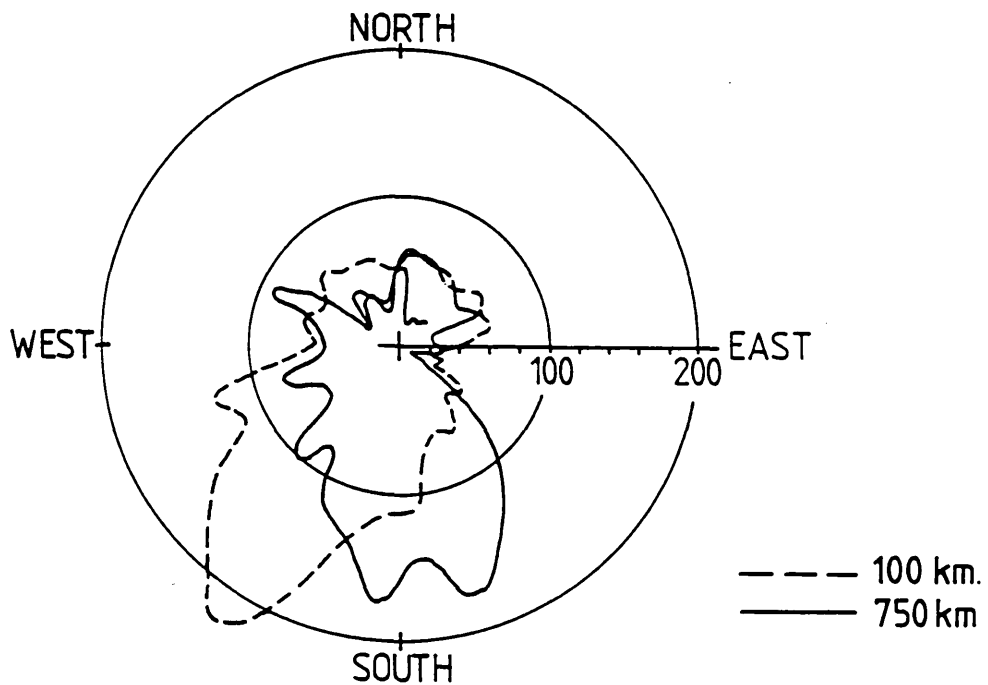


(b) in wet conditions.

Figure 2.2 : Proportion of MESOS geostrophic windroses greater than 5 m/s directed from 30° sectors at Mol during 1973.



(a) releases from Mol during 1973.



(b) releases from Cadarache during 1976.

Figure 2.3 : Trajectory roses at various distances from the source, for releases of duration 3 hours.

throughout this chapter because puffs released when source winds are rather low are likely to meander, so resulting exposures are less highly correlated with the initial direction of the puff. Basing the analysis on these proportions rather than the full windrose results in a slightly better fit of the eventual model for exposure probabilities.

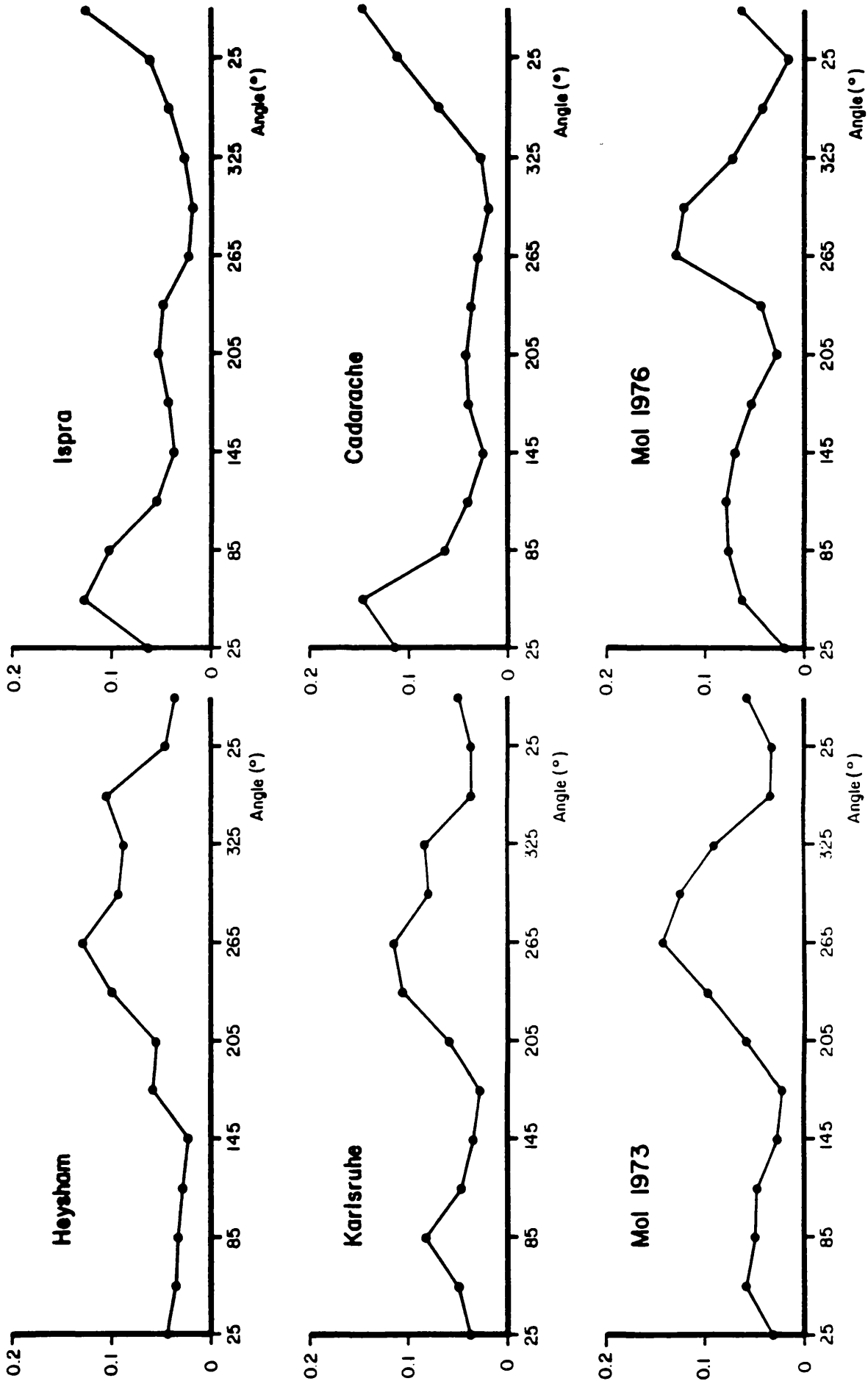
The degree of dependence on distance and source windrose may vary with year - 1973 and 1976 had rather different weather patterns - and geographical features of the source. The sources in Northern Europe - Heysham, Mol and Karlsruhe - generally experience different weather to those in the Mediterranean area - Cadarache and Ispra. Typical windroses differ, suggesting that trajectory characteristics too may differ between the two regions. Figure 2.4(a) shows that most of the higher geostrophic winds at Ispra and Cadarache are from the north-east; whereas those at the other sites tend to be associated with the general south-westerly airflow across Northern Europe, with a secondary peak to the east at Mol and Karlsruhe. The wet windroses in Figure 2.4(b) show the differences even more markedly.

Exposure probabilities depend on release duration. Material released over longer periods is more dispersed because of variation in source and travel conditions, and consequently exposure probabilities increase.

Factors upon which the probabilities may depend are known in statistical terms as covariates or explanatory variables, whereas the probabilities themselves may be termed response variables. The following issues are of main interest:

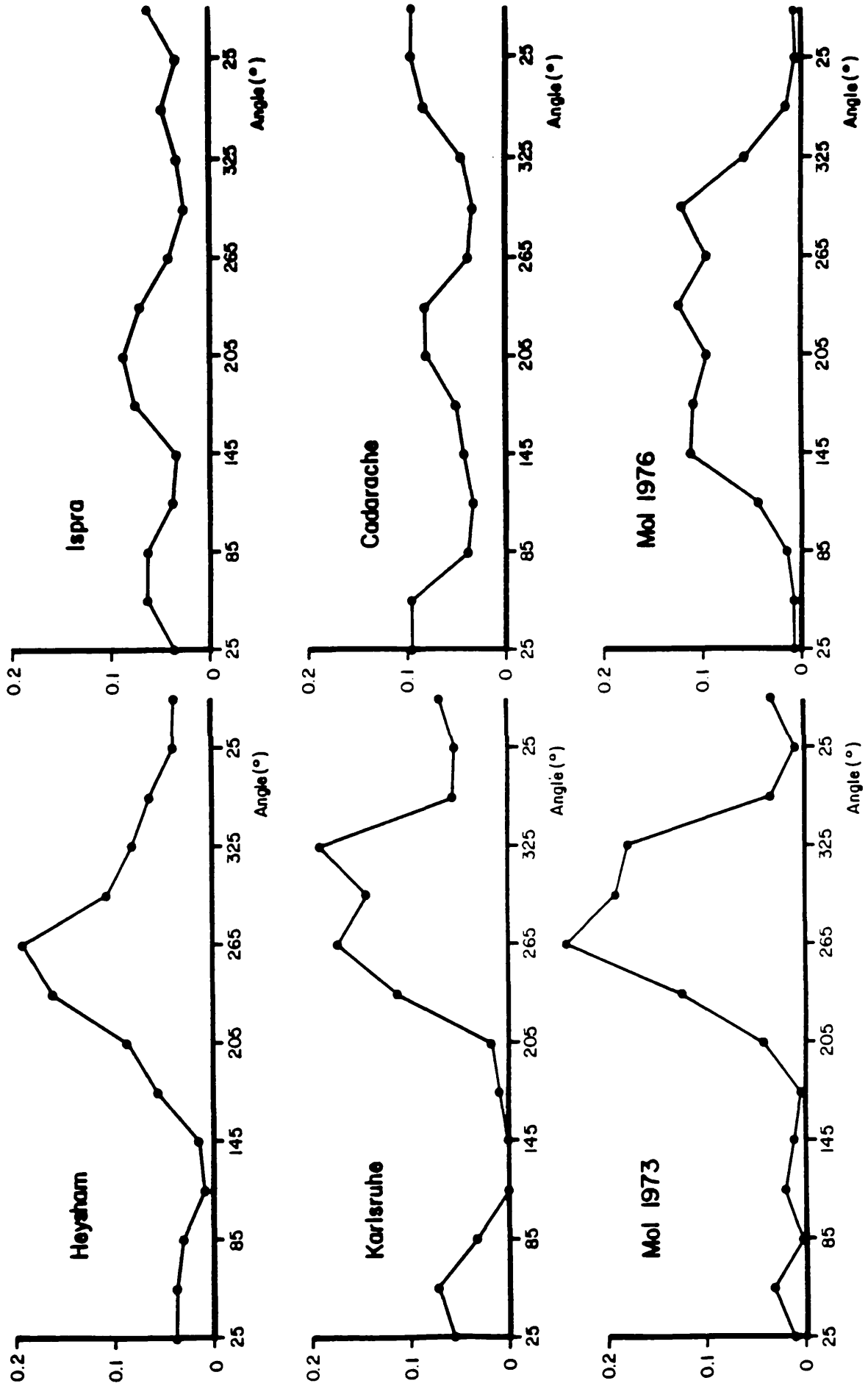
(a) to what extent do the exposure probabilities depend on these factors, and which of them are most important?

(b) can the probabilities be succinctly summarized in terms of



(a) in all weather.

Figure 2.4 : Proportions of MESOS geostrophic windroses greater than 5 m/s directed from 30° sectors.



(b) in wet conditions.

Figure 2.4 : Proportions of MESOS geostrophic windroses greater than 5 m/s directed from 30° sectors.

changes in the covariates?

(c) how may the dependence of the probabilities on the covariates be interpreted in terms of physical processes and effects? These questions are answered in Sections 2.3 and 2.4, after the tools used to investigate them have been described in Section 2.2. Section 2.5 discusses windrose data needed to use the model developed in the previous two sections, which is then verified against Hannover and Stuttgart data in Section 2.6.

2.2 Regression analysis: tools, diagnostics, and transformations

The material in this section is included to make the thesis fairly self-contained. It is germane to much of the rest of this chapter and the next, and underpins the material in Chapters 6, 7, and 8. Familiarity with the basic idea of least squares regression is assumed.

A multiple regression model is one in which the n uncorrelated random quantities Y_i with common variance σ^2 are thought to depend linearly on the fixed vectors x_i of dimension p ; Y_i has expected or mean value

$$EY_i = x_i^T \beta = \sum x_{ij} \beta_j.$$

Here β is a p -vector of unknown parameters β_j to be estimated from the observed values y_i of the Y_i , which control the extent to which they depend on the x_i . For n such observations y_i and their covariate vectors x_i the least squares estimate $\hat{\beta}$ of β is

$$\hat{\beta} = (X^T X)^{-1} X^T y,$$

where X is the $n \times p$ matrix $\{x_{ij}\}$ of rank p and y is the $n \times 1$ vector $\{y_i\}$. The estimate $\hat{\beta}$ minimizes the sum of squared differences

$$\sum_{i=1}^n (y_i - x_i^T \beta)^2,$$

hence its name. The $p \times p$ matrix $X^T X$ will have an inverse if the observed data are adequate to estimate all the parameters β_j , which is henceforth assumed. The number of 'independent' observations - in a sense - remaining in the data when p parameters are estimated from n original observations is the number $n-p$ of 'degrees of freedom' of the model. This does not include allowance for estimating σ^2 .

Two regression equations are said to be nested if one reduces to the other by restricting some parameter values or their combinations. The model $EY_i = \beta_1 x_{i1}$ is nested within the model $EY_i = \beta_0 + \beta_1 x_{i1}$ since the first is obtained by restricting β_0 to be zero in the second. Extending a model equation by including extra parameters will reduce the residual sum of squares

$$\sum_{i=1}^n (y_i - x_i \hat{\beta})^2$$

since more parameters may vary: this will improve the fit of the model to the data. However the improvement in fit may only be small, so some check is needed to see if the extra parameters are worth including. This check is provided by the so-called 'F-test' which compares the residual sums of squares for nested models. The basic idea is to see if the increase in goodness of fit - measured by the difference between residual sums of squares - could have happened by chance were the simpler of the nested models true. For example, suppose that under Model A, $EY_i = \beta_0$; whereas under Model B, $EY_i = \beta_0 + x_{i1} \beta_1$. The estimates β_A and β_B are found, and it is of interest to determine if $\beta_1 = 0$; i.e. if the observations y_i depend on the x_{i1} . The reduction in sum of squares by fitting the extra parameter β_1 is

$$\sum_{i=1}^n (y_i - \beta_{0A})^2 - \sum_{i=1}^n (y_i - \beta_{0B} - x_{i1} \beta_{1B})^2,$$

which is $ss_A - ss_B$, say. This is likely to be large if the y_i do

indeed depend on the x_i , and to be small if not, suggesting that tests to see if β_B is non-zero be based upon it. If Model A is correct then the so-called F-statistic

$$\frac{(ss_A - ss_B)/q}{ss_B / (n - p_B)}$$

has the $F_{q, n-p_B}$ -distribution, which depends on q and $n-p_B$. Here q is the difference in the numbers of parameters in the two models and p_B is the number of parameters in Model B; so in this example $q=1$ and $p_B=2$. These statistics are used to compare the fit of different regression equations, and are helpful in the search for equations which fit data well with as few parameters as possible.

When an equation has been selected as good by comparison of F-statistics, it is then usual to assess how well the model fits the data by seeing how close the fitted and observed values of the y_i are. The fitted - or predicted - value of Y_i given x_i and $\hat{\beta}$ is $\hat{y}_i = x_i^T \hat{\beta}$, which has variance

$$\text{var}(\hat{y}_i) = \sigma^2 x_i^T (X^T X)^{-1} x_i = \sigma^2 v_i,$$

say. The differences $y_i - \hat{y}_i$, the unstandardized residuals, are uncorrelated with the y_i and may be used to check the fit of the regression equation. They may be plotted against the fitted values \hat{y}_i or the x_{ij} for different values of j to reveal the extent to which the data support assumptions made for the purpose of the analysis.

However the variance of $y_i - \hat{y}_i$ is $\sigma^2(1-v_i)$, which is not constant; this may lead to difficulties in interpreting the plots. It is usually better to plot the standardized residuals

$$r_i' = \frac{y_i - \hat{y}_i}{s\sqrt{(1-v_i)}}$$

instead. Here s^2 is the residual mean squared error estimate of σ^2 :

i.e. $s^2 = (n-p)^{-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2$. The r_i ' all have mean zero and variance unity when the assumptions made for the analysis are valid. If in addition - as is commonly assumed - each Y_i is Normally distributed with mean $x_i^T \beta$ and variance σ^2 , a plot of the ordered r_i ' against Normal order statistics should be a straight line of unit slope and zero intercept. This helps to detect cases where Normal assumptions are invalid.

There may be observations y_i which do not accord with the posited linear regression - outliers. Such aberrant data may be detected by considering the cross-validatory or jackknife residuals

$$r_i^* = \frac{y_i - \hat{y}_i}{s_{(i)} \sqrt{(1-v_i)}} ,$$

where $s_{(i)}^2$ is the estimate of σ^2 based on the regression equation but with the i^{th} datum y_i omitted. Large values of r_i^* may indicate errors of measurement or recording, or more importantly in many applications may identify values or areas of values of x_i where the postulated equation is a poor approximation to the data.

In addition to outlying values of the observations y_i , there may be outlying or atypical values of the explanatory variables. The vectors of covariates x_i may be so arranged - by chance or design - that their corresponding y_i have more or less influence on the regression equation. An example is the case $EY_i = \beta_0 + \beta_1 x_i$, where $\sigma^2=1$, $n=10$, and $x_i=0$ for $i=1, \dots, 9$ and $x_{10}=10$. The observed value y_{10} of Y_{10} contains almost all the sample information about the true value of β_1 , which cannot be estimated if Y_{10} is unobserved for any reason. Contrast this with the case where $x_i=1$ for each i , which is plainly much more balanced: the equation will be little disturbed if any single y_i is omitted. The effect on the parameter estimate of deleting the i^{th} datum may be seen by considering the statistics

$$T_i = \text{sgn}(r_i^*) \sqrt{\left\{ \frac{(n-p)}{p} \frac{v_i}{(1-v_i)} \right\} |r_i^*|},$$

which will be large in magnitude for observations whose influence on the fitted equation is big.

The r_i^* and the T_i may be plotted in the same way as the r_i' to find unusual observations y_i or design points x_i respectively.

In many circumstances it will be natural to model the y_i as linear functions of the x_i , particularly if physical or dimensional considerations so dictate. However in other cases, for example if on dimensional grounds it seems likely that a relationship

$$y_i \propto x_{i1}^{\beta_1} x_{i2}^{\beta_2} x_{i3}^{\beta_3}$$

holds, then a transformation of the y_i - in this case logarithmic - is strongly suggested. Such a transformation may render the regression more nearly linear; or may make the error distribution more symmetric; or may stabilize the variance of the y_i ; or may remove from the model equation interactions between different covariates to give a more plausible and parsimonious regression; or it may do all these at once. A power-law transformation may be sought on empirical grounds by following the procedure due to Box and Cox(1964). They suggested that the regression equation be applied to

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda & (\lambda \neq 0) \\ \log(y) & (\lambda = 0) \end{cases}$$

instead of y , and the value of λ be chosen which made the data most nearly Normally distributed with constant variance under the assumed regression model. They quantified this by choosing the value $\hat{\lambda}$ of λ which maximized the joint probability density - the 'likelihood' - of the original data. This choice of λ often has the desirable properties outlined above.

The books by Seber(1977) and Draper and Smith(1981) give recent accounts of mathematical and statistical aspects of linear regression analysis. The outline above is based on part of Atkinson(1982). Cook and Weisberg(1982) and Atkinson(1985) give full accounts of the topics discussed. Many computer programs are available for such analysis. I used GLIM (Baker and Nelder, 1978).

2.3 Exposure probabilities for releases of duration three hours

The exposure probabilities depend on source-receptor distance and source windrose. A form such dependence might take is suggested by the following argument.

According to the statistical theory of diffusion, with the idealized mathematical assumption of uniform diffusivity, the lateral spread of a plume of material at a receptor downwind of a point source is proportional to the square root of the source-receptor distance. In practice simple models of Gaussian plume type assume a power-law relationship - usually established empirically - between lateral spread and distance. Suppose then that the lateral spread of a plume at a distance d from the source is αd^γ . Then the plume subtends an angle $\alpha d^\gamma / 2\pi d = \alpha d^{\gamma-1} / 2\pi$ at the source. Typical values of γ would be between one-half and one. If the plume is straight it will expose a receptor at distance d and angle θ only if the wind is directed into the interval $\theta \pm \phi$, where $\phi = \alpha d^{\gamma-1} / 2\pi$. Thus the probability that the receptor will be exposed is roughly

$$\int_{\theta-\phi}^{\theta+\phi} p(u) du,$$

where $p(u)$ is the probability that the source wind is directed in the direction u . This is approximately $2\phi p(\theta)$, provided the source windrose is fairly smooth. Or

$$\log\{\text{exposure probability}\} \approx \log\{\alpha/2\pi\} + \log\{p(\theta)\} + (\gamma-1)\log\{d\},$$

suggesting that the model

$$\log\{\text{exposure probability}\} = \beta_0 + \beta_1\log\{p(\theta)\} + \beta_2\log\{d\}$$

be fitted to the data. Here the parameters β are to be estimated from the data, and $p(\theta)$ is the probability that the source windrose is directed towards the receptor. The inclusion of the additional parameter β_1 is prompted by the consideration that far away from the source the source windrose is unlikely to have the strong influence assumed above: β_1 is expected to be between zero - when the source windrose would have no effect on the exposure probabilities - and one - were the simplistic argument above valid. As mentioned above, the windrose $p(\cdot)$ is restricted to those occasions when the windspeed is 5 m/s or more, in order to exclude situations in which the puff is unlikely to have a straight-line trajectory due to starting in slack conditions.

The model

$$\log\{\text{exposure probability}\} = \beta_0 + \beta_1\log\{p(\theta)\} + \beta_2\log\{d\} + \varepsilon,$$

where the ε 's are uncorrelated with common unknown variance, was fitted to the 96 exposure probabilities corresponding to releases of three-hour duration from the sources in Table 1.2. Model fit as measured in terms of sums of squares was good, but inspection of the residuals revealed that the best such equation tended to underestimate exposure probabilities at distances up to about 500 km, and to overestimate them at longer distances. The estimate of β_1 was about 0.36, and values of $\hat{\beta}_2$ were about -0.35 for northerly and -0.61 for Mediterranean sources - which correspond to values of γ about 0.65 and 0.39 respectively. These are rather low values of γ for spread of plume width with distance over this range: the work of

Doury(1976) suggests that in moderate winds γ is one or so for travel times of up to a few days and somewhat less thereafter, depending on the scale of the turbulence the plume experiences. The pattern of the residuals confirms this, and thus it is overoptimistic to suppose that a common γ at all ranges will adequately fit the data. This could be overcome by fitting to the probabilities different values of γ over different ranges of d , but this has the drawback that estimation at long ranges - where data are scarcer and more likely to be affected by long-range effects such as trajectory turning - is likely to be inaccurate; furthermore it increases the complexity of the model - which, using Occam's razor, is in principle undesirable unless a more complicated model is likely to account for the data appreciably better. Another possibility is to seek a simple function of d which initially grows more quickly than $d^{0.6}$ but whose growth slows down as d increases.

A function appropriate on those grounds is $\text{dexp}\{-\epsilon d\}$ for a small positive value of ϵ . It grows a little more slowly than d initially, reaches a maximum at $d = 1/\epsilon$, and then slowly decreases - which is hard to envisage happening to a plume. However, over the range of d and ϵ encountered here, this turns out to be a good approximation to the effect of distance d on exposure probabilities, despite this difficulty in its physical interpretation. It suggests that the equation

$$\log\{\text{exposure probability}\} = \beta_0 + \beta_1 \log\{p(\theta)\} + \beta_2 d \quad \dots 2.3.1$$

be fitted to the data. When this relationship is fitted by ordinary least squares, residual plots for the model equation indicate that there are three outlying observations: at the receptors 1300 km east of Heysham and north-west of Karlsruhe, and 700 km south of Ispra. Their probabilities are respectively too high, too low,

and too low to be accounted for by the equation. The receptors are all quite far from their sources and close to the edge of the map, so it is not unreasonable to exclude them from further analysis. The high value east of Heysham may be due to the convergence of easterly-bound trajectories mentioned by Smith(1980); and the low value north-west of Karlsruhe due to air masses from the south-west diverting trajectories travelling long distances to the north and north-west from the Continent.

Further evidence for these to be regarded as aberrant is found when the Box-Cox procedure discussed in Section 2.2 is applied to the original probabilities of exposure using this model equation. The parameter λ estimated from the entire set of probabilities takes the rather unlikely value $\hat{\lambda} \approx 0.5$. This implies that

$$\text{exposure probability} = \{ \beta_0 + \beta_1 \log\{p(\theta)\} + \beta_2 d \}^2,$$

which is not a very plausible relationship: rather than ultimately increase, exposure probabilities ought to die away for low values of $p(\theta)$ and long distances d . However when the three outliers are left out the situation resolves itself and $\hat{\lambda} \approx 0$, which implies that the logarithmic scale is indeed appropriate for the data and regression. This illustrates the fact that the choice of a transformation may be strongly influenced by only a few observations and unsupported by the bulk of the data. Figure 2.5 shows plots of the maximized log probability density for fixed values of λ , for both the full and reduced datasets. Especially in the light of the reasoning which began this section, both curves rule out the use of the untransformed data - regression of the probabilities themselves on the covariates could lead to prediction of negative probabilities. The three-hour exposure probabilities are quite small - in the range 0-0.25 or thereabouts - and the prediction of inadmissible values would be

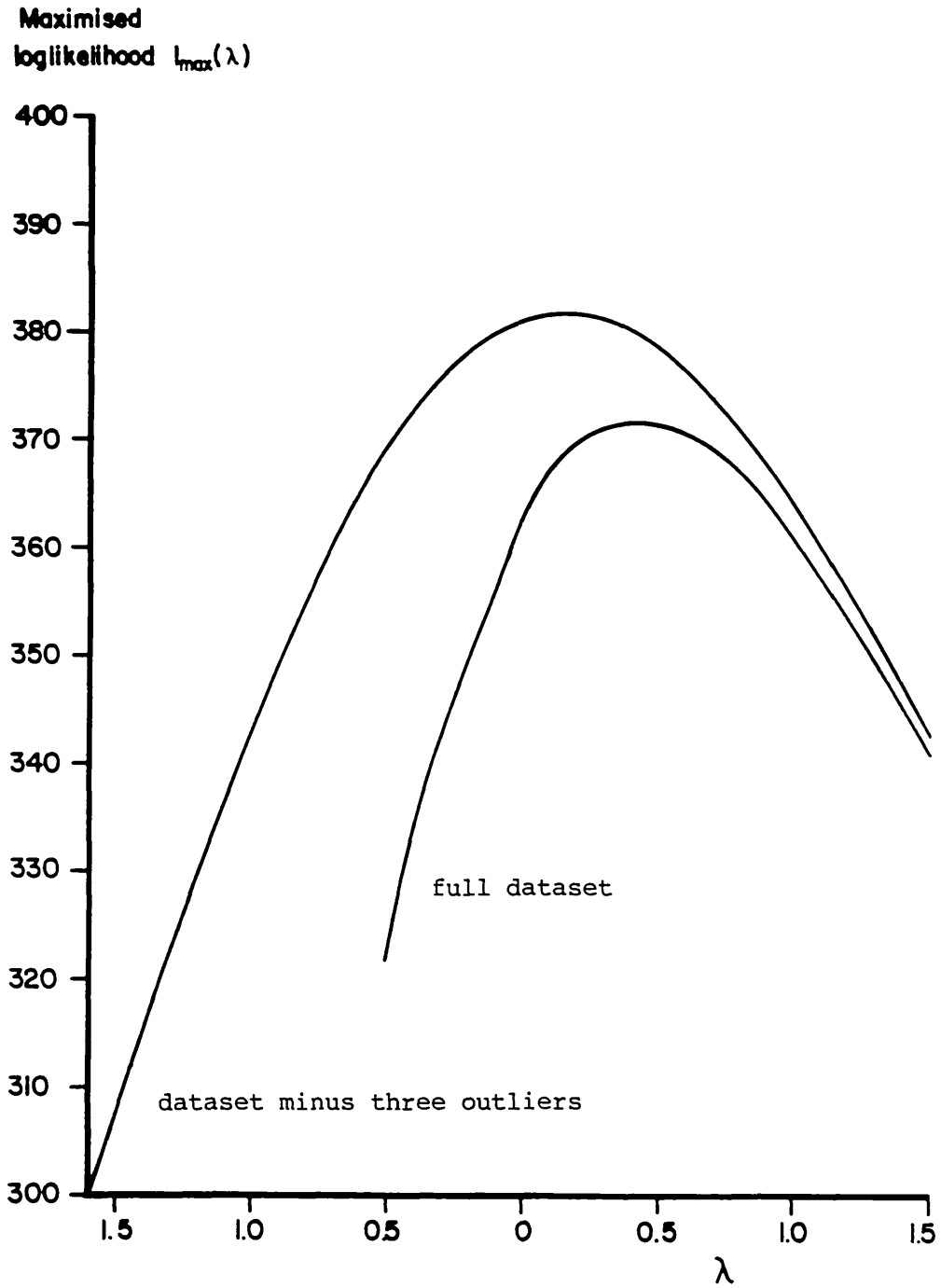


Figure 2.5 : Partially maximized loglikelihood $l_{\max}(\lambda)$ for Box-Cox power transformation of probabilities of exposure to air contamination.

certain were the regression equation based on the original data.

More complex models than 2.3.1 above may be fitted to the data. For example, a model equation with different values of β_0 at the five different sources, different values of β_1 in the two years, and two different values of β_2 at Mediterranean and other sources may be written in an obvious shorthand as

$$\text{SOUR} + \text{YEAR.P} + \text{MED. DIST},$$

whereas the model equation at 2.3.1 is simply

$$\text{P} + \text{DIST}.$$

Model 2.3.1 has 90 degrees of freedom; the more complex one has $84 = 93$ (96 observations - 3 outliers) - 5 (different values of β_0 at each source) - 2 (1973 and 1973 values of β_1) - 2 (Mediterranean and non-Mediterranean values of β_2).

Numerous models were fitted to the log-exposure probabilities. Their nesting structure, including their residual sums of squares and associated degrees of freedom, is shown in Figure 2.6. An arrow from one box to another denotes nesting: it points to the box representing the simpler model. The best-fitting model is $\text{SOUR} + \text{P} + \text{SOUR.DIST}$, with residual sum of squares 2.697 on 82 degrees of freedom. The F-test for comparing this model with the basic model 2.3.1 is

$$\frac{(4.012 - 2.697)/8}{2.697/82} = 4.998,$$

indicating that the model is a significant improvement over 2.3.1.

The corresponding statistic which compares $\text{MED} + \text{WIND} + \text{MED.DIST}$ with 2.3.1 is 16.46, a dramatic improvement over the simpler model. However the residual sum of squares is not further significantly reduced by allowing different values of β_0 and β_2 at each source.

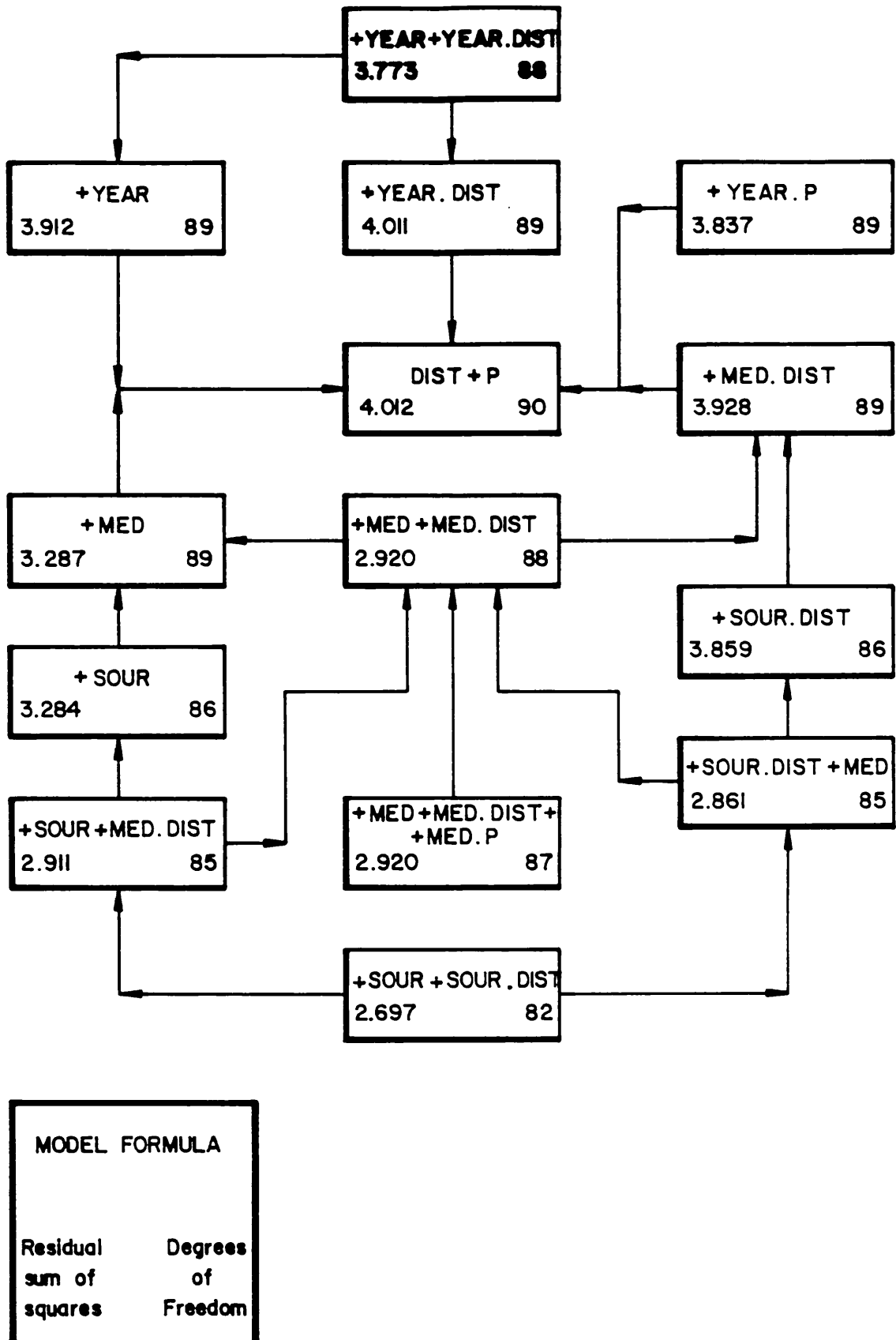


Figure 2.6 : Nesting structure of regression models fitted to log-probabilities of exposure to air contamination.

The practical import of Figure 2.6 is that the linear regression model of this type which best represents the data is the model MED + P + MED.DIST. This allows different means β_0 and effects of distance β_2 at Mediterranean and non-Mediterranean sources, but fits the same effect β_1 of source windrose at all five sources. Allowing parameters to vary from year to year or source to source gives no real improvement in fit. The windroses for Mol in 1976 and 1973 are not identical but are fairly similar, giving grounds for a belief that yearly differences at sources are embodied in their windroses.

The regression equation for Mol, Karlsruhe, and Heysham is

$$\log\{\text{exposure probability}\} = 0.284\log\{p(\theta)\} - 1.12 - 0.110d \quad \dots 2.3.2,$$

(0.0349) (0.108) (7.24×10⁻³)

where the standard errors of the parameter estimates are below the estimates themselves. The equation for Ispra and Cadarache is

$$\log\{\text{exposure probability}\} = 0.284\log\{p(\theta)\} - 0.739 - 0.148d \quad \dots 2.3.3.$$

(0.0349) (0.123) (8.85×10⁻³)

The estimated value of σ is 0.1822. Details of the regression equations are displayed in Table 2.1.

For the Mediterranean sources the overall mean is higher than for the others, but the probabilities fall more rapidly with distance. For a given value of $p(\theta)$, exposure probabilities at distances up to about 700 km for releases from sources in the Mediterranean area are higher than those from other sources, but they are lower thereafter. The value 0.284 of $\hat{\beta}_1$ suggests that source geostrophic windrose $p(\theta)$ is connected to exposure probabilities by roughly a cube- or fourth-root law, consistent with the argument above that β_1 should lie between zero and one. The value $\hat{\beta}_1=0.3$ expresses the effect of synoptic divergence and turning of the trajectories over the distances of hundreds of kilometres of interest

Parameter	<u>North</u>		Parameter	<u>Mediterranean</u>	
	Estimate	s.e.		Estimate	s.e.
β_0	-1.12	0.108	β_0	-0.739	0.123
β_1	0.284	0.0349	β_1	0.284	0.349
β_2	-0.110	0.00724	β_2	-0.148	0.00885

Correlation matrices of estimates

	<u>North</u>				<u>Mediterranean</u>		
	β_0	β_1	β_2		β_0	β_1	β_2
β_0	1.0	0.926	-0.367	β_0	1.0	0.898	-0.313
β_1		1.0	-0.0642	β_1		1.0	0.0439
β_2			1.0	β_2			1.0

Estimate of σ is $S = 0.1822$

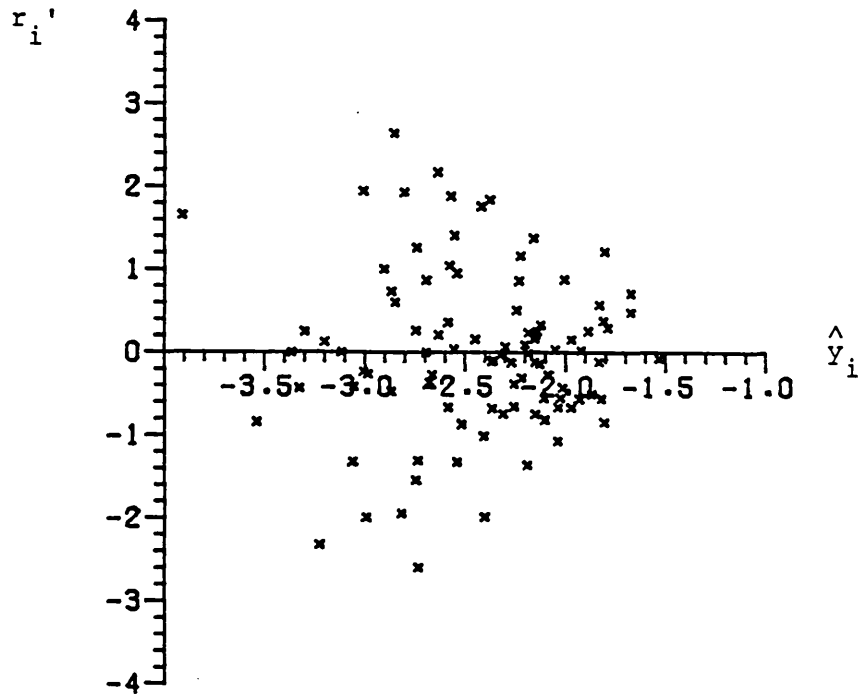
Table 2.1 Details of regression equations for probabilities of exposure in all conditions due to releases of duration three hours.

here. Over shorter distances use of similar models could be expected to produce values of $\hat{\beta}_1$ closer to but still less than one. The values of the estimates $\hat{\beta}_1$ suggest that rapid broadening of the puffs stops at distances of about 900 km and 700 km from northerly and Mediterranean sources respectively; which may be interpreted to imply that relatively less large-scale turbulence affects releases in the Mediterranean region than in northern Europe.

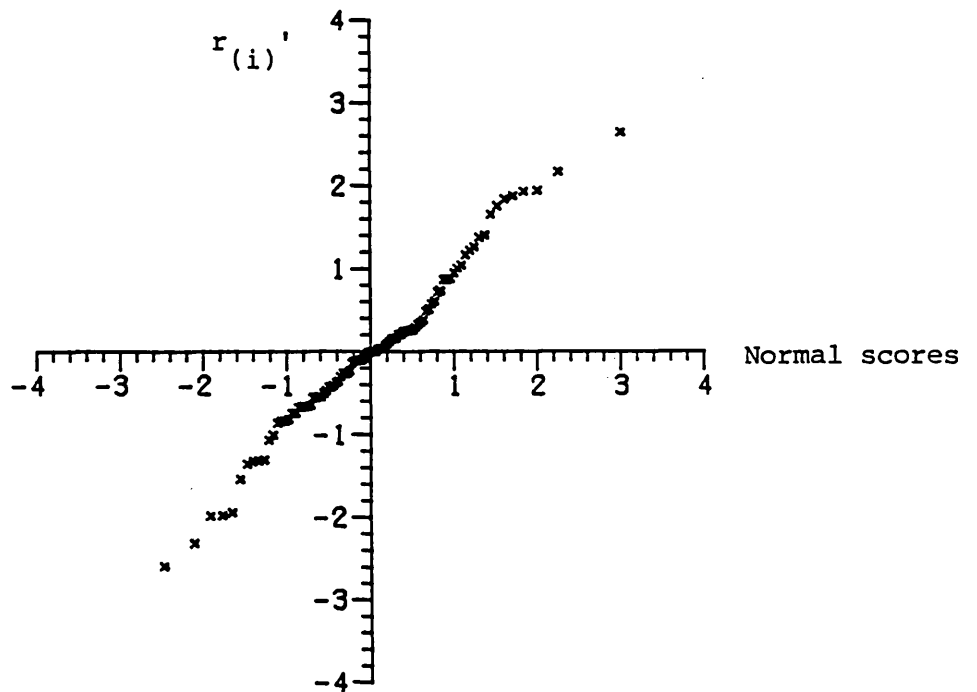
The plot of standardized residuals r_i' against fitted values \hat{y}_i in Figure 2.7(a) shows a slight departure from randomness insofar as the plot tapers off to the right and indicates that the spread of the log-probabilities decreases as they themselves increase.

Qualitatively this means that confidence intervals for probabilities predicted from the regression equation will be a little too wide for higher predicted values and too narrow for lower ones. Since the effect is small and any improvement would not be big, it is not worth adjusting the model to take it into account. Inspection of the numerical values of the r_i' reveals small systematic but not statistically significant departures from the fitted equations. Any attempt to take account of them would lead to a more complex model with no real gain to the user. In fact the model fits five parameters to 93 observations to soak up over 80% of the absolute variation in the data.

The plot of the ordered r_i' against Normal order statistics in Figure 2.7(b) is close to a straight line and shows that the residuals are roughly Normal. The plot of the statistics T_i in Figure 2.7(c) shows two rather anomalous points, which correspond to the receptors 1500 km north of Ispra and 1300 km south-east of Heysham. Dropping them from the data and refitting the regression model has little effect on the estimated parameters. Note that they belong to receptors far from their respective sources and out of the

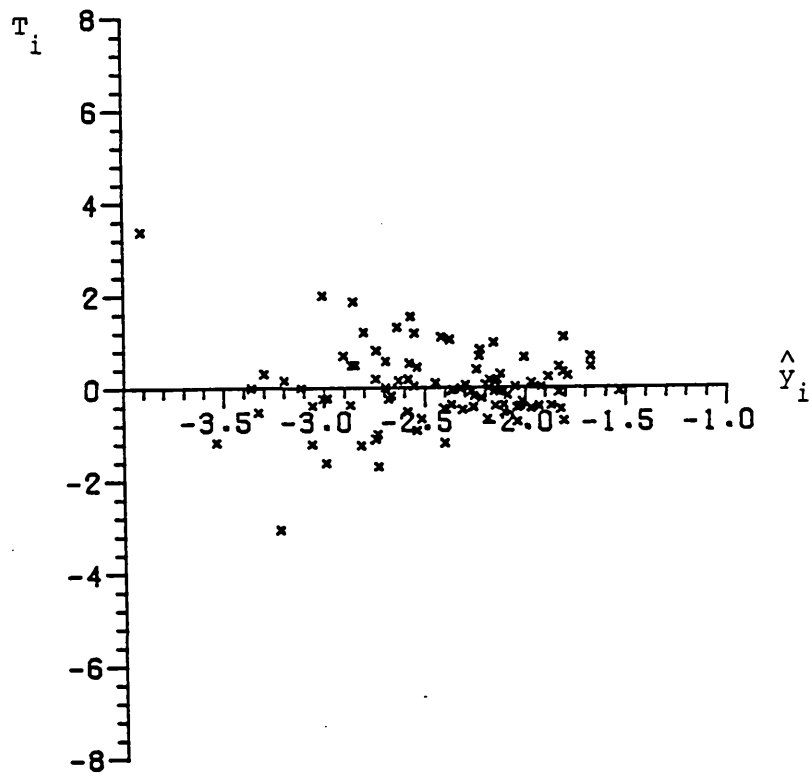


(a) standardized residuals r_i' vs. fitted values \hat{y}_i .

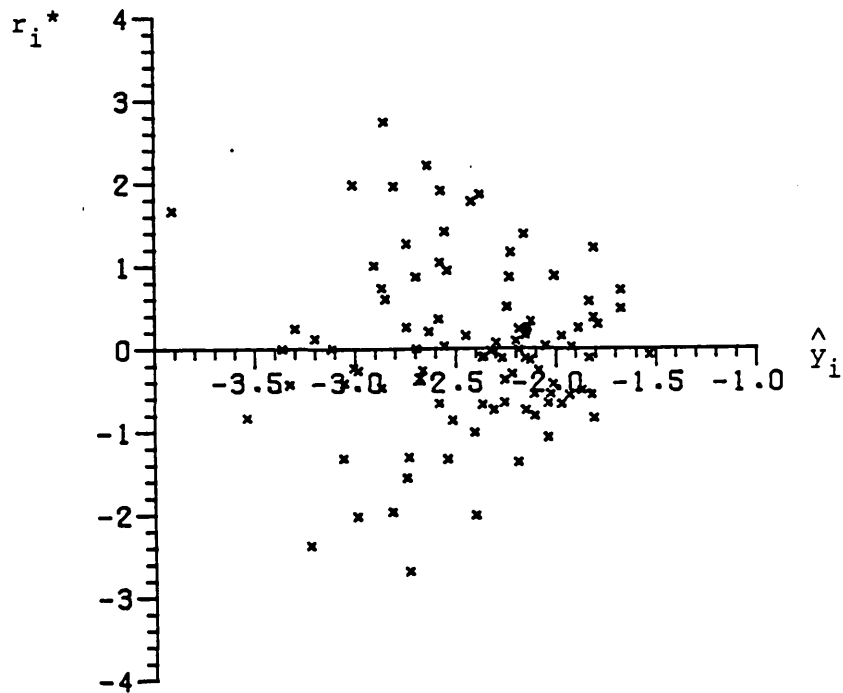


(b) ordered standardized residuals $r(i)'$ vs. Normal order statistics.

Figure 2.7 : Residual plots for regression of log-probabilities of exposure to air contamination.



(c) modified Cook statistics T_i vs. fitted values \hat{y}_i .



(d) jack-knifed residuals r_i^* vs. fitted values \hat{y}_i

Figure 2.7 : Residual plots for regression of log-probabilities of exposure to air contamination.

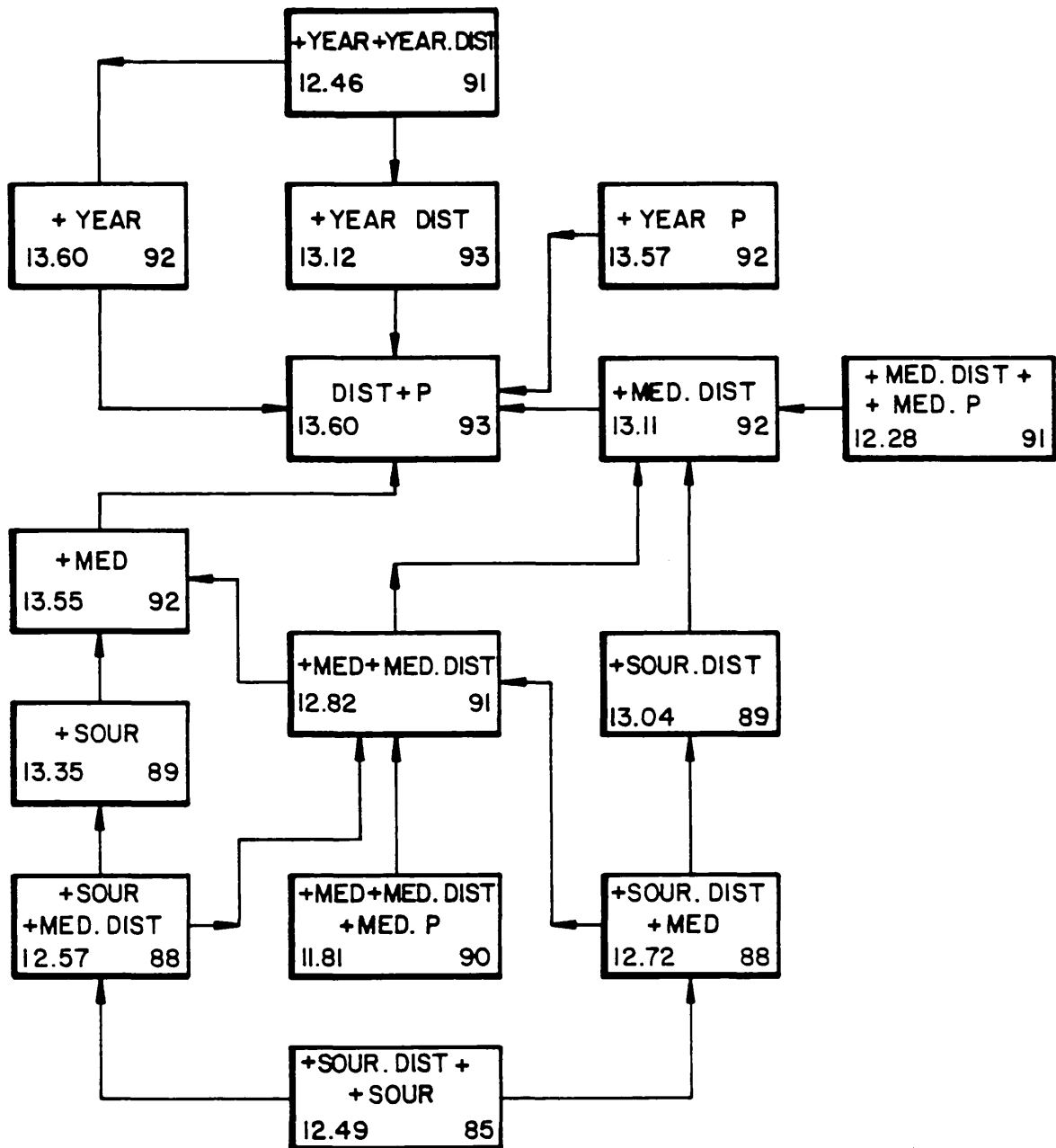
range over which the regression equations must be good. The jackknifed residuals in Figure 2.7(d) show no outliers.

The shape of the residual plot Figure 2.7(a) suggests that the binomial distribution might provide an alternative model for probabilities of exposure. However fits of such a model in GLIM show that the data are very overdispersed relative to the binomial distribution. Moreover since exposures due to separate three-hour releases are not independent the basis for use of the binomial distribution must be regarded as suspect.

A similar set of regression analyses performed on the 96 probabilities of exposure to wet deposition used the source windrose $p_w(\theta)$ restricted to winds of speed at least 5 m/s and occasions when it was raining at the source. Strictly speaking, $p_w(\theta)$ does not exactly measure the probable incidence of rainbearing air masses at the source, but it provides the best information easily available about it. Parallel regressions were performed using the source geostrophic windrose under all conditions, and the entire wet source windrose, but that using $p_w(\theta)$ gave the best fitting equations, and is recorded here. The regression using $p(\theta)$ was markedly worse than both the others, with substantially larger residual sum of squares - an unsurprising result.

The nesting diagram for models fitted to the log-wet exposure probabilities using $p_w(\theta)$ is shown in Figure 2.8. Details of the 'best' model, MED + MED.P + MED.DIST. are displayed in Table 2.2. It is clear that the data are more variable than the exposure probabilities in all conditions - the residual sums of squares for the models are about three times larger. The estimated value of σ is 0.3622, almost exactly twice its value of 0.1822 for the previous regression.

The fitted equation for Mol, Karlsruhe, and Heysham is



MODEL FORMULA	
Residual sum of squares	Degrees of Freedom

Figure 2.8 : Nesting structure of regression models fitted to log-probabilities of exposure to wet deposition.

Parameter	<u>North</u>		<u>Mediterranean</u>		
	Estimate	s.e.	Parameter	Estimate	s.e.
β_0	-2.487	0.1370	β_0	-0.6393	0.5929
β_1	0.2873	0.03547	β_1	0.8191	0.1883
β_2	-0.06376	0.01308	β_2	-0.1208	0.01763

Correlation matrices of estimates

	<u>North</u>			<u>Mediterranean</u>		
	β_0	β_1	β_2	β_0	β_1	β_2
β_0	1.0	0.8227	-0.3936	β_0	1.0	0.9833
β_1		1.0	0.0819	β_1		1.0
β_2			1.0	β_2		1.0

Estimate of σ is $S = 0.3622$

Table 2.2 Details of regression equations for probabilities of exposure in wet conditions due to releases of duration three hours.

$$\log\{\text{wet exposure probability}\} = \frac{0.2878 \log\{p_w(\theta)\}}{(3.547 \times 10^{-2})^w} - 2.487 - 0.0637d$$

(0.1370) (1.308 × 10⁻²)

.... 2.3.4

and that for Ispra and Cadarache is

$$\log\{\text{wet exposure probability}\} = \frac{0.8191 \log\{p_w(\theta)\}}{(0.1883)} - 0.6393 - 0.1208d$$

(0.5929) (1.763 × 10⁻²)

.... 2.3.5.

These equations show that patterns of behaviour for exposures to wet deposition differ for releases from different parts of Europe. Wet exposure probabilities for sources not in the Mediterranean area display quite low correlations with their source windroses and a fairly weak - but significant - decline with distance d . In contrast, wet exposure probabilities for sources in the Mediterranean area are highly correlated with $p_w(\theta)$ and drop more quickly as source-receptor distance increases. This suggests that trajectories spread and meander more in wet conditions in Northern Europe, but are narrower and straighter in similar conditions in the Mediterranean basin. For $p_w(\theta) \approx 0.2$, wet exposure probabilities for releases from Mediterranean sources are higher than for their more northerly counterparts by a factor about 2.0-1.5, decreasing as d increases; the corresponding factor for $p_w(\theta) \approx 0.1$ is 1.5-1.0 or so; and for $p_w(\theta) \approx 0.05$ the probabilities for releases from Mediterranean sources exceed those for northerly ones only up to about 450 kilometres, and thereafter are lower. The values of the estimates $\hat{\beta}_2$ suggest that rapid broadening of the puffs in wet conditions stops at distances of about 1500 km and 800 km respectively for releases from northerly and Mediterranean sources. These distances are larger than the corresponding values for exposures in all conditions and suggest that turbulent effects on dispersion occur on a larger scale in wet than in dry conditions, particularly over northern Europe, and to a lesser

extent in the Mediterranean area.

Residuals for the regression model are plotted in Figure 2.9. There are no obvious outliers or very influential points, but there are a number of rather large negative residuals. These correspond to probabilities at receptors mostly far to the south or west of their sources, and imply that trajectories travelling long distances to the south and west in wet conditions are slightly less frequent than the equations predict. This is consistent with the eastward-bound passage of wet air masses over Northern Europe. The residuals do not invalidate the fitted equations - which should be interpreted and used with care rather than treated as a *deus ex statistica*.

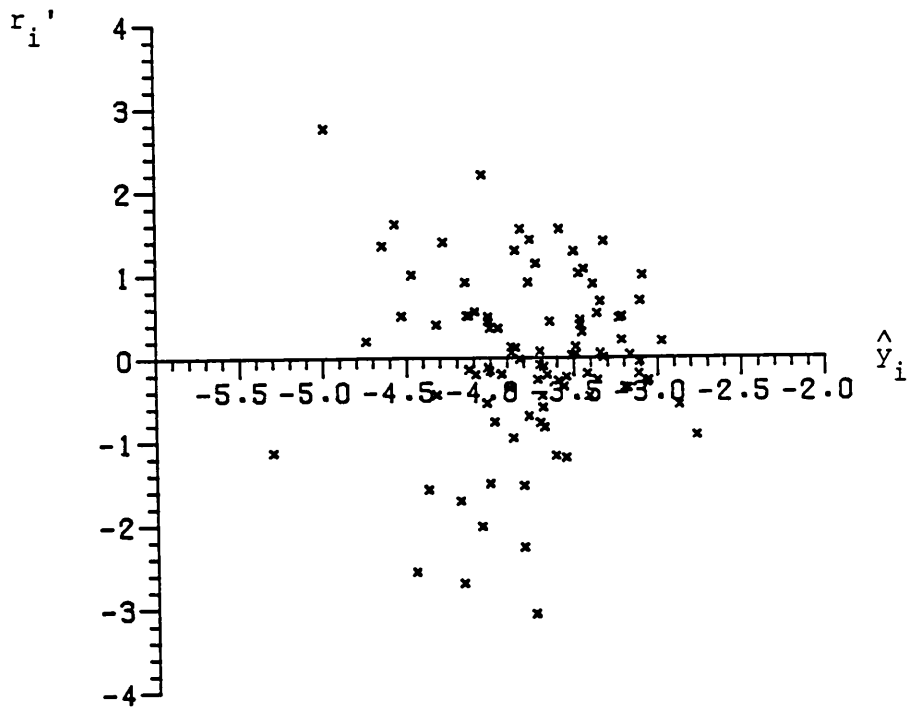
Confidence intervals for the MESOS values of exposure probabilities P_3 for releases of duration three hours may be based on the fact that $\log\{\hat{P}_3\}$ is roughly Normally distributed with mean

$$\beta_0 + \beta_1 \log\{p(\theta)\} + \beta_2 d \quad \dots \quad 2.3.6$$

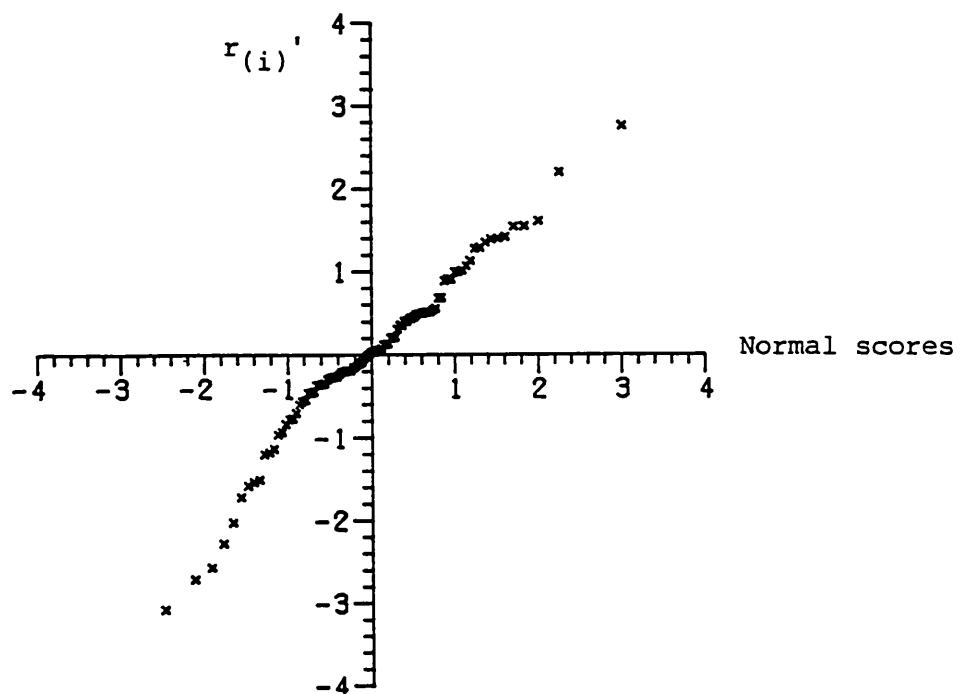
and variance $\tau^2 = \text{Var}\{\log(\hat{P}_3)\}$, with

$$\begin{aligned} \text{Var}\{\log(\hat{P}_3)\} = & \text{Var}(\hat{\beta}_0) + 2\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)\log\{p(\theta)\} + 2\text{Cov}(\hat{\beta}_0, \hat{\beta}_2)d \\ & + \text{Var}(\hat{\beta}_1)\log\{p(\theta)\}^2 + 2\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)\log\{p(\theta)\}d \\ & + \text{Var}(\hat{\beta}_2)d^2 \quad \dots \quad 2.3.7. \end{aligned}$$

Conditionally on the value of $\log\{\hat{P}_3\}$, the MESOS log-probability may be thought of as Normally distributed with mean $\log\{\hat{P}_3\}$ and variance σ^2 . Thus unconditional confidence intervals for $\log\{P_3\}$ may be found from the Normal distribution with mean $\log\{\hat{P}_3\}$ and variance $\omega^2 = \hat{\sigma}^2 + \tau^2$, with $\hat{\sigma}^2$ taken to be the estimate of σ^2 obtained from the regression model. That is, $\hat{\sigma}^2 = (n-p)^{-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2$. Hence P_3 lies in the interval $(\hat{P}_3 \exp(z_\alpha \omega), \hat{P}_3 \exp(-z_\alpha \omega))$ with probability approximately $1-2\alpha$, where $\Phi(z_\alpha) = \alpha$, and $\alpha < 0.5$.

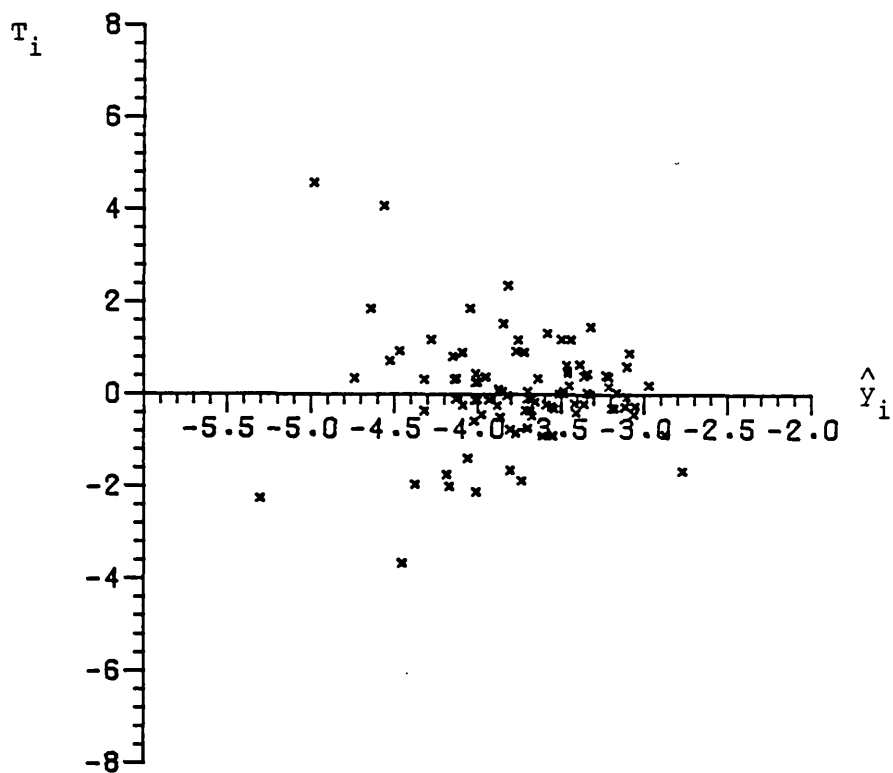


(a) standardized residuals r_i' vs. fitted values \hat{y}_i .

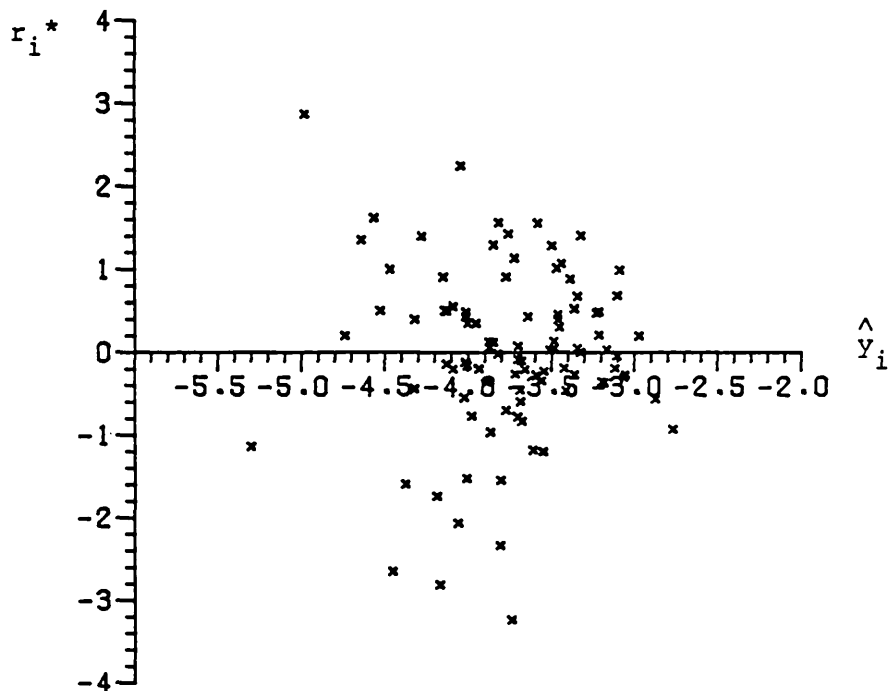


(b) ordered standardized residuals $r(i)'$ vs. Normal order statistics.

Figure 2.9 : Residual plots for regression of log-probabilities of exposure to wet deposition.



(c) modified Cook statistics T_i vs. fitted values \hat{y}_i .



(d) jack-knifed residuals r_i^* vs. fitted values \hat{y}_i .

Figure 2.9 : Residual plots for regression of log-probabilities of exposure to wet deposition.

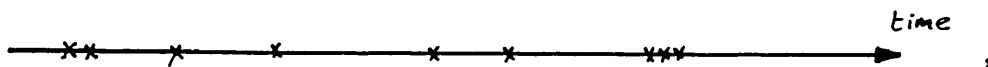
2.4 Exposure probabilities for releases of longer duration

Wrigley(1982) proposed that probabilities of exposure for releases of duration longer than three hours be related to those for release duration three hours by the power-law

$$P_t = P_3(t/3)^\delta .$$

Here P_u is the probability of exposure of a given type at a receptor due to releases of duration u hours at some fixed source. Typically $\delta \approx 0.5$. Although simple and appealing this formula has the obvious disadvantage that if used for large values of t it predicts probabilities greater than one as it has not the right asymptotic behaviour. This difficulty may be overcome by considering the following argument.

Suppose that a series of infinitesimal puffs is released from a source, and that the release times of those which later expose a remote receptor are noted. Suppose in addition that the probability of an exposure of the receptor by a puff released over a short time interval of length Δt is $\rho \Delta t$, and that exposures to different puffs are independent. The process of release times might look like



and is a homogeneous Poisson process of rate ρ . It has the property that the number of infinitesimal puffs released during an interval of length t which later expose the receptor has the Poisson distribution

$$\text{Prob(number of exposures = } k \text{)} = \frac{(\rho t)^k}{k!} \exp(-\rho t)$$

for non-negative values of k . It follows that the probability of at least one exposure due to the entire release is

$$1 - \exp(-\rho t) .$$

For a given positive value of ρ this rapidly approaches one as t increases: if $\rho \approx 0.1$ is the probability of exposure at a MESOS receptor due to a three-hourly release, then the probability of exposure due to a daily release would be about 0.55, whereas a typical value in the MESOS exposure database is 0.3. Clearly this is because exposures at receptors are not random. Short-term persistence of weather conditions leads to correlation among successive puff trajectories and so the probability of exposure to releases over longer periods rises more slowly than the form above.

The argument nevertheless suggests the following somewhat empirical modification of the equation:

$$P_t = 1 - \exp\{ -P_3(t/3)^\delta \},$$

for values of t greater than three hours.

Note as an aside that the Palm-Khintchine equations (Cox and Isham, 1980) imply that no stationary orderly process exists with

$$\text{Prob}\{ \text{no points in } (0,t) \} = \exp(-\rho t^\delta),$$

but this is a quibble since exposures of a receptor are not stationary, nor orderly, nor even a point process.

The modified form for P_t has the desirable property that for all values of $P_3 > 0$, $t > 3$, and $\delta > 0$ it lies in the range zero to one. Moreover for $P_3 \approx 0.25$ or less and fairly small values of t

$$P_t = 1 - \exp\{ -P_3(t/3)^\delta \} \approx P_3(t/3)^\delta,$$

which agrees with with the formula which ushered in this section.

The value of δ may be estimated from the MESOS data by noticing that

$$\log\{ -\log(1-P_t) \} - \log\{ P_3 \} = \delta \log\{t/3\}.$$

Regression of the known left hand side of this equation on $\log\{t/3\}$

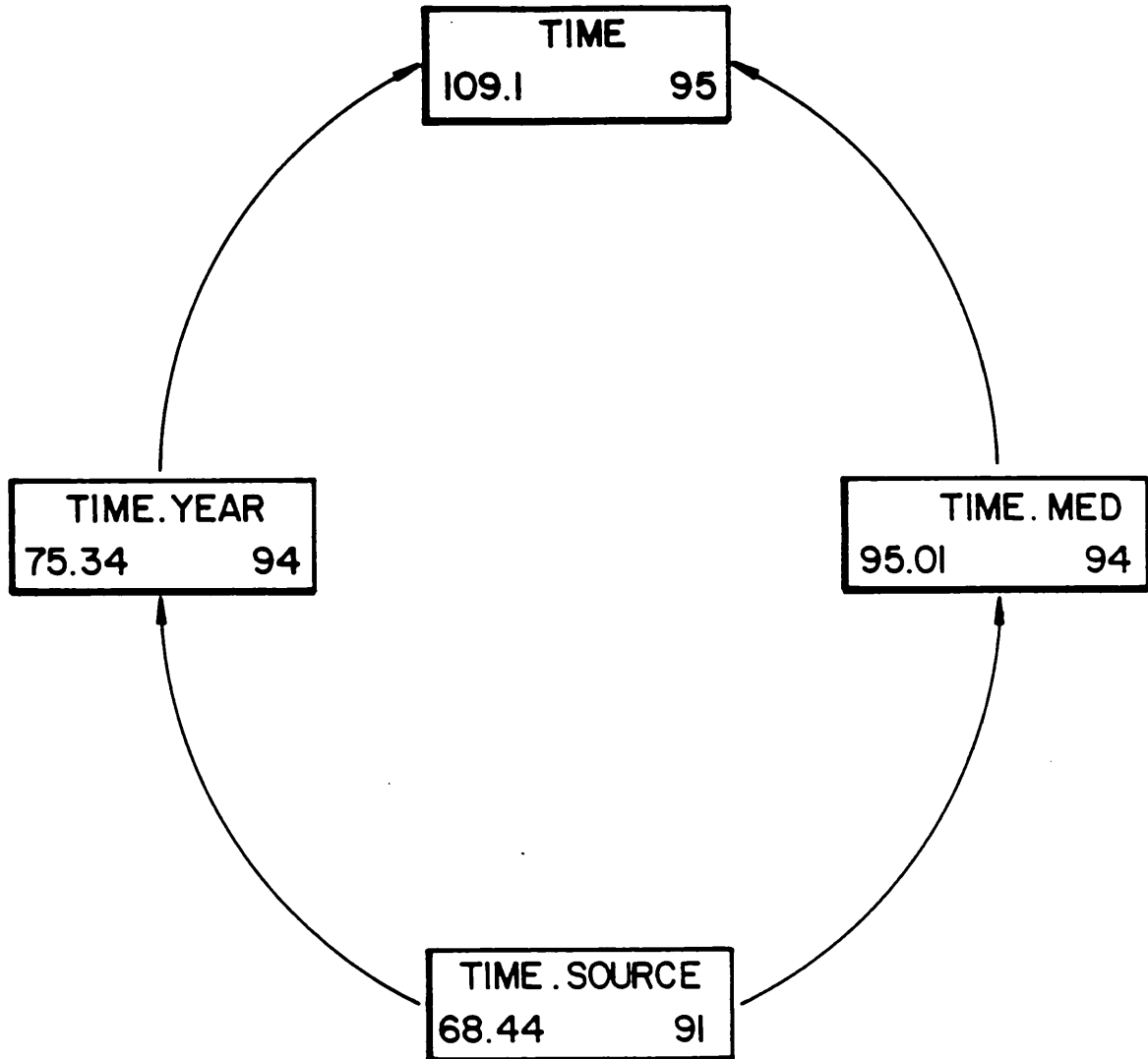
enables δ to be estimated.

Processes at different receptors were assumed approximately independent in Section 2.3, but this cannot even be roughly true of the values of P_s and P_t for different values of s and t at a single receptor. For each of the 96 receptors for which three-hourly data are available, a longer release duration $t = 6, 12, 24, 72, \text{ or } 168$ hours was randomly chosen and the probability P_t found from the MESOS data. The randomization was balanced to enable good estimation of source, year, or 'Mediterranean' effects on δ . A different randomization led to almost exactly the same values of $\hat{\delta}$ - which depend little on the particular randomization used.

Rather than the actual values of P_3 , the estimates P_3 predicted from the equations derived in Section 2.3 were used, so that a true idea was obtained of the variability which could eventually be expected in the estimates of P_t . This hardly affects estimates of δ at all, but it inflates their variance.

The regressions on $\log\{t/3\}$ are very strong. A nesting diagram and parameter estimates for probabilities of all types of exposure are shown in Figure 2.10 and Table 2.3. Statistically the best model is that which allows δ to be different in each year, but although an interesting result this is not useful since a model is needed for predictions in arbitrary years. The model which allows δ to depend on whether or not the source is Mediterranean is adopted instead. The values of $\hat{\delta}$ are close to the values expected; roughly 0.62 for more northerly sources, but about 0.58 for those in the Mediterranean basin. This indicates that probabilities for non-Mediterranean sources increase rather faster with release duration. An interpretation of this is that correlation between successive trajectories tends to be higher for releases from sources in the Mediterranean area because typical weather conditions there are

Figure 2.10 : Nesting structure for dependence of probabilities of exposure to air contamination on release duration.



Model	Parameter	Estimate	s.e.
Time	δ	0.6040	0.0062
Time.Med	δ_{Uniform}	0.6209	0.0074
	$\delta_{\text{Mediterranean}}$	0.5763	0.0094

Table 2.3 : Parameter estimates for dependence of probabilities of exposure in all conditions on release duration.

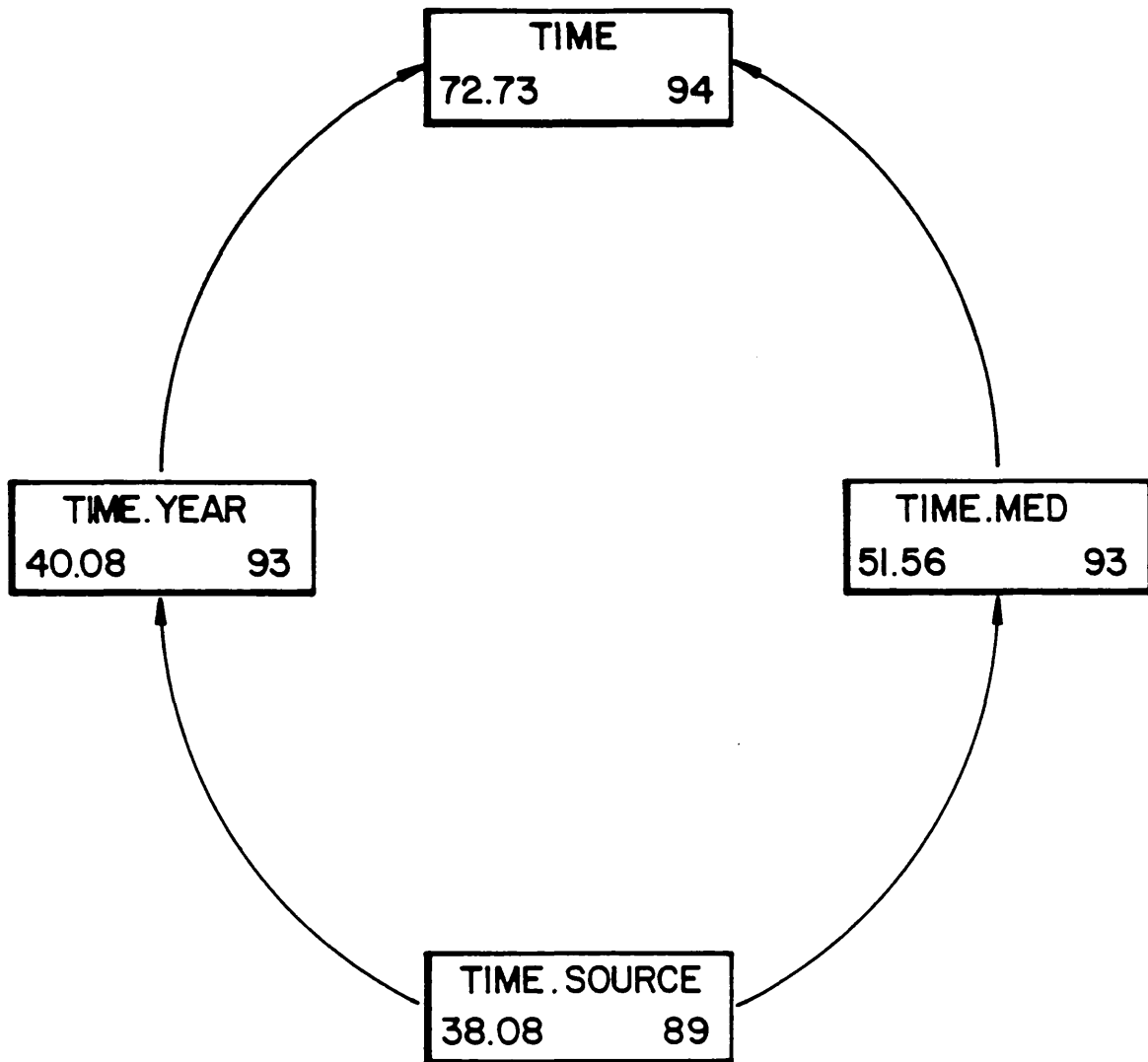
influenced by factors different to those affecting more northerly sources. This is probably confounded to some extent with the effect on atmospheric dispersion of the unusual series of anticyclones over north-west Europe in the summer of 1976 - so-called blocking situations - which would have led to higher correlation between trajectories initiated at sources there, explaining why the model allowing yearly variation of δ fits well.

The corresponding diagram and estimates for exposures to wet deposition may be found in Figure 2.11 and Table 2.4. They have the same pattern as those in Figure 2.10, but the sums of squares are lower, indicating a better overall fit of the model. Once again the equation to be preferred allows δ to depend on source location. The values of $\hat{\delta}$ are higher: about 0.77 in the north and 0.69 in the Mediterranean. Introduction of an extra element - rainfall at the receptor - makes wet exposure more nearly random than exposure in all conditions, hence the slightly higher values of $\hat{\delta}$. The values of $\hat{\delta}$ again imply that trajectories are more highly correlated for 1976 and Mediterranean sources.

The residual plots for the regressions display an excellent fit, except for an outlier in the wet exposure probabilities. It represents wet exposure at the receptor 800 kilometres south-east of Ispra due to releases of duration one week. Its probability is too high to be properly fitted by the model, and has been discarded.

Estimates of exposure probabilities at remote receptors are of little use if there is no idea - in the form of confidence intervals or the like - of their likely variability. Such intervals are easily obtained - as outlined above - for three-hour releases, but are harder to find for longer releases. Suppose that a predicted probability of exposure at some receptor due to a three-hour release is \hat{P}_3 . Given \hat{P}_3 and $\hat{\delta}$ as fixed, the estimated probability of

Figure 2.11 : Nesting structure for dependence of probabilities of exposure to wet deposition on release duration.



Model	Parameter	Estimate	s.e.
Time	δ	0.746	0.0071
Time . Med	δ_{Uniform}	0.771	0.0073
	$\delta_{\text{Mediterranean}}$	0.695	0.0099

Table 2.4 : Parameter estimates for dependence of probabilities of exposure in wet conditions on release duration.

exposure due to a t-hour release is

$$\hat{P}_t = 1 - \exp\{ -\hat{P}_3(t/3)^{\hat{\delta}} \}.$$

If there are n_t periods of length t hours in a year, and if exposures in each of them are independent, then the variance of \hat{P}_t is the usual binomial formula $P_t(1-P_t)/n_t$, with P_t the true t-hour exposure probability. However this does not allow for the variability of \hat{P}_3 . A set of approximate confidence intervals which allow for this is found by recalling that for any random variable Z,

$$\text{Var}(\hat{P}_t) = E_Z[\text{Var}(\hat{P}_t | Z=z)] + \text{Var}_Z[E(\hat{P}_t | Z=z)].$$

Now $\log(\hat{P}_3) + \hat{\delta}\log(t/3)$ is roughly Normal with mean μ' and variance about τ^2 , say, since the variance of $\hat{\delta}$ is very small compared with that of $\log(\hat{P}_3)$. Suppose that Z has the standard Normal distribution, in which case

$$\hat{P}_t = 1 - \exp\{ -\exp\{ \mu' + \tau Z \} \} = g(\mu' + \tau Z),$$

say. Then

$$E_Z[\text{Var}(\hat{P}_t | Z=z)] = (n_t \sqrt{2\pi})^{-1} \int_{-\infty}^{\infty} \exp(-z^2/2) g(\mu' + \tau z)(1-g(\mu' + \tau z)) dz,$$

and moreover

$$\begin{aligned} \text{Var}_Z[E(\hat{P}_t | Z=z)] &= (\sqrt{2\pi})^{-1} \int_{-\infty}^{\infty} \exp(-z^2/2 - 2\exp\{\mu' + \tau z\}) dz \\ &\quad + \{ (\sqrt{2\pi})^{-1} \int_{-\infty}^{\infty} \exp(-z^2/2 - \exp\{\mu' + \tau z\}) dz \}^2; \end{aligned}$$

thus if

$$J_k(\mu', \tau) = (\sqrt{2\pi})^{-1} \int_{-\infty}^{\infty} \exp(-z^2/2 - k\exp\{\mu' + \tau z\}) dz,$$

the unconditional variance of \hat{P}_t is

$$\text{Var}(\hat{P}_t) = \{ J_1(\mu', \tau) - J_2(\mu', \tau) \} / n_t + J_2(\mu', \tau) - J_1(\mu', \tau)^2 \dots 2.4.1.$$

The idea now is to use approximate Normal confidence intervals for the true value of P_t - to say that an approximate $(1-2\alpha) \times 100\%$

confidence interval for the true value of P_t is $\hat{P}_t \pm z_\alpha \sqrt{\text{Var}(\hat{P}_t)}$, where $\Phi(z_\alpha) = \alpha$, and $\alpha < 0.5$.

A good approximation to the integral $J_k(\mu', \tau)$ is needed. This can be found using a saddlepoint expansion: if

$$K = (\sqrt{2\pi})^{-1} \int_{-\infty}^{\infty} \exp(-h\{z\}) dz,$$

then if z^* is such that $h'(z^*) = 0$ and some regularity conditions hold,

$$K \approx \exp(-h\{z^*\}) / \sqrt{h''\{z^*\}}.$$

In this case $h(z) = z^2/2 + k \exp(\mu' + \tau z)$, so after some differentiation and substitution it turns out that

$$J_k(\mu', \tau) \approx \exp\{-v^{*2} - v^*/\tau\} / \sqrt{(1 + \tau v^*)} \quad \dots \quad 2.4.2,$$

where v^* is the unique solution of the equation

$$\log\{v\} = \mu' + \log\{k\tau\} - \tau v \quad \dots \quad 2.4.3.$$

Note that v^* is positive and always exists if $\tau > 0$; v^* is easily found graphically or by a simple bisection search on a programmable pocket calculator. It generally lies close to zero for the range of values of μ' and τ of interest here. Approximate intervals based on these equations are within ± 0.001 of those based on more sophisticated 'exact' calculations, which are not reported here, so these approximate intervals are quite adequate to the job for which they are intended.

2.5 A taxonomy of windroses

Statistical analysis of the exposure probabilities suggests different types of puff trajectory behaviour depending broadly on whether or not the source is in the Mediterranean basin. Here geostrophic windrose data needed to apply the model equations to releases from sources in Western Europe are displayed and discussed. The emphasis is on classification of the windroses for the present purpose rather than on wholly meteorological issues, although Figure 2.12 is of interest in its own right. This section is based partly on Manning(1984), who extracted the windroses from the MESOS database and interpreted them.

Figure 2.12 shows for a 7×13 grid over Europe the proportion of winds of speed 5 m/s or more compared with all windspeeds, directed into 30° sectors, for winds in all conditions and winds in wet conditions. The grid elements are 2° latitude by 4° longitude. The map area and grid elements are displayed in Figure 2.13, which also shows ground over 3000' - about 1 km - high. The windrose for each element is an average of those at sixteen points spacially evenly within the element, deduced from pressure fields found by polynomial interpolation between appropriately adjusted measurements at surface stations and weather ships throughout the year 1976. The original observations are 'present weather' data recorded every three hours; conditions are deemed 'wet' at a point on the grid element if precipitation is observed there or nearby.

Note that the windroses are not probability density functions. The value corresponding to θ° on any one of them is an estimate of the annual average proportion of geostrophic winds of strength 5 m/s or more in wet and in all conditions directed into a sector of arc 30° centred on θ . In the notation of Section 2.3, they are roughly

8W

4W

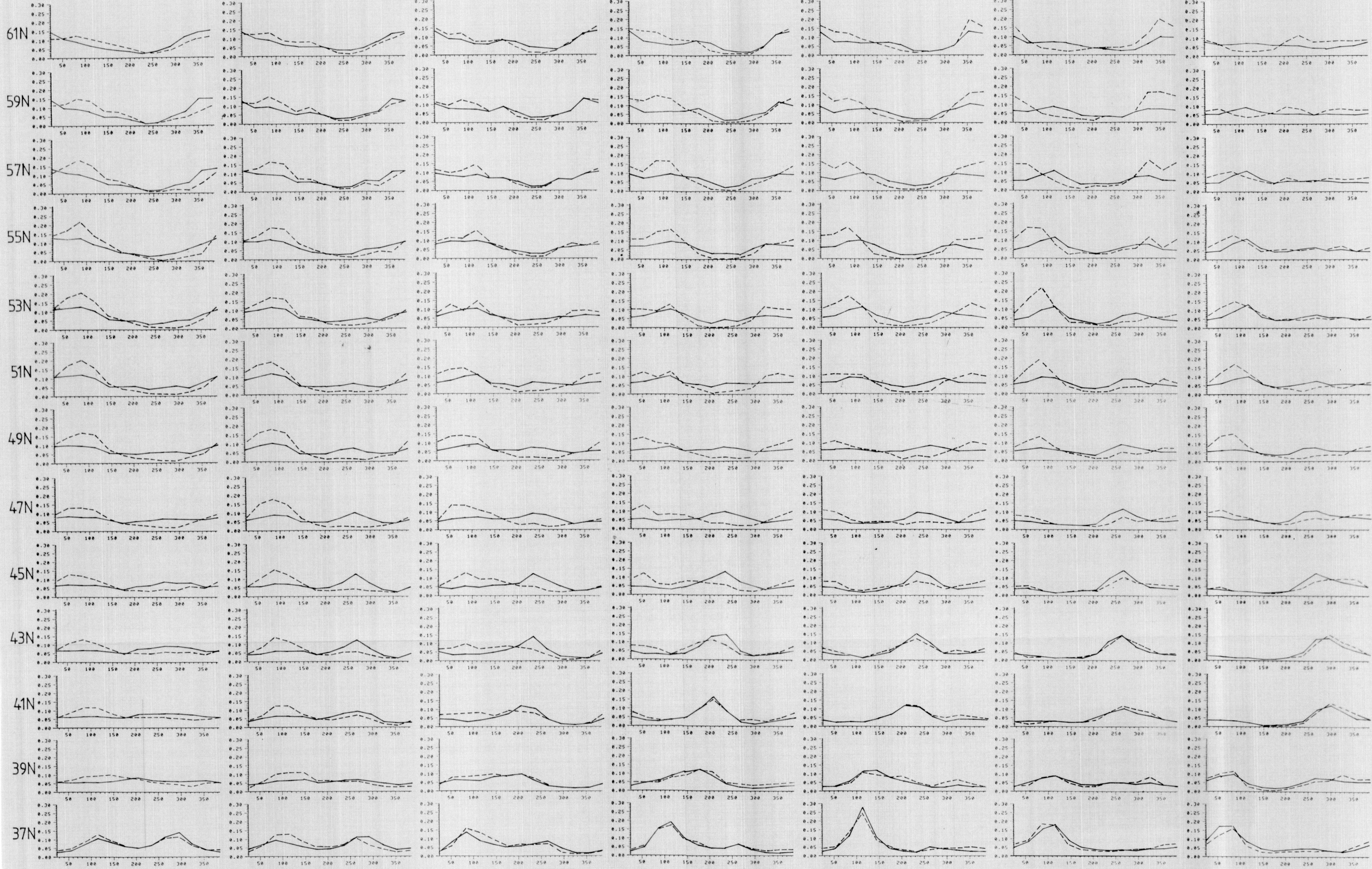
0

4E

8E

12E

16E



ANGLE

ANGLE

ANGLE



ANGLE

ANGLE

ANGLE

ANGLE

Proportion of winds > 5 M/S into 30 degree sectors, interpolated between their centres :

key
 ALL =  WET = 

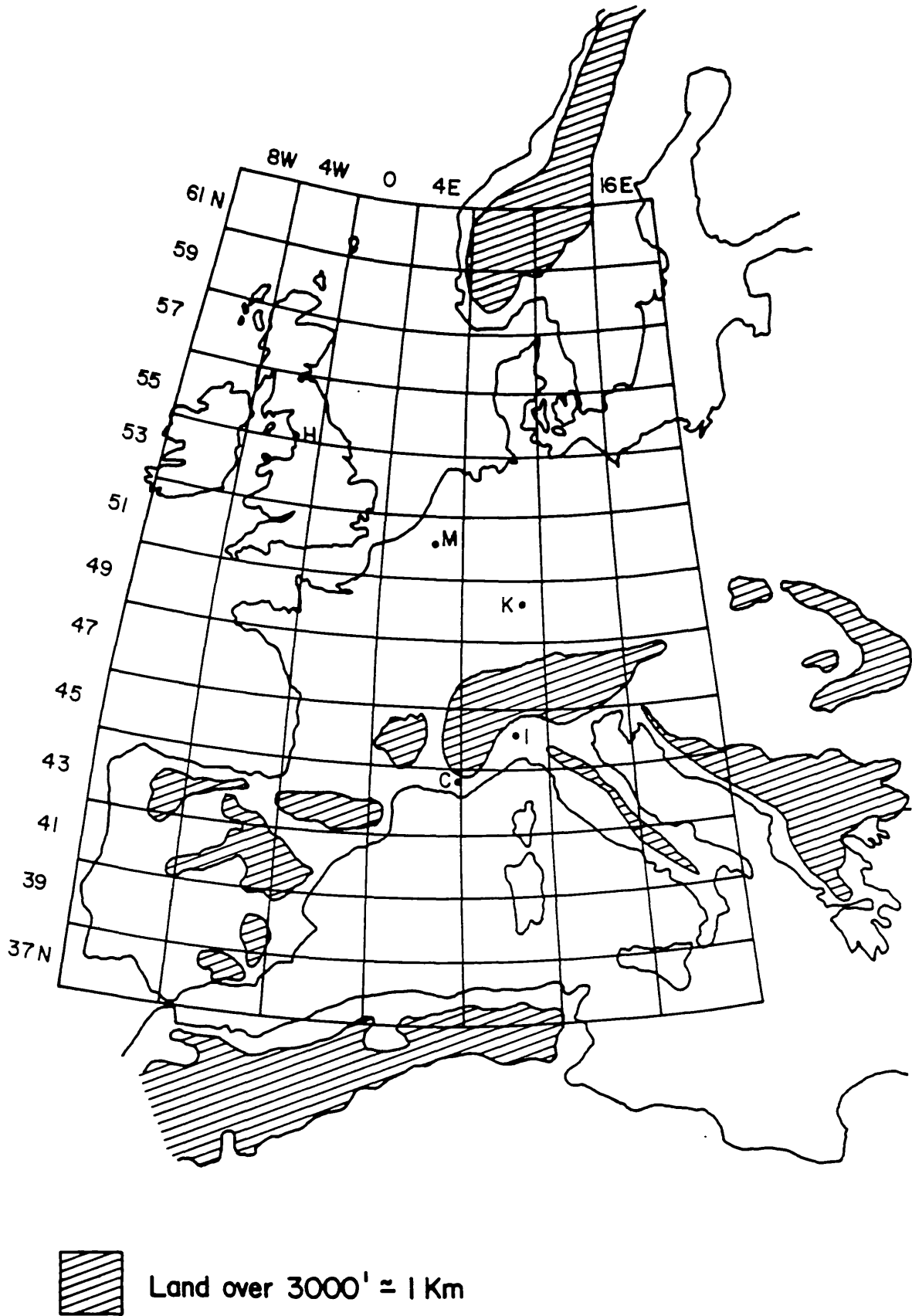


Figure 2.13 : Map area and grid elements for windrose classification.

$$\int_{\theta-\eta}^{\theta+\eta} p(u) du$$

and

$$\int_{\theta-\eta}^{\theta+\eta} p_w(u) du,$$

where $n=2\pi \times 15/360$. It is assumed that variation in $p(\theta)$ over arcs of 30° or less is irrelevant to the purpose at hand.

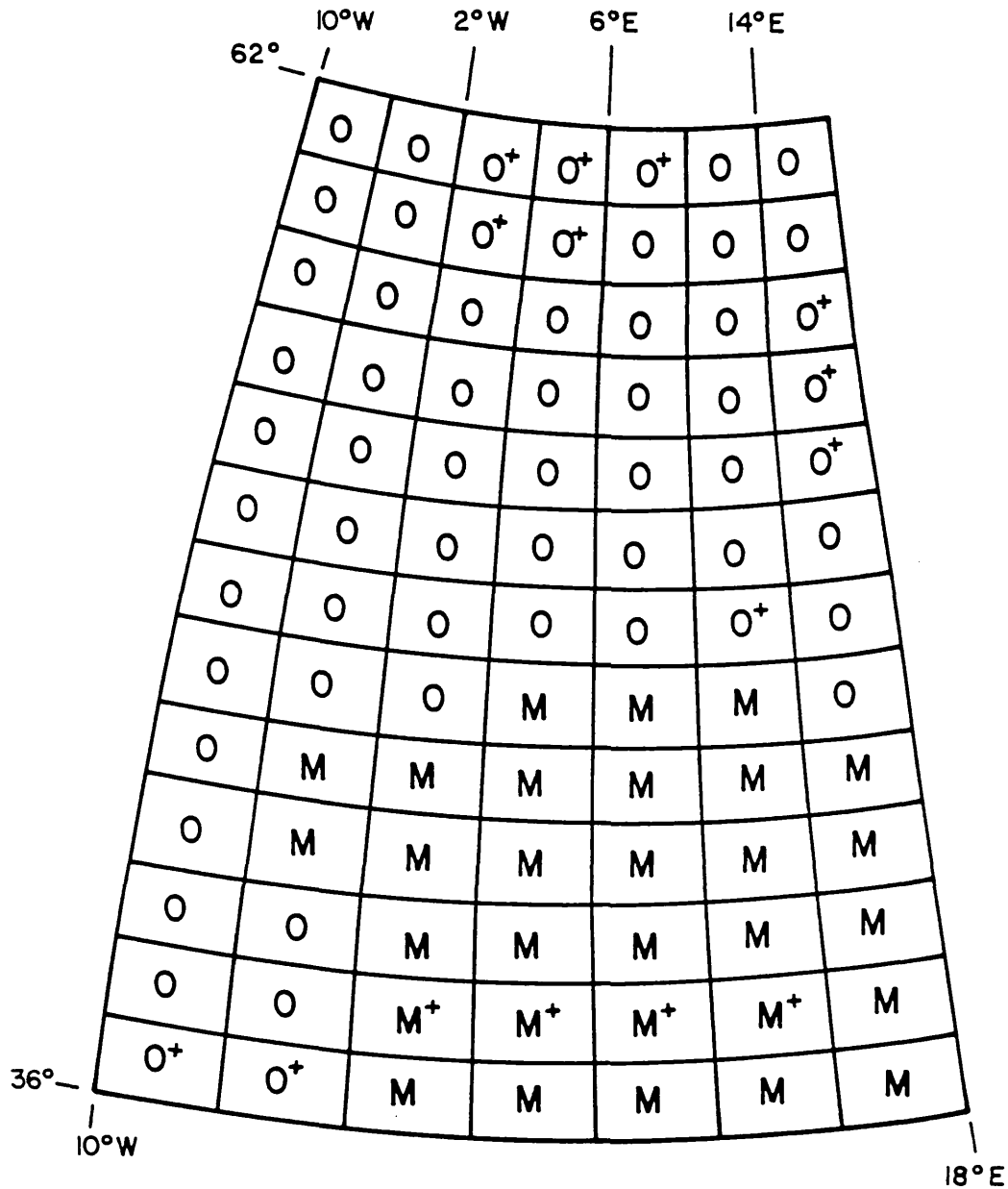
The windroses in all conditions have been classified twice. Figure 2.14 shows a naive classification of them into roughly uniform (O) or Mediterranean (M) types; whereas Figure 2.15 shows a more detailed classification taking into account the qualifications below.

Figure 2.12 shows a few almost truly uniform windroses: at (16E,61N) behind the Scandinavian mountains; over the Spanish peninsular at (8W,41N), (8W,39N), and (4W,39N); and at (4E,49N) and (8E,49N).

Windroses over Britain and the North Sea show the expected slight maximum for westerly winds, but are grouped with the uniform windroses. The same applies to those just south of the Norwegian mountains, which have a secondary maximum for easterly winds. Windroses to the west of these mountains shows few winds from the east and just south of east. Of the five affected - (0,61N), (4E,61N), (0,59N), (4E,59N), and (8E,61N) - the first four lie over the North Sea, but the fifth - in Norway - though tentatively classed as uniform, may need separate consideration if the regression equations are to be used for sites in that area.

Three windroses for the area towards the north of the eastern boundary of the map - (16E,53N), (16E,55N), and (16E,57N) - also have winds mainly from one direction. They have been classified as uniform but may need separate treatment. The two more northerly ones lie mostly over the Baltic Sea and the last lies over Poland.

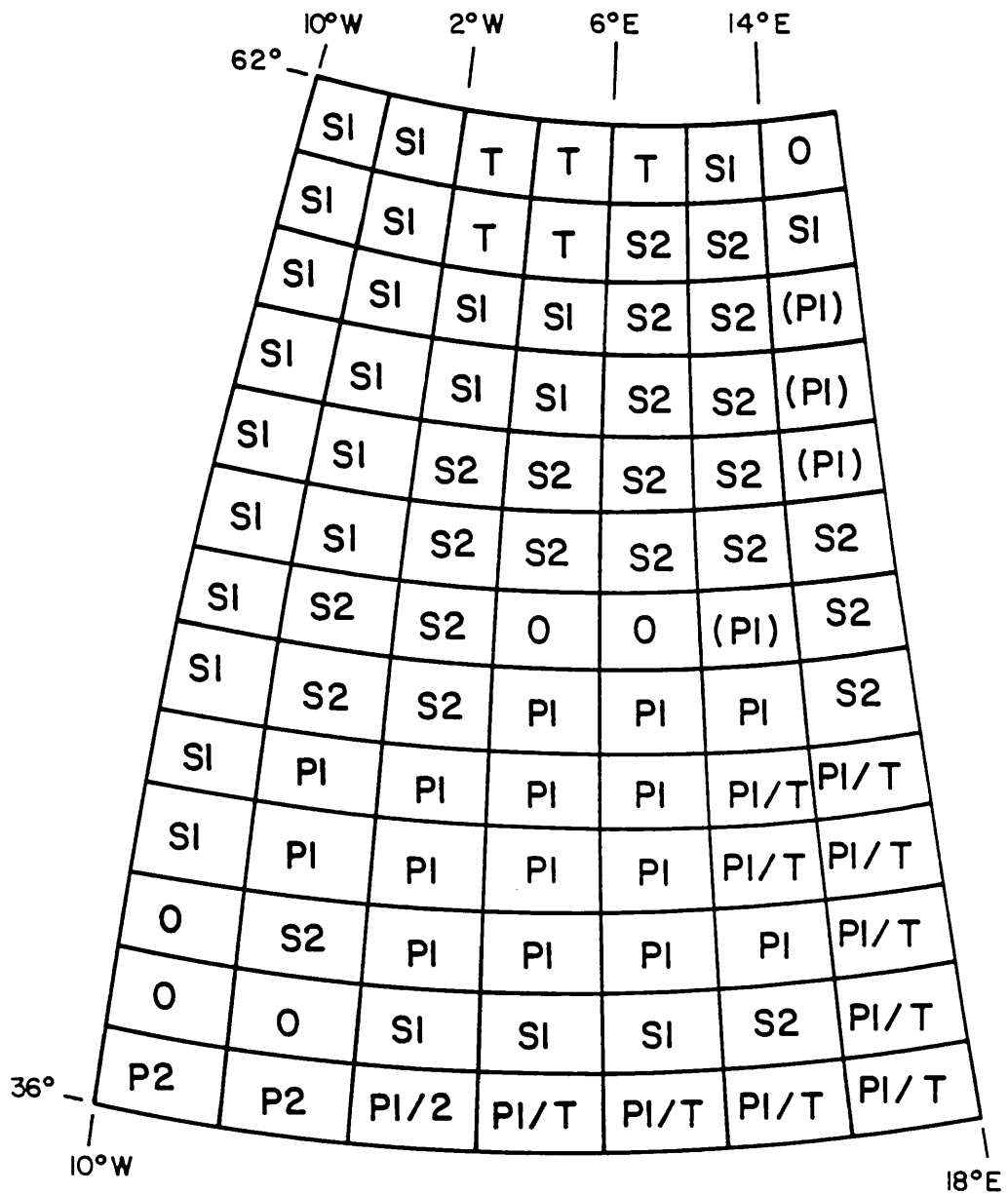
The windroses for the north of Europe may be contrasted with those in the Mediterranean area, the Alps, the Pyrenees, and Southern



Key to Symbols

- O Uniform
- M Mediterranean
- + See more detailed classification

Figure 2.14 : A simple classification of geostrophic winds of speed 5 m/s or more in all conditions.



Key to Symbols

- O Uniform
- S1 Roughly sinusoidal with single peak
- S2 Roughly sinusoidal with double peak
- P1 Single sharp peak
- P2 Two sharp peaks
- T Transitional

Figure 2.15 : A detailed classification of geostrophic winds of speed 5 m/s or more in all conditions.

France, which tend to be strongly dominated by winds from a single direction - which however varies in different parts of the region. All grouped together as one category, they are dominated by cyclogenesis in the lee of the Alps or a local climatology due to a moist basin of warm water surrounded almost entirely by mountain barriers. The two grid elements over southern Spain and Portugal - (8W,37N) and (4W,37N) - do not fall into this class as they show strong contributions from both westerly and easterly winds in the Straits of Gibraltar. However they are treated as Mediterranean in the simple classification in Figure 2.14.

Figure 2.16 shows a naive classification of wet geostrophic windroses into northerly (O) and Mediterranean (M); and Figure 2.17 shows a more detailed classification.

The pattern the wet windroses take over north-west Europe is consonant with the passage of wet air-masses from the Atlantic: the windroses generally have a peak corresponding to the arrival of moist air from the west and south-west, and a dip for easterly winds. This is the general pattern - albeit with local variations - north of a line roughly joining Galicia at (8W,43N) to the north of the Alpine barrier at (8E,49N) and then to (16E,49N), for areas in the body of Europe unaffected by the Scandinavian mountains. It is more obvious over land than over the sea, for two possible reasons: 'present weather' data is scarcer over the sea, so the database is less accurate; and - more plausibly - an air mass from any direction arriving at any point out to sea is moist, but this is not the case for points over land. Over continental Europe, for example close to the Alps at (8E,49N), (8E,51N), and (4E,51N), wet winds tend to be associated with south-south-easterly as well as south-westerly winds, due to thunderstorms in anticyclonic conditions through the hot summer of 1976.

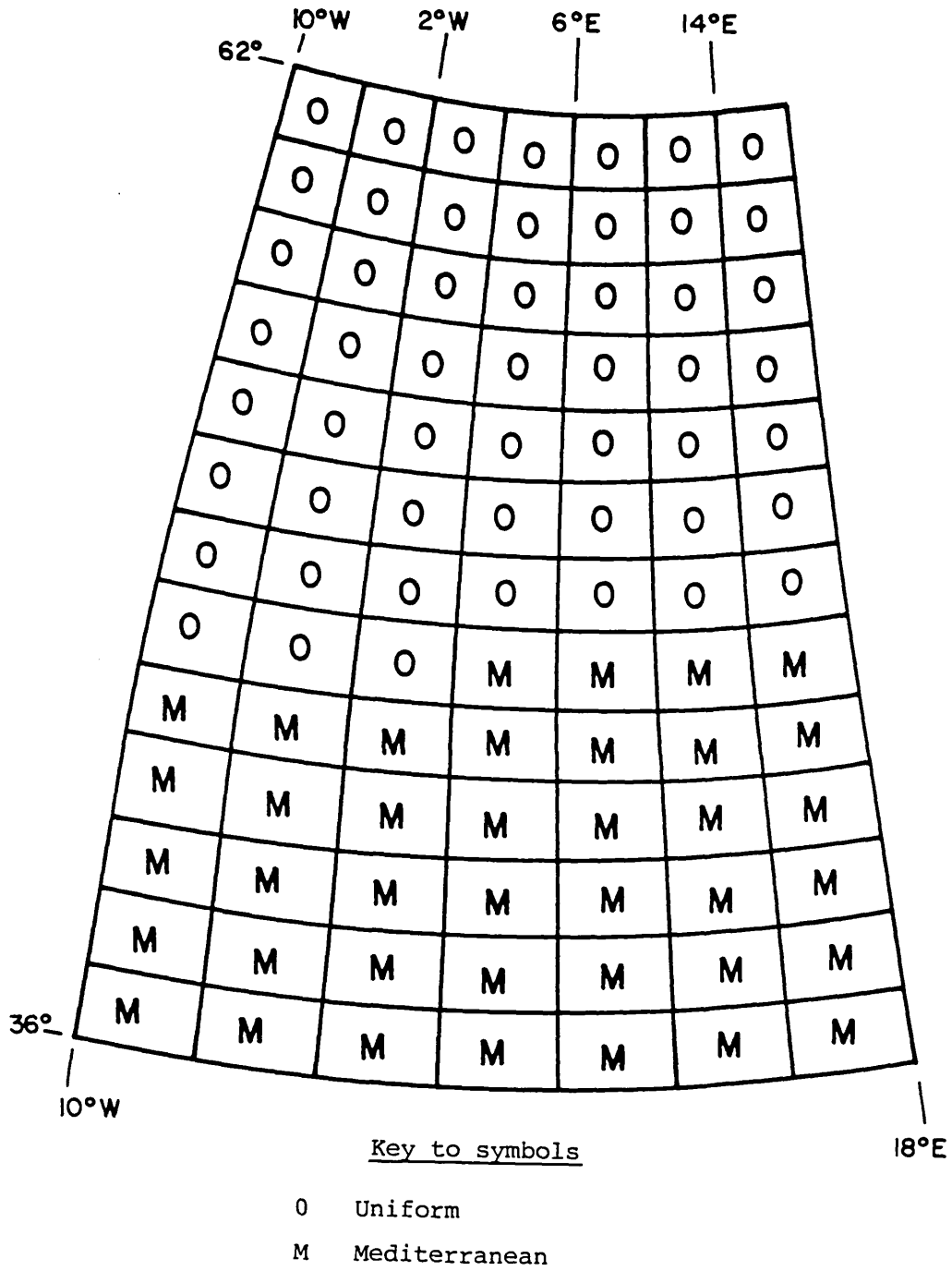
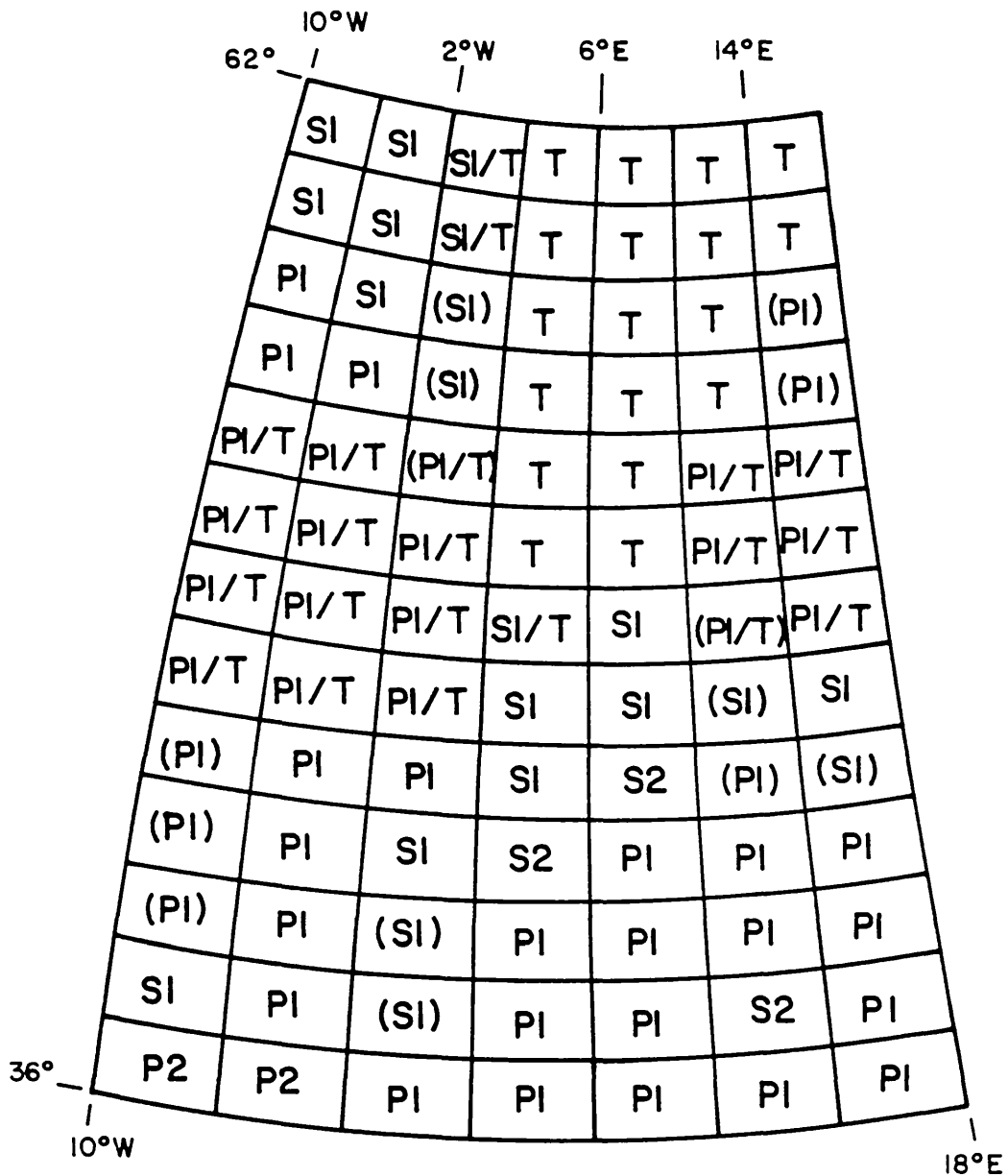


Figure 2.16 : A simple classification of geostrophic winds of speed 5 m/s or more in wet conditions.



Key to Symbols

- 0 Uniform
- S1 Roughly sinusoidal with single peak
- S2 Roughly sinusoidal with double peak
- P1 Single sharp peak
- P2 Two sharp peaks
- T Transitional

Figure 2.17 : A detailed classification of geostrophic winds of speed 5 m/s or more in wet conditions.

The grid elements surrounding the Scandinavian mountains have wet windroses dominated by moist air masses from the south-west and from south-east over the Baltic Sea. This explains the roses for the elements (16E,61N), (16E,59N) and (16E,57N), particularly since air arriving from the west is there relatively dryer due to the orographic effects of the Norwegian mountains.

The wet windroses over northern Europe, and especially to the north-west, are in general exaggerated and skewed versions of the roses in all conditions; the windroses in all and in wet conditions for grid elements over the Mediterranean and Italy tend to coincide closely with one another. This suggests that winds in dry conditions in much of this region are light and uniformly distributed. However the windroses do not follow each other so closely over the bulk of the Spanish peninsular, although that at (8W,37N) near the Straits of Gibraltar is an exception; nor near Marseille - (4E,43N) - where the windrose is affected by both the Alps and the Pyrenees; nor over the Alps themselves; nor over Austria and Yugoslavia at (16E,45N) and (16E,47N), which are transitional between typical northerly and Mediterranean windroses.

If it is intended to use the model described in Sections 2.3 and 2.4 to predict exposure probabilities for releases from a source in the map area in Figure 2.13, then the classifications of the grid element in which the source lies should be found from Figures 2.14 and 2.16. The equations appropriate to the types of the windroses in the grid element should then be used to predict probabilities of exposures in wet and in all conditions. For exposures to time-integrated air concentrations or dry deposition, for example, the equation for Heysham, Karlsruhe, and Mol should be used if the grid element is type O, but the equation for Cadarache and Ispra should be used if the element is type M. A similar procedure should be

followed for exposures to wet deposition. It may be useful to compare the results of both equations in cases where the more detailed classification indicates some uncertainty about source windrose type. It seems plausible that if a source is not exclusively dominated by either a northerly or a Mediterranean climatology, then dispersion characteristics of releases initiated there will tend to lead to probabilities which lie between the values predicted by the different equations.

2.6 A verification study

The accuracy of the model described above may be assessed by comparing its predictions with MESOS probabilities for releases from Hannover and Stuttgart computed using the 1973 database. These MESOS probabilities were not used to fit the statistical equations, so to an extent such comparison provides independent verification of them. The study is incomplete insofar as data for releases from sources in the Mediterranean basin - apart from Cadarache and Ispra - have not been obtained using MESOS. However there seem to be no a priori reasons why qualitatively different conclusions would be reached were such data available for comparison to be made.

The comparison is made for exposures in all conditions in Table 2.5. Agreement is generally very good for releases from both Hannover and Stuttgart. The maximum error for releases of duration three hours is about 6%, and for those of duration one week is about 25%; but most sets of probabilities compare much more favourably. Systematic errors arise because the simple predictive equations used take no account of trajectory persistence in certain directions and of topographic effects. Thus exposure probabilities for three-hour releases from Hannover are underpredicted by 2% or so to the east and are overpredicted by a similar amount to the north and west. There

Receptor	Release duration (hours)					
	3	6	12	24	72	168
1 100 km N	10.6	15.1	21.8	31.7	56.5	85.2
-	12.8	17.9	26.2	37.3	60.3	79.0
2 200 km N	7.4	10.0	16.1	23.8	46.2	72.8
	11.5	16.2	23.8	34.2	56.2	75.3
3 400 km N	5.3	8.0	12.1	18.2	38.1	62.4
	9.2	13.2	19.6	28.5	48.5	67.5
4 800 km N	4.1	5.9	9.1	15.1	27.0	41.6
	5.9	8.7	13.1	19.4	34.8	51.5
5 100 km E	16.5	23.4	34.4	48.3	76.1	94.7
	15.2	20.8	30.2	42.5	66.5	84.3
6 200 km E	15.5	22.0	31.7	44.7	74.0	91.6
	13.6	18.9	27.5	39.1	62.5	80.9
7 400 km E	12.3	18.4	27.3	40.8	69.7	89.2
	10.9	15.5	22.8	32.8	54.4	73.6
8 600 km E	11.2	16.2	25.3	38.0	62.9	83.2
	8.8	12.6	18.7	27.3	46.8	65.6
9 100 km S	11.7	16.5	24.2	32.5	56.0	77.8
	10.8	15.3	22.5	32.5	54.0	73.1
10 200 km S	10.6	14.9	21.0	27.3	49.2	70.8
	9.7	13.8	20.4	29.7	50.1	69.2
11 400 km S	9.0	12.5	17.0	23.4	47.1	73.3
	7.8	11.3	16.8	24.6	42.8	61.1
12 800 km S	2.9	4.4	7.4	11.9	27.4	51.5
	5.0	7.4	11.2	16.6	30.2	45.6
13 100 km W	11.6	16.2	23.2	31.7	54.3	75.8
	13.7	19.0	27.7	39.3	62.8	81.2
14 200 km W	10.6	14.9	21.6	29.7	48.3	68.4
	12.3	17.2	25.2	36.1	58.7	77.6
15 400 km W	9.4	11.8	17.0	23.8	44.1	63.4
	9.9	14.1	20.8	30.2	50.9	69.9
16 800 km W	6.4	8.8	12.1	17.0	32.5	55.5
	6.4	9.3	14.0	20.6	36.7	53.9

Table 2.5 (a) : Comparison of MESOS and statistically predicted %-probability of exposure at 16 receptor points for several release durations : Hannover dry deposition.

MESOS
 statistical prediction } %-probability

Receptor	Release duration (hours)					
	3	6	12	24	72	168
1 100 km NE	15.9	21.7	30.9	40.9	67.4	92.6
-	12.5	17.5	25.6	36.5	59.3	78.1
2 200 km NE	14.3	21.1	27.1	38.0	58.8	82.0
	11.2	15.8	23.2	33.4	55.3	74.4
3 400 km NE	11.3	15.4	21.2	31.3	50.7	74.8
	9.0	12.9	19.1	27.8	47.6	66.5
4 800 km NE	7.7	10.3	14.5	21.4	38.2	58.0
	5.8	8.5	12.8	19.0	34.0	50.5
5 100 km SE	15.0	21.1	28.3	42.3	69.7	88.2
	10.8	15.3	22.5	32.5	54.0	73.1
6 200 km SE	11.7	16.6	23.8	34.0	61.0	82.7
	9.7	13.8	20.4	29.7	50.1	69.2
7 400 km SE	5.1	8.1	12.5	18.2	35.2	57.0
	7.8	11.3	16.8	24.6	42.8	61.1
8 600 km SE	1.6	3.0	5.0	8.7	21.1	43.6
	6.2	9.1	13.7	20.3	36.1	53.2
9 100 km SW	14.5	19.3	27.7	38.0	67.0	86.2
	13.1	18.3	26.7	38.0	61.2	79.8
10 200 km SW	14.8	19.8	27.7	39.6	67.9	89.7
	11.8	16.6	24.3	34.8	57.1	76.2
11 400 km SW	14.4	19.4	27.1	37.2	68.3	89.7
	9.5	13.5	20.0	29.1	49.3	68.4
12 600 km SW	13.7	18.8	24.0	33.3	60.1	86.7
	7.6	11.0	16.4	24.1	42.1	60.3
13 100 km NW	11.4	16.1	23.2	33.6	58.0	79.3
	13.1	18.3	26.7	38.0	61.2	79.8
14 200 km NW	9.6	14.0	20.4	29.7	54.1	78.8
	11.8	16.6	24.3	34.8	57.1	76.2
15 400 km NW	7.1	10.0	13.9	20.2	38.2	62.4
	9.5	13.5	20.0	29.1	49.3	68.4
16 800 km NW	4.2	6.2	9.3	14.3	27.9	45.1
	6.1	8.9	13.4	19.9	35.5	52.3

Table 2.5 (b) : Comparison of MESOS and statistically predicted %-probability of exposure at 16 receptor points for several release durations : Stuttgart dry deposition.

MESOS
 statistical prediction } %-probability

is quite close agreement between MESOS and statistically predicted probabilities to the south of Hannover, except at 800 km where the statistical model cannot allow for diversion of trajectories around the Alps.

Exposure probabilities for three-hour releases from Stuttgart show a similar pattern. Probabilities to the north-west are over-predicted and those to the north-east are underpredicted, by about 2% or so. However the Alps play a more prominent role in determining the probabilities of exposure to the south-east and south-west of Stuttgart. Three-hour exposure probabilities to the south-west are underpredicted by an amount increasing from 1% to 6% as source-receptor distance increases and trajectories are increasingly diverted around the Alps. Corresponding probabilities to the south-east are underpredicted north of the Austrian Tyrol and overpredicted south of it, because of the blocking and diverting effect the Alps and Tyrol have on puff trajectories. Exposure probabilities for releases of longer duration generally agree well even in cases where there are likely to be big topographic blocking effects.

MESOS and statistically predicted exposure probabilities for wet deposition due to releases from Hannover and Stuttgart are compared in Table 2.6.

Effects such as orographic rainfall lead to a more complicated pattern of MESOS wet exposure probabilities, which do not necessarily fall systematically with distance. This contrasts with the behaviour of probabilities of exposure in all conditions - except where topographic diversion of trajectories is particularly strong, as it is south-west of Stuttgart, for example.

North and west of Hannover, wet exposure probabilities for all release durations tend to be overpredicted by the statistical model - by 1-2% for releases of duration three hours, and by 10-15% for

Receptor	Release duration (hours)					
	3	6	12	24	72	168
1 100 km N	2.7 4.3	4.2 7.0	7.0 11.6	12.7 19.0	27.8 38.9	54.0 61.2
2 200 km N	1.9 4.0	3.2 6.6	5.8 11.0	10.7 18.0	24.4 37.0	45.6 58.8
3 400 km N	1.9 3.5	3.3 5.8	5.2 9.7	9.5 16.0	22.7 33.4	42.1 54.2
4 800 km N	2.0 2.7	3.1 4.5	5.6 7.6	9.5 12.6	21.0 27.0	36.2 45.4
5 100 km E	4.6 4.6	7.5 7.6	13.3 12.6	23.4 20.5	53.0 41.4	80.7 64.2
6 200 km E	5.1 4.3	8.3 7.1	14.1 11.8	23.4 19.3	51.8 39.4	81.2 61.9
7 400 km E	3.5 3.8	6.1 6.3	10.9 10.5	19.8 17.2	45.8 35.7	80.7 57.2
8 600 km E	3.3 3.4	5.1 5.6	8.5 9.3	14.3 15.3	32.5 32.2	55.0 52.6
9 100 km S	2.7 2.1	4.2 3.5	7.0 5.9	12.7 9.8	30.4 21.4	55.5 37.1
10 200 km S	2.3 2.0	3.8 3.3	6.8 5.5	12.3 9.3	25.3 20.3	44.6 35.3
11 400 km S	2.7 1.7	4.4 2.9	7.0 4.9	11.9 8.2	26.5 18.1	48.1 31.8
12 800 km S	0.8 1.3	1.5 2.3	2.8 3.8	4.4 6.4	10.7 14.3	21.8 25.7
13 100 km W	1.8 2.9	3.0 4.8	5.2 8.0	8.7 13.2	21.4 28.2	36.7 47.0
14 200 km W	1.8 2.7	2.5 4.5	4.6 7.5	7.6 12.5	15.0 26.7	27.3 42.9
15 400 km W	1.9 2.4	2.9 3.9	5.2 6.6	8.4 11.0	17.1 23.9	27.8 40.8
16 800 km W	1.9 1.8	2.7 3.1	4.4 5.2	6.4 8.7	14.2 19.1	23.8 33.4

Table 2.6 (a) : Comparison of MESOS and statistically predicted %-probability of exposure at 16 receptor points for several release durations : Hannover wet deposition.

MESOS
 statistical prediction } %-probability

Receptor	Release duration (hours)					
	3	6	12	24	72	168
1 100 km NE	4.6	7.0	11.1	16.7	35.2	62.4
-	4.0	6.6	11.1	18.1	37.3	59.2
2 200 km NE	3.1	4.9	8.2	14.3	29.7	49.1
	3.8	6.3	10.4	17.1	35.5	56.9
3 400 km NE	3.1	4.6	7.4	12.7	27.5	47.6
	3.3	5.5	9.2	15.2	32.0	52.4
4 800 km NE	1.6	2.6	4.6	8.0	18.1	27.8
	2.6	4.3	7.2	12.0	25.8	43.7
5 100 km SE	6.8	10.7	16.3	25.7	54.1	82.7
	3.3	5.5	9.2	15.1	31.8	52.1
6 200 km SE	5.0	8.0	13.3	21.0	44.7	72.8
	3.1	5.2	8.6	14.3	30.2	49.9
7 400 km SE	1.9	3.1	5.4	8.7	19.4	35.2
	2.7	4.6	7.6	12.7	27.1	45.5
8 600 km SE	0.5	0.8	1.4	2.8	7.3	17.4
	2.4	4.0	6.8	11.2	24.3	41.4
9 100 km SW	3.9	5.9	10.1	17.0	37.8	64.9
	2.5	4.2	7.1	11.9	25.5	43.2
10 200 km SW	4.0	6.1	9.9	16.3	33.5	60.0
	2.4	4.0	6.7	11.2	24.1	41.2
11 400 km SW	2.8	4.3	7.4	12.3	28.4	51.5
	2.1	3.5	5.9	9.9	21.6	37.3
12 600 km SW	2.8	4.0	5.8	9.5	20.2	37.7
	1.9	3.1	5.2	8.8	19.3	33.7
13 100 km NW	1.9	3.3	5.8	9.5	21.9	39.7
	3.5	5.8	9.6	15.9	33.2	53.9
14 200 km NW	1.8	3.0	5.2	9.5	23.3	41.6
	3.3	5.4	9.1	15.0	31.5	51.7
15 400 km NW	1.7	3.1	5.2	8.0	19.8	36.7
	2.9	4.8	8.0	13.3	28.3	47.3
16 800 km NW	1.6	2.4	3.8	6.4	13.4	21.3
	2.2	3.7	6.3	10.5	22.8	39.1

Table 2.6 (b) : Comparison of MESOS and statistically predicted %-probability of exposure at 16 receptor points for several release durations : Stuttgart wet deposition.

MESOS
 statistical prediction } %-probability

releases of duration one week. The amount of overprediction depends on the exact siting of the receptor: for short releases exposure probabilities fall little and may rise with distance from the source. This is consistent with the observation in Section 2.3 that puffs tend to spread faster in wet conditions, due to bigger turbulent effects of fronts and depressions - especially for releases from northerly sources. Note also that even if receptors which lie at different distances from the source have the same three-hour exposure probability, their exposure probabilities for releases of longer duration may differ: this is plain, for example, at the receptors 200 km and 400 km north; and 100 km and 200 km west; and 400 km and 800 km west of Hannover. The first and third of these pairs of exposure probabilities show different rates of increase with release duration - though their three-hour exposure probabilities are the same - due to greater clustering of wet exposures to the west of the source, where dispersion episodes leading to contamination in wet conditions are as rare as in the north, but more prolonged.

South and east of Hannover, MESOS and statistically predicted three-hour exposure probabilities are comparable, but within 400 km of the source exposure probabilities for releases of longer duration tend to be underpredicted by up to 20-25% at duration one week because the statistical model does not allow for the weaker effects of clustering in these directions.

Broadly similar comments apply to probabilities for releases from Stuttgart, which again are overpredicted by 1-2% to the north-west. Agreement is good to the north-east. Once again the diverting effect on trajectories of the Alps and Austrian Tyrol is manifest, particularly 600 km south-east of the source. Probabilities south of Stuttgart differ by 2-3% for releases of duration three hours, and by up to 30% - but usually 10-20% - for releases of duration one week.

release duration	Hannover		Stuttgart	
	all	wet	all	wet
3 hours	4	1	4	1
6 hours	3	0	3	1
12 hours	2	0	2	1
1 day	1	0	1	1
3 days	0	0	1	1
1 week	0	0	0	1

Table 2.7: Numbers out of 16 receptors at which MESOS exposure probabilities lie outside their 95% confidence intervals

Table 2.7 shows for the sources Hannover and Stuttgart the numbers out of their 16 receptors at which the MESOS exposure probabilities lie outside their 95% confidence intervals. The intervals for wet exposure probabilities are plainly very good, but those for exposures in all conditions underestimate for short release durations the systematic effect of trajectory blocking by mountains. For longer release duration the intervals are good despite blocking effects.

The comparisons in this section show that little information is lost by using the statistical model to predict exposure probabilities, although at receptors which lie in regions of unusual trajectory divergence due to the effects of mountain barriers there may be systematic effects on probabilities which cannot be taken into account in such a general and simple set of predictive equations. Confidence intervals for the MESOS probabilities for releases of duration up to about twelve hours based on their statistical estimates are accurate except where blocking effects are great, and those for longer release durations are adequate even if blocking is severe.

3. THE DISTRIBUTION OF LEVELS OF EXPOSURE

3.1 Introduction, and exploratory analysis

This chapter of the thesis deals with the analysis of the levels of contamination experienced when a receptor is exposed. Its aims are: to summarize the exposures in the MESOS database in a parsimonious and mathematically tractable form which does not drastically misrepresent the data; then to generalise this by finding explicit ways to represent the effects of changes in the nuclide parameters, source-receptor distance, and release duration on the exposure distributions; and finally to check that the eventual model is adequate for the job for which it is intended. There are roughly 300,000 positive exposures in the database being considered here: six sources each with sixteen receptors each experiencing an average 200 positive exposures to each of twelve combinations of nuclide and exposure mode. The most obvious - and possibly the only - way to summarize the data is to fit a suitable parametric distribution to the exposure distributions at each receptor, and then to explore how the parameters of the fitted distributions themselves vary. It is over-optimistic to hope that a two-parameter distribution - or one with three or more parameters - will give a uniformly good representation of exposures which depend on a multitude of varying factors. However a single family of distributions has to be used so that the work is coherent: the effects on exposure levels of altering nuclide parameters, release duration, and source-receptor distance must be capable of explicit representation.

Typically there are at least three steps to carry out when it has been decided to fit probability distributions on an empirical basis to large datasets of any provenance. These are:

- (a) exploration of the data and its properties and a comparison

of different distributions to assess their suitability;

(b) choice and fitting of a suitable distribution;

(c) confirmation that the chosen distribution is indeed

appropriate, and in particular that such departures as may occur will not materially alter any conclusions resulting from the analysis.

The division is somewhat arbitrary since the results of (c) may demand that work begins again at (a) or that (b) be reconsidered. Often the last two phases may be carried out together. However there is a clear conceptual distinction between the open-endedness of (a) and the narrower outlook of (c). This section concentrates on (a) and (b), and the next deals with (c).

Many parametric probability distributions have been proposed for fitting to air pollution data, partly because physical considerations do not seem to favour any particular distribution uniquely, and partly because of the multitude of different types of data and reasons for the fitting. Since such data are by their nature the result of many complex phenomena, it does not seem reasonable to expect that any particular distribution will describe them all uniformly better than all its competitors.

There is a large body of literature dealing with the frequency distributions of air pollutants. Georgopoulos and Seinfeld(1982) give a critical review which concentrates on pollution from areal sources, discuss the fitting of several distributions to such data, and give examples of their use. The same eclectic approach is taken by Holland and Fitz-Simons(1982), who describe a computer program for fitting and assessing the fit of the Normal and three-parameter log-Normal, the three-parameter gamma and Weibull, and the Beta and Johnson S_B distributions. Pollack(1975) discusses concentration frequency distributions for point sources, and derives the log-Normal or log-chi-squared distributions for them - depending on the

assumptions made - by a mixture of theoretical and empirical arguments. However these arguments apply to time-averaged rather than time-integrated air concentrations. The distinction is subtle: a three-hourly time-averaged concentration is the average concentration observed at a receptor over a three-hour period - perhaps due to the passage of a number of different puffs - whereas a three-hourly time-integrated concentration is that due to a single puff released over a period of three hours, integrated over the period the puff takes to cross the receptor, whatever its length. Data for continuous releases, whether from point or area sources, are almost by definition time-averaged, since the time at which the pollutant was released cannot usually be determined from its concentration profile at the receptor.

In the absence of either a direct link between these different types of data or compelling physical arguments in favour of a particular distribution for time-integrated air concentrations, an empirical approach to the MESOS data is adopted.

A plot which distinguishes distributions of different types, based on comparison of their lower-order moments, is described by Cox and Oakes(1984). Let $f(y)$ be the probability density of a positive variable. Its variance and skewness are

$$\mu_2 = \int_0^{\infty} (y-\mu_1)^2 f(y) dy$$

and

$$\mu_3 = \int_0^{\infty} (y-\mu_1)^3 f(y) dy$$

respectively, where $\mu_1 = \int y f(y) dy$ is the mean of the distribution. Then the coefficient of variation γ and standardized skewness γ_3 of the distribution are $\sqrt{\mu_2/\mu_1}$ and $\mu_3/\mu_2^{3/2}$. From dimensional considerations they do not depend on the scale of the distribution, but only its shape, so that comparison of them for different

distributions provides an idea of their relative shapes: γ summarises the spread of the distribution relative to its mean, and γ_3 its degree of asymmetry. For example γ and γ_3 are respectively: 1 and 2 for the exponential distribution; $1/\sqrt{\alpha}$ and $2/\sqrt{\alpha}$ for the two-parameter gamma distribution with shape parameter α ; and $\gamma = \sqrt{w-1}$ and $\gamma_3 = (w+2)\sqrt{w-1}$ for the log-Normal distribution with $w = \exp(\sigma^2)$ where σ is the shape parameter of the distribution.

Figure 3.1 shows γ_3 plotted against γ for a few distributions. The Weibull and gamma lines intersect at the exponential distribution. The general qualitative picture to emerge is that the log-logistic distribution is most skewed of all for a given value of γ , and that the gamma and Weibull distributions are less skewed depending on the value of γ . Only the Weibull is capable of negative skewness.

The idea now is that the sample values

$$C = \frac{\sqrt{\{(n-1)^{-1} \sum (Y_i - \bar{Y})^2\}}}{\bar{Y}}$$

and

$$C_3 = \frac{n^{-1} \sum (Y_i - \bar{Y})^3}{\left\{ (n-1)^{-1} \sum (Y_i - \bar{Y})^2 \right\}^{3/2}}$$

of γ and γ_3 be found - here \bar{Y} is the mean of the simple random sample Y_1, \dots, Y_n - and plotted on such a graph.

Figure 3.2(a) shows the plot for the sixteen receptors for Mol for which dry deposition exposures due to three-hourly releases of $I_{131}(g)$ are available. Most of the points cluster closely together, but the four closest to Mol have higher skewnesses due to a number of individually higher observations at the receptors. The high values do not seem aberrant in this case. The equivalent plot for Ispra $I_{131}(p)$ wet deposition is displayed in Figure 3.2(b). Here random scatter is larger because there are fewer observations in each

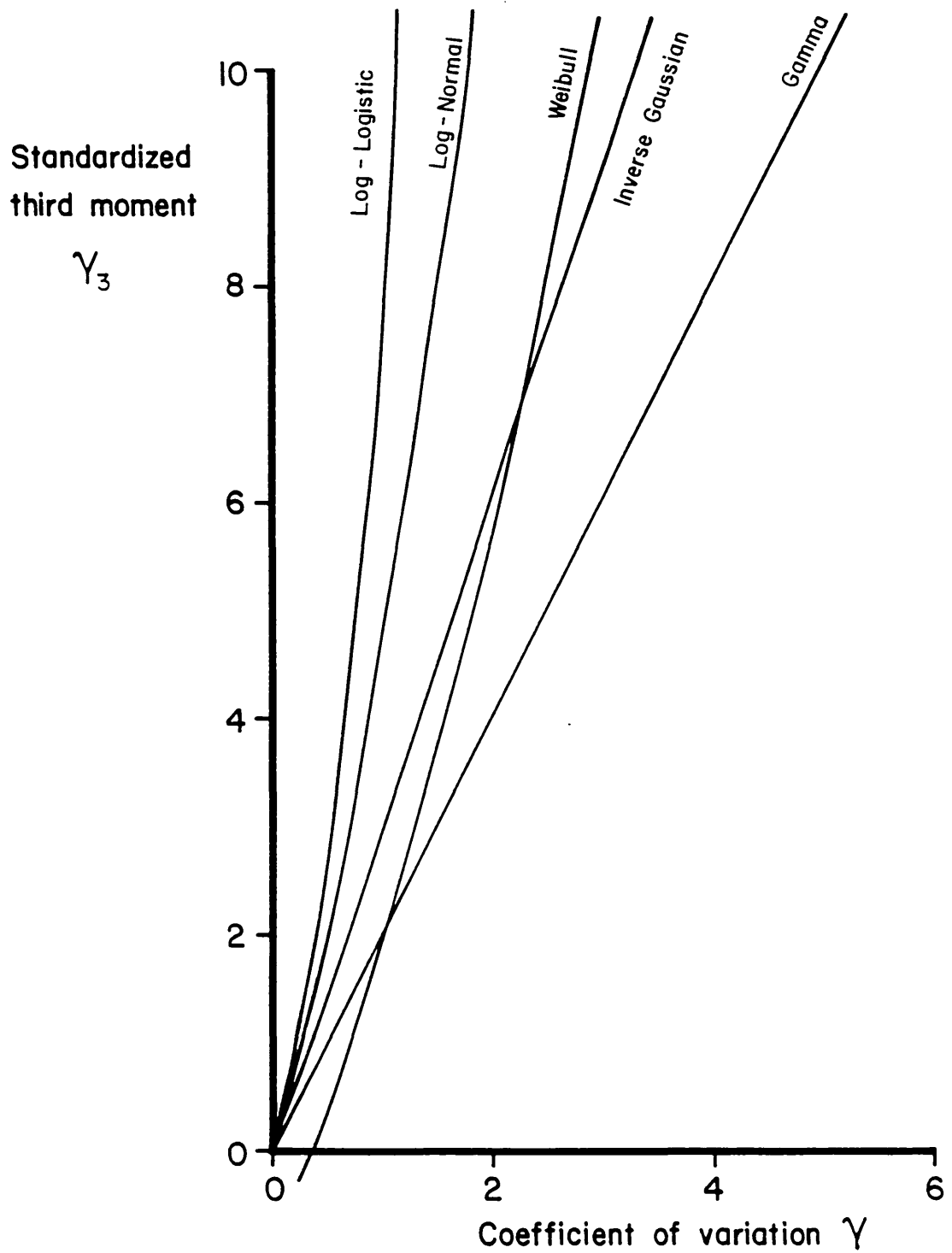
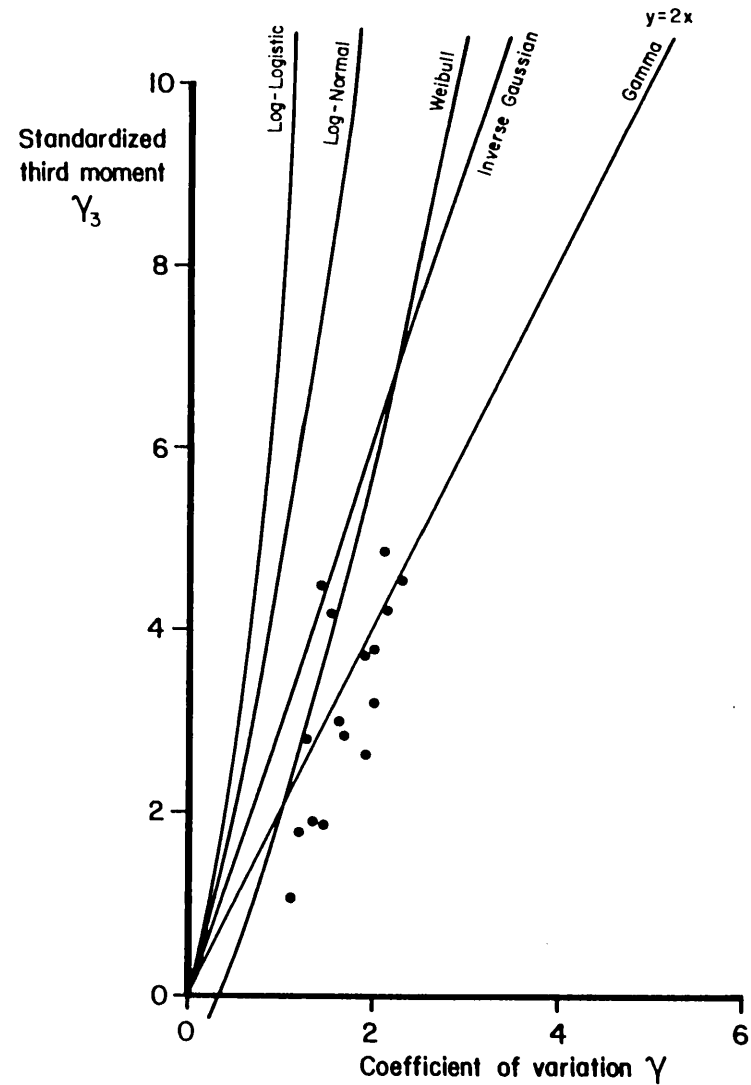
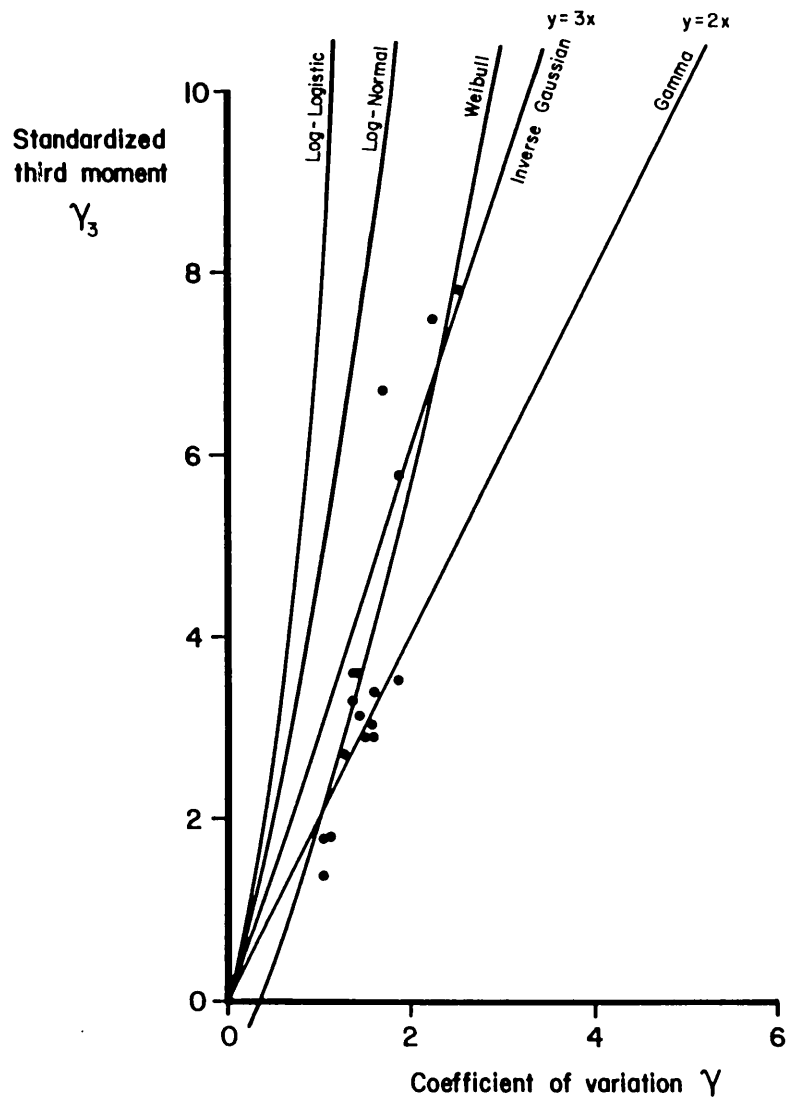


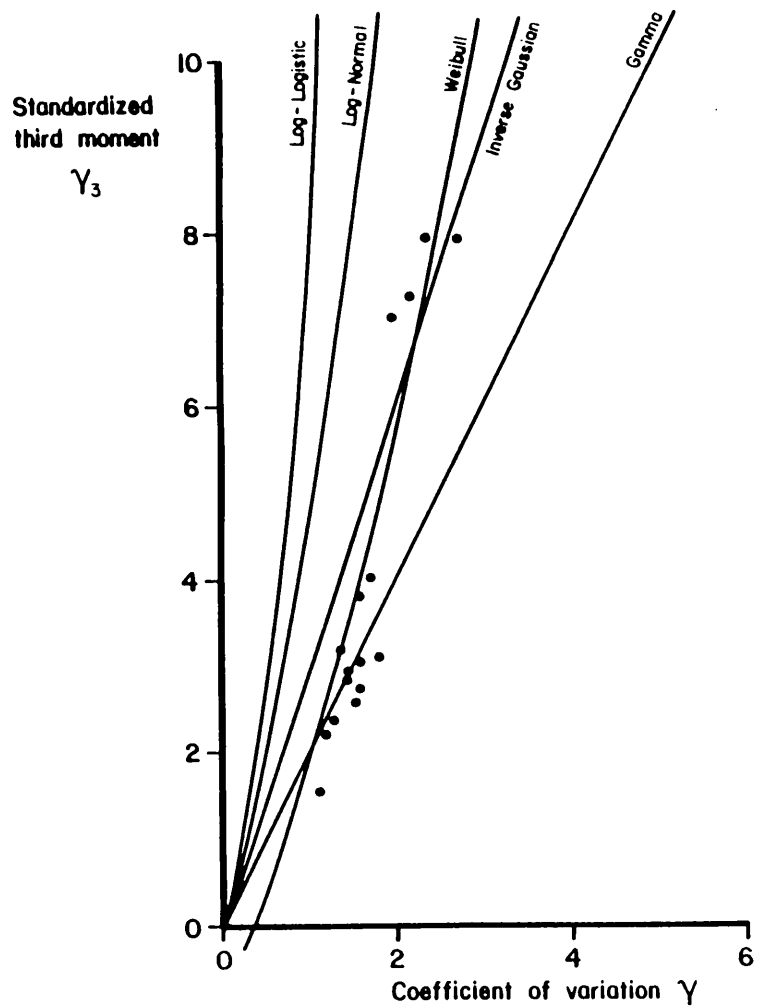
Figure 3.1 : Comparison of moments for different distributions.



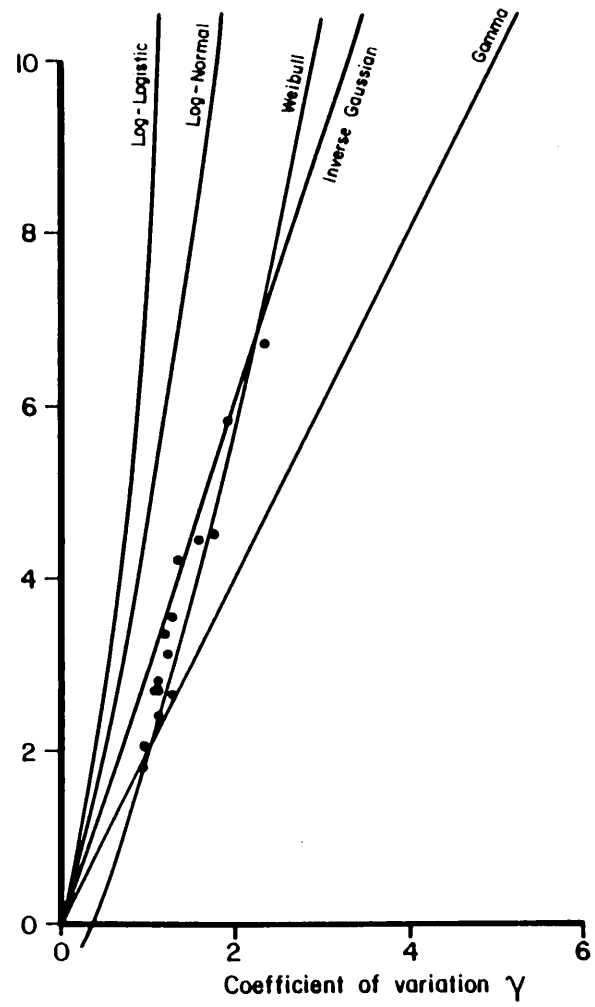
(a) Mol 1976, I_{131} (a) dry deposition.

(b) Ispra, I_{131} (p) wet deposition.

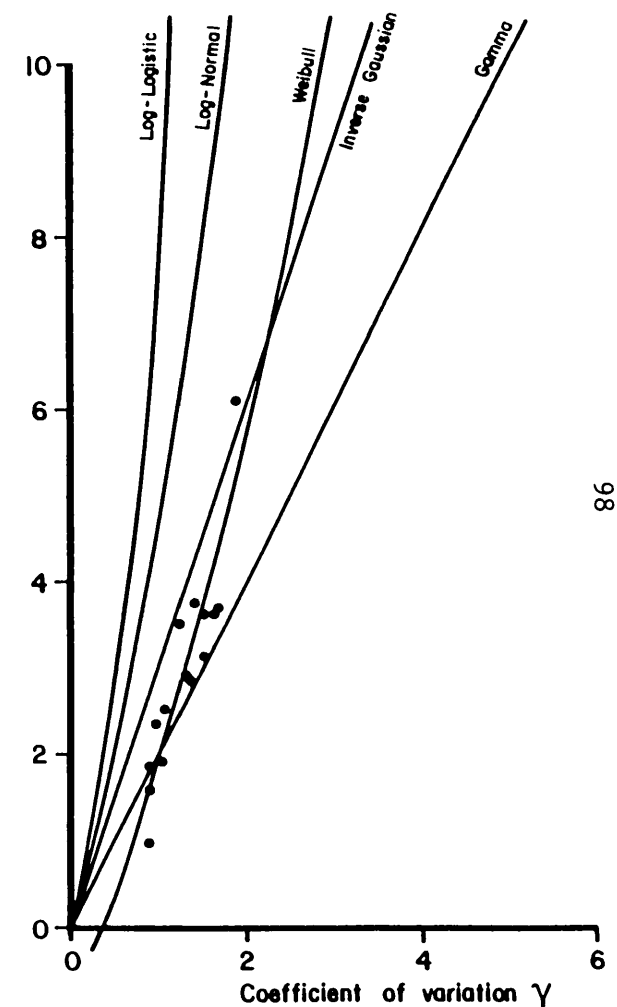
Figure 3.2 : Standardized moment plots for exposure distributions due to three-hour releases.



(c) Mol 1976, $I_{131}(p)$ air contamination.



(d) Mol 1973, Kr_{85} air contamination.



(e) Karlsruhe, Xe_{133} air contamination.

Figure 3.2 : Standardized moment plots for exposure distributions due to three-hour releases.

dataset. Both these graphs and the others in Figure 3.2 show that the Weibull or possibly the gamma distributions offer the most promising fit to the data.

Three points should be noted: the sample values of γ and γ_3 will usually provide biased estimates of their true values; they may have big variances; and sample moments are rather sensitive to outlying observations - although this may prove useful for an exploratory procedure since it can alert the user to oddities of the datasets. This means that such plots should be interpreted with caution, especially if there are only a few sets of data or if they themselves are small. However they provide a good map for the exploration of big databases.

Figure 3.3, taken from Wrigley(1982), shows histograms of exposure data for both three-hourly and daily unit releases of pollutants from several sources. They have a variety of shapes, but many have an extended lower tail, which suggests that the log-Normal and log-logistic distributions are unlikely to fit them well. The Normal distribution is a non-starter. The effect of decay on the exposures is evident from a comparison of Figures 3.3(a) and 3.3(b): the lower Xe_{135} exposures are smeared to the left by up to three orders of magnitude; and because of the separation between exposures from direct and indirect trajectories the histograms tend to be bimodal, especially close to the source. The histograms for the daily releases are very similar, which suggests that a distribution suitable for exposures from short-term releases is likely to fit those from longer ones.

Wrigley(1982) found that the Weibull distribution

$$P(Y \leq y) = 1 - \exp\{ -(y/\mu)^\alpha \}$$

$$(y > 0; \alpha, \mu > 0)$$

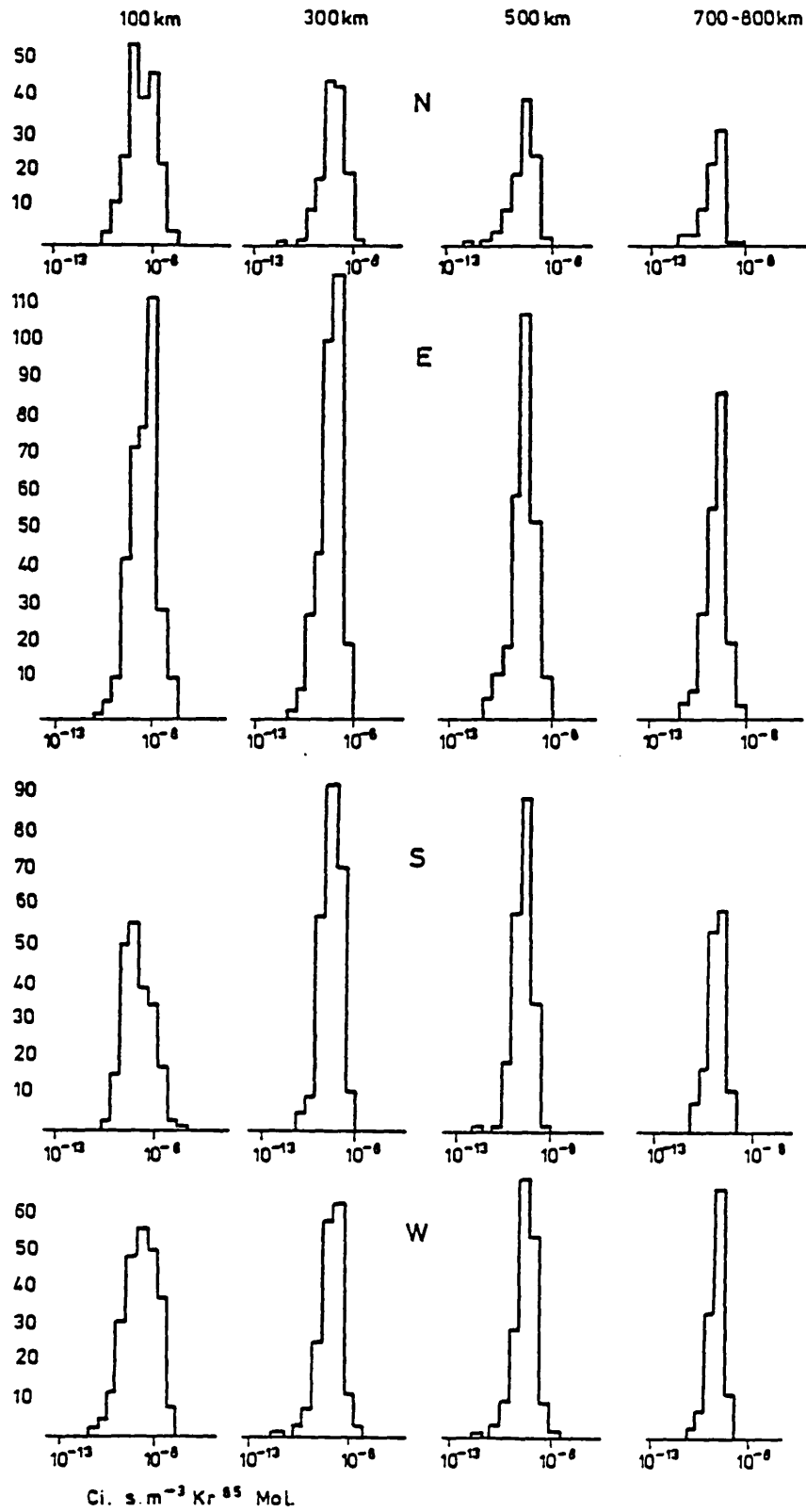


Figure 3.3 : Histograms of exposure distributions.

(a) time-integrated air concentrations due to three-hourly unit releases of Kr_{85} from Mol during 1973.

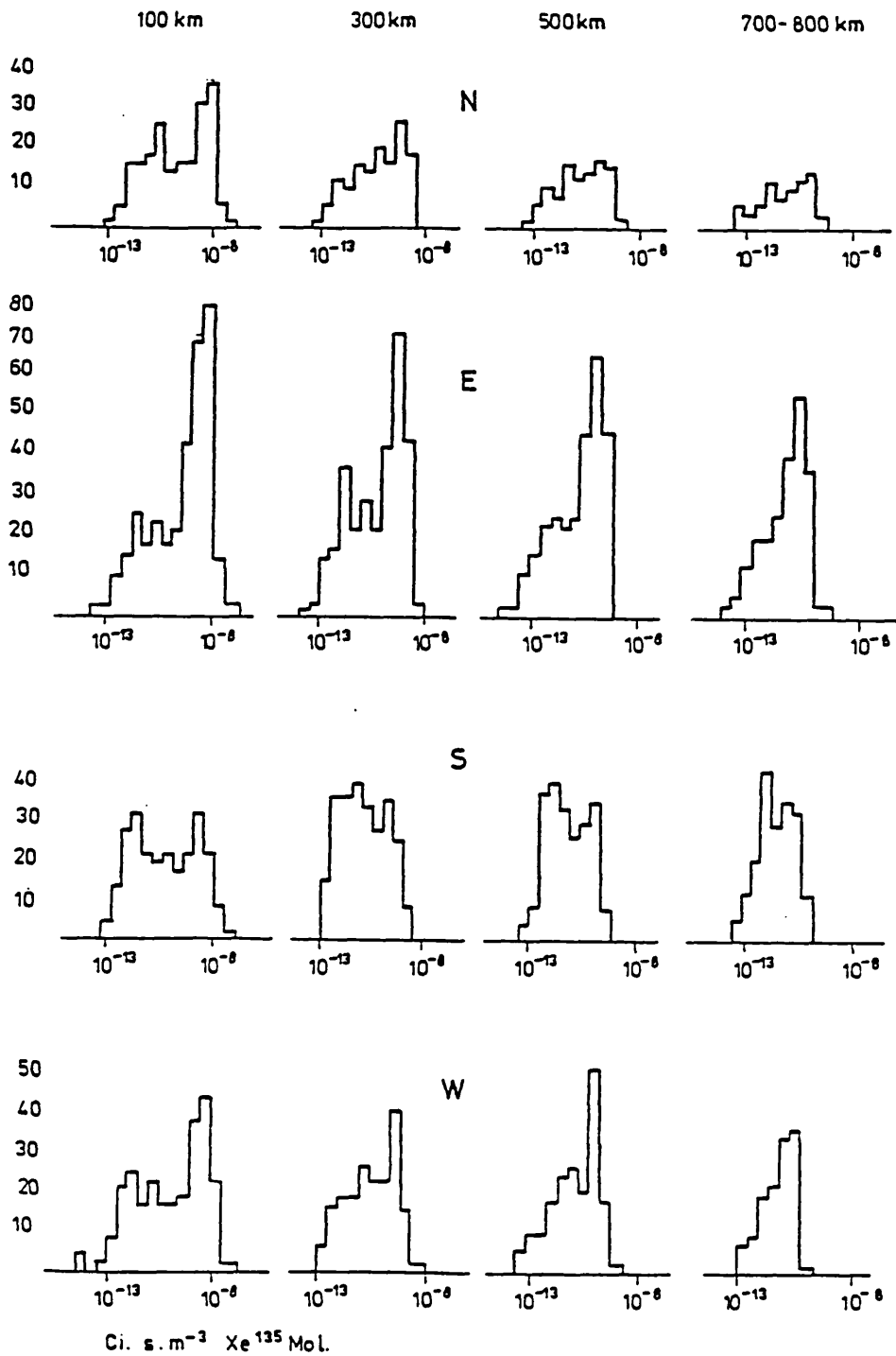


Figure 3.3 : Histograms of exposure distributions.

(b) time-integrated air concentrations due to three-hourly unit releases of Xe_{135} from Mol during 1973.

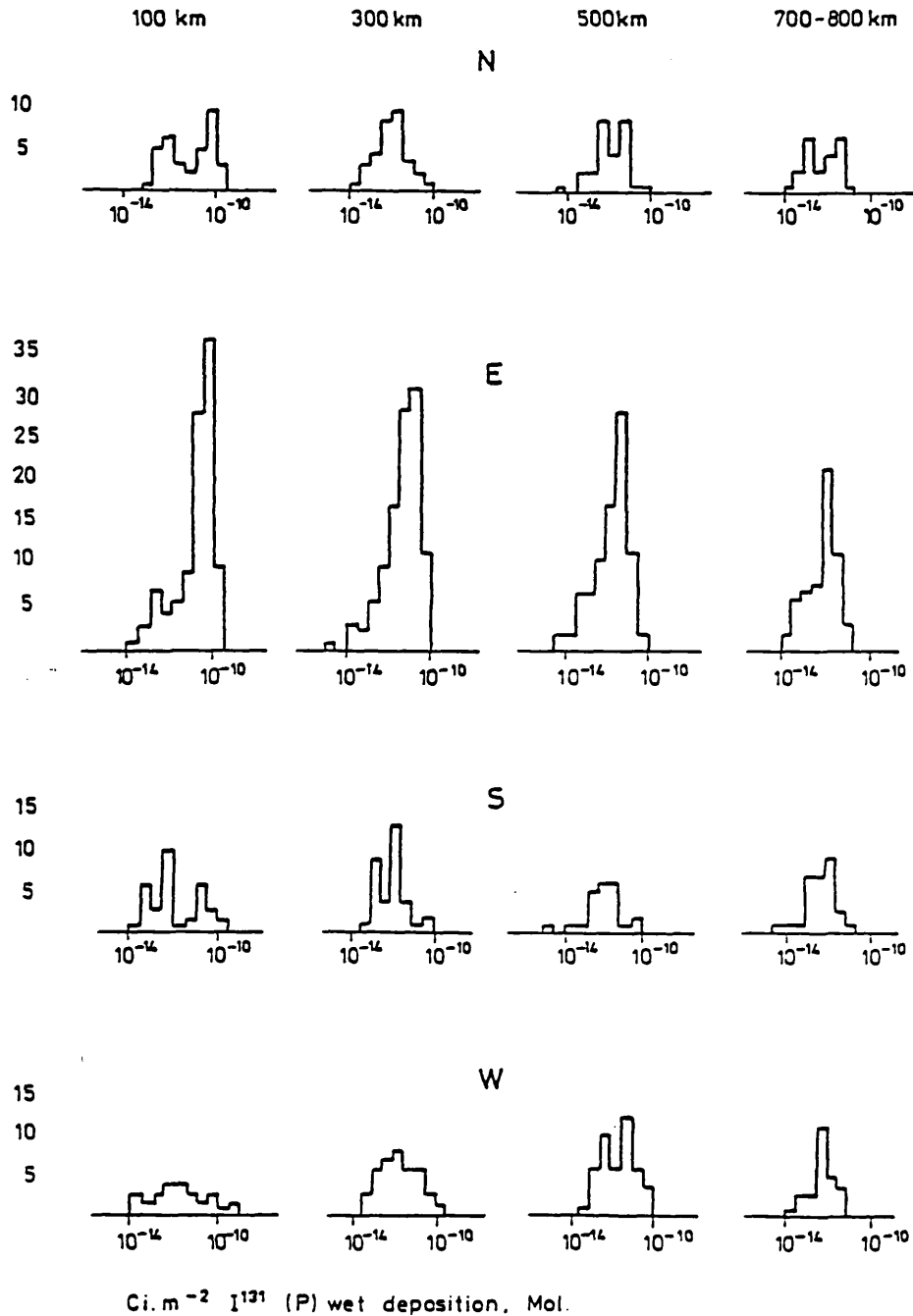


Figure 3.3 : Histograms of exposure distributions.

(c) wet deposition due to three-hourly unit releases of I_{131} (P) from Mol during 1973.

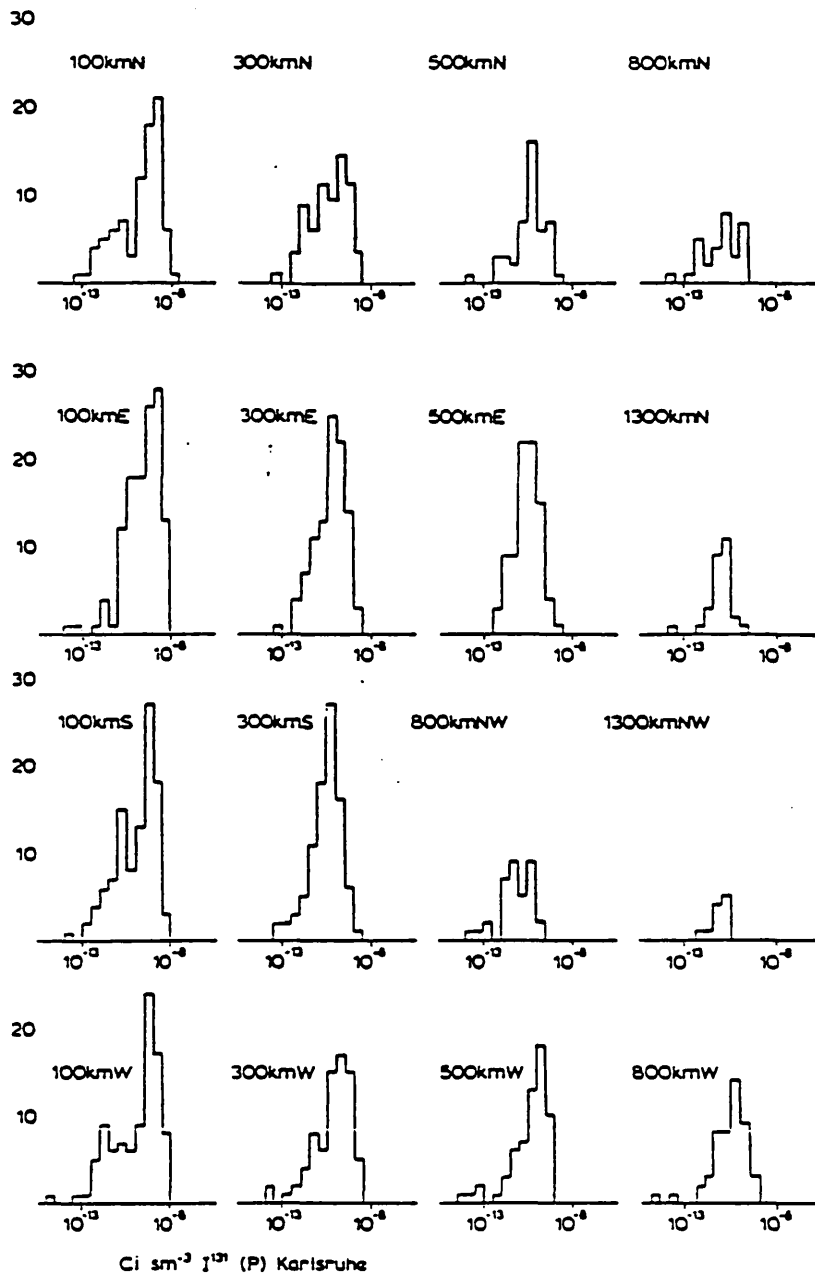
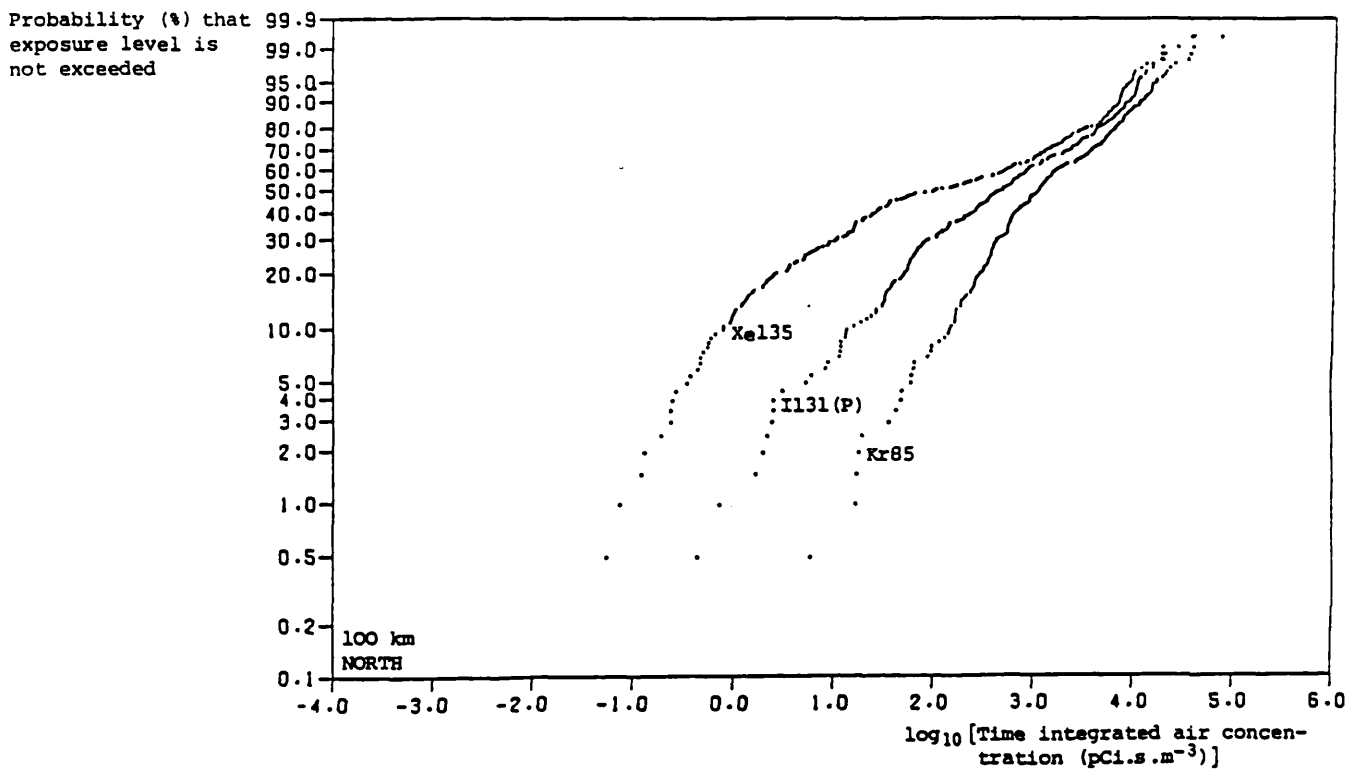


Figure 3.3 : Histograms of exposure distributions.

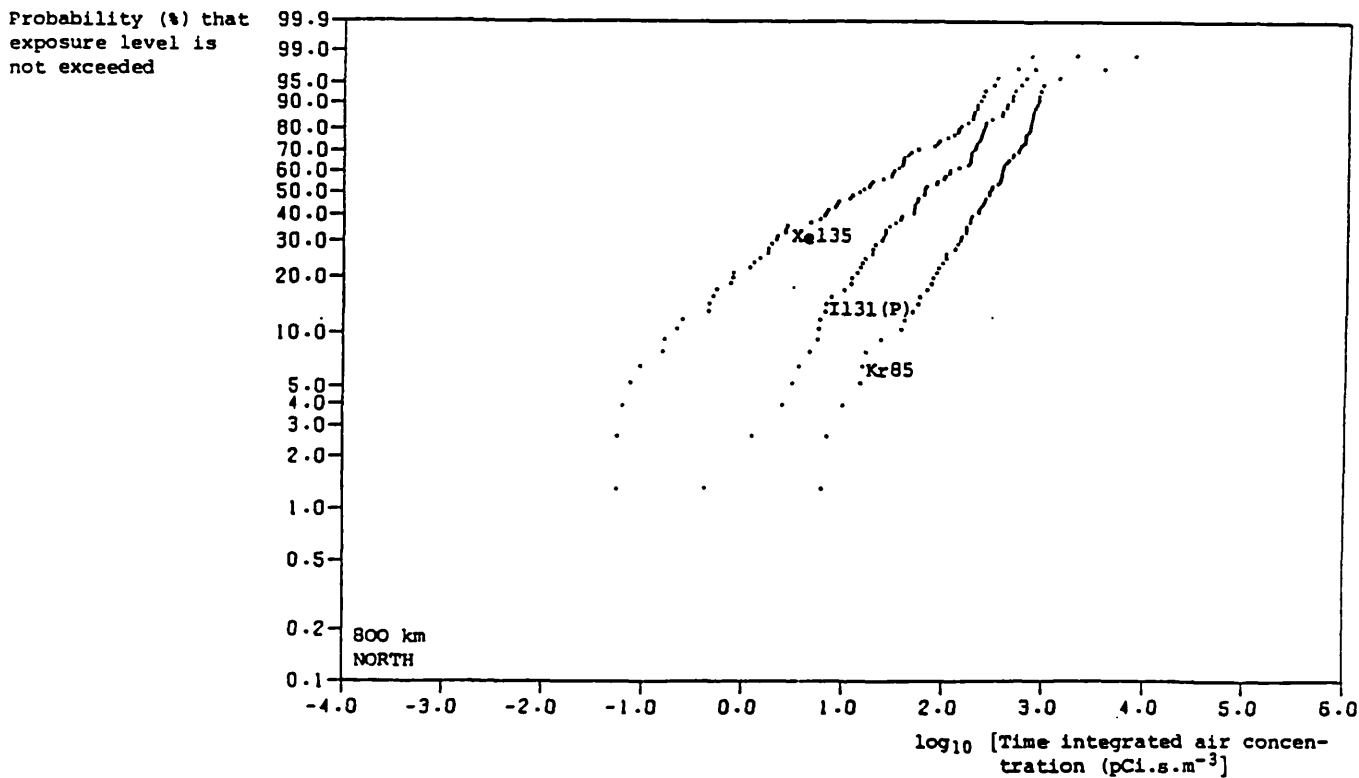
(d) time-integrated air concentrations due to daily unit releases of $\text{I}_{131}(\text{p})$ from Karlsruhe during 1973.

was appropriate for MESOS exposure distributions, particularly for slowly depleting nuclides such as Kr_{85} , and that its fit generally improved with distance from the source. Figure 3.4 shows ordered exposures plotted on a Weibull scale. Those for $I_{131}(p)$ and Kr_{85} are not too far from linearity, but that for Xe_{135} 100 kilometres north of Mol strikingly shows the bimodality of the exposure distribution. Upper outliers are obvious in Figure 3.4(b). The largest two arise from the same dispersion incident on 26th October 1973. High exposures are considered from a statistical viewpoint in Chapter 8.

As a result of this exploratory work, Weibull distributions were fitted by maximum likelihood to a large number of sets of MESOS exposure data, using the GLIM algorithm described by Aitken and Clayton(1980). Some of the parameter estimates are shown in Table 3.1. The estimated values of α are almost all less than one, indicating that the exposure distributions are more spread out than the exponential. This is consistent with Figure 3.2. Comparison of the estimates for the Kr_{85} results for Mol in both years shows that the shape parameters are generally quite close, but that the scale parameters μ tend to be higher in 1976 by a factor of up to 1.4. The values of α for the depleting nuclides are lower than for Kr_{85} because the distributions are more extended. The shape parameter estimates for the Xe_{135} distributions are very low indeed and reflect the marked effect of its short half-life on exposures.



(a) 100 km North of Mol.



(b) 800 km North of Mol.

Figure 3.4 : Cumulative frequency distributions of nuclide exposures for unit releases over three hours from Mol during

Receptor	Mol 1973 Kr ₈₅ air		Mol 1976 Kr ₈₅ air		Mol 1976 Xe ₁₃₅ air	
	$\hat{\alpha}$	$\hat{\mu} \times 10^3$	$\hat{\alpha}$	$\hat{\mu} \times 10^3$	$\hat{\alpha}$	$\hat{\mu} \times 10^4$
1	.63	2.99	.66	3.67	.35	7.16
2	.84	1.38	.78	1.81	.44	1.76
3	.89	.672	.78	1.02	.45	.673
4	.80	.420	.83	.579	.54	.228
5	.71	3.45	.76	4.99	.45	15.0
6	.93	1.15	.85	1.48	.51	2.38
7	.95	.691	.86	1.01	.54	.910
8	.92	.506	.92	.506	.61	.284
9	.59	2.22	.67	4.11	.38	7.45
10	.96	.966	.82	1.50	.44	1.56
11	1.06	.646	.94	.964	.51	.461
12	1.18	.410	1.05	.627	.57	.154
13	.62	3.76	.65	3.77	.34	6.83
14	.85	1.39	.81	1.61	.42	1.31
15	.91	1.08	.82	.983	.46	.469
16	1.33	.600	1.14	.652	.56	.347

Table 3.1: Estimated Weibull parameters for MESOS exposure distributions

Units for $\hat{\mu}$ are $\text{Cism}^{-3} \times 10^{-6}$

3.2 Confirmatory analysis

The process of elimination which arrived at the Weibull distribution for MESOS exposures was described in the previous section. Here its adequacy as a summary of MESOS three-hour exposure distributions is carefully checked. Six datasets of exposures at sixteen receptors each are used in order to cover a wide variety of nuclides and sources: air contamination data for releases of Kr₈₅ from Heysham, Xe₁₃₃ from Karlsruhe, Xe₁₃₅ from Cadarache, and I₁₃₁(p) from Mol through 1973; dry deposition for I₁₃₁(g) released from Mol throughout 1976; and wet deposition for I₁₃₁(p) released from Ispra. At each of the sixteen receptors the two-parameter Weibull distribution fitting the data 'best' - the maximum likelihood estimate - is compared with the exposure distribution itself, both visually and using a sensitive statistical test of fit.

It was stated in Chapter 1 that it is important to be sure that any discrepancies between fitted and MESOS distributions are not physically important, whatever their statistical size. For this reason the visual comparison of the distributions is more relevant here than the calculation of a statistical test, however powerful or useful it may be: a single number cannot contain the information in a pertinent plot. Moreover, since real observations will never be exactly distributed according to any mathematical formula, some lack of fit will always be found in very large datasets. A test statistic provides a rule of thumb rather than a strict prescript. However it may give valuable help in seeking out inadequacies of fit, then assessed graphically.

Stephens(1977) gives significance points of tests for the goodness of fit of the Gumbel distribution

$$H(y) = \exp\{ -\exp(- \tau(y-\zeta)) \}$$

Receptor	Heysham Kr ₈₅ air	Karlsruhe Xe ₁₃₃ air	Mol 1973 I ₁₃₁ (p) air	Mol 1976 I ₁₃₁ (g) dry	Cadarache Xe ₁₃₅ air	Ispra I ₁₃₁ (p) wet
1	.21	.93*	1.48°	2.80°	4.10°	.29
2	.65	.42	.59	.69	1.80°	.37
3	.83*	.85*	.39	1.45°	1.65°	.49
4	.35	.28	.40	1.08°	.58	.69
5	.87*	.92*	2.97°	2.09°	2.38°	.36
6	.89*	.76	2.79°	.51	1.43°	.35
7	.76*	1.46°	2.51°	.30	4.47°	.29
8	.47	1.05°	.48	.55	1.85°	.49
9	.39	3.20°	1.80°	1.01*	1.37°	.18
10	.48	1.73°	.40	.48	8.87°	.96*
11	.30	1.64°	.57	.64	3.02°	.26
12	.64	.73	.72	.34	3.51°	.64
13	.28	.75	1.32°	1.10°	1.48°	.44
14	2.23°	.89*	.30	.40	1.63°	.59
15	1.33°	.60	.95*	.76*	2.10°	.67
16	.49	1.11°	1.09°	.29	.53	.55

Table 3.2: Anderson-Darling statistics for exposure data

(a) release duration three hours

* significance level between .05 and .01

° significance level less than .01

Receptor	Heysham Kr ₈₅ air	Karlsruhe Xe ₁₃₃ air	Mol 1973 I ₁₃₁ (p) air	Mol 1976 I ₁₃₁ (g) dry	Cadarache Xe ₁₃₅ air	Ispra I ₁₃₁ (p) wet
1	.26	.23	.64	.70	1.53°	.27
2	.22	.24	.28	.23	.81	.54
3	.55	.92*	.29	.31	1.00*	.48
4	.48	.29	.61	.47	1.00*	.45
5	.50	.77*	1.39°	.40	1.08°	.49
6	.49	.85*	1.09°	.46	.44	.23
7	.22	.86*	1.68°	.22	2.11°	.19
8	.30	.58	.45	.48	.94*	.68
9	.30	.51	.59	.46	1.05°	.49
10	.38	.55	.61	.37	4.81°	.29
11	.50	.64	.38	.41	1.27°	.29
12	.28	.69	.50	.39	2.00°	.48
13	.62	.47	.88*	.43	.53	.56
14	.94*	.48	.21	.46	.97*	.34
15	.66	.69	.25	.33	.81*	1.19°
16	.64	.74	.23	.97*	.34	.41

Table 3.2: Anderson-Darling statistics for exposure data

(b) release duration one day

* significance level between .05 and .01

° significance level less than .01

Receptor	Heysham Kr ₈₅ air	Karlsruhe Xe ₁₃₃ air	Mol 1973 I ₁₃₁ (p) air	Mol 1976 I ₁₃₁ (g) dry	Cadarache Xe ₁₃₅ air	Ispra I ₁₃₁ (p) wet
1	.19	.27	.24	.27	.89*	.52
2	.47	.48	.27	.39	.58	.46
3	.25	.96*	.40	.25	.65	.25
4	.41	.20	.24	.26	.64	.27
5	.29	.21	.80*	.50	.64	.35
6	.25	.36	.90*	.28	.70	.27
7	.37	.09	1.13°	.35	1.48°	.52
8	.36	.29	.69	.27	.89*	.61
9	.69	.27	.50	.19	.36	1.04°
10	.33	.31	.31	.63	2.17°	.83*
11	.42	.34	.22	.36	.59	.41
12	.28	.38	.64	.27	1.30°	.25
13	.56	.17	.77*	.39	.46	.56
14	1.04*	.77*	.80*	.31	.87*	.39
15	.39	.41	.23	.42	.39	.85*
16	.36	.37	.46	.44	.20	.28

Table 3.2: Anderson-Darling statistics for exposure data

(c) release duration one week

* significance level between .05 and .01

° significance level less than .01

$$(y \in \mathbb{R} ; \tau > 0 , \zeta \in \mathbb{R})$$

with scale and location parameters τ and ζ , to simple random samples, based on their empirical distribution functions. The tests he describes are all based on the fact that if the random variable Y has a Gumbel distribution, then $H(Y)$ is uniformly distributed in the unit interval. Hence the transformed values $U_i = H(Y_i)$ of a simple random sample Y_i ($i=1, \dots, n$) form a sample of size n from the unit uniform distribution, and a variety of tests of their uniformity can be constructed, although they are not directly applicable in cases where the parameters τ and ζ are unknown. However if estimates $\hat{\tau}$ of τ and $\hat{\zeta}$ of ζ are available, the 'configuration' of the sample Z - whose n elements are $Z_i = \hat{\tau}(Y_i - \hat{\zeta})$, - is invariant with respect to the parameters ζ and τ . That is, it is independent of their true values, and hence so are the distributions of tests based on Z for the 'Gumbelness' of the Y_i . This is essentially the same as noticing that the Y_i should lie close to a straight line on Gumbel plotting paper, and testing for this independently of its slope and intercept. The tests may be extended immediately to the Weibull distribution by using the fact that if Y is Gumbel with parameters τ and ζ , then $\exp(-Y)$ is Weibull with scale and shape parameters τ and $\exp(-\zeta)$. Stephens gives the significance points of a number of closely related tests, but the one used here is the Anderson-Darling statistic

$$-n^{-1} \sum_{i=1}^n \{ (2i-1) \log U_{(i)} + (2n-2i+1) \log(1-U_{(i)}) \},$$

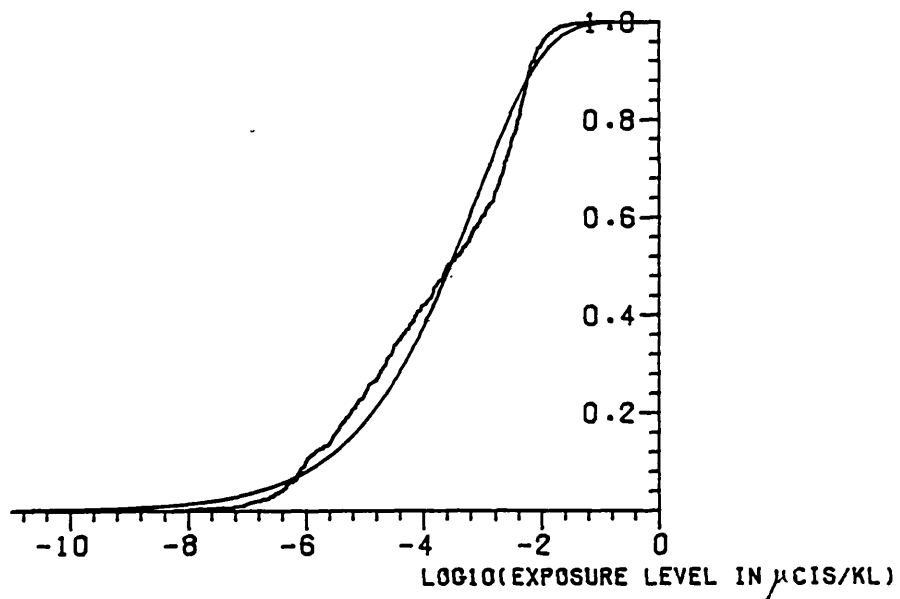
where $U_{(i)}$ are the ordered values of the $U_i = \exp\{ -\exp(-Z_i) \}$, which is equally sensitive to departures from 'Weibullness' throughout the range of Y .

Table 3.2 shows for the data described above the numbers of these statistics significant at various probability levels, for unit releases of duration three hours, one day, and one week. Were the

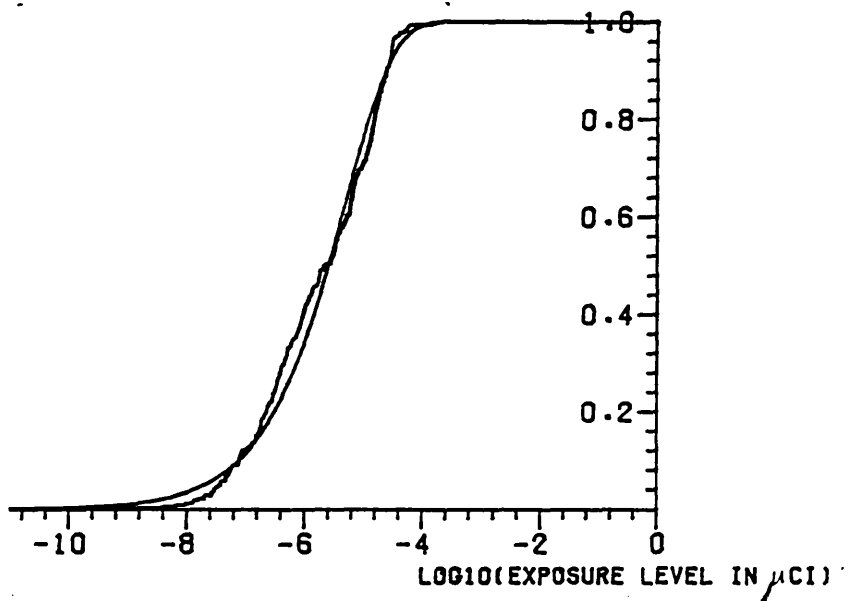
data Weibull, roughly 5 - rather than the observed 46 - of the 96 statistics in Table 3.2(a) would be significant at the 5% level, and 1 - rather than 35 - at the 1% level. The fit is clearly quite poor, nor does it improve much with distance except at ranges of 1000 kilometres or more. The Weibull fit to the Xe_{135} data is worst, and to the wet deposition data best. A poor fit to the Xe_{135} exposures was expected because of their bimodality.

With slight variations the same story is told by Tables 3.2(b) and (c). The fit improves with release duration, but 23 and 17 of the sets of 96 statistics in them are significant at the 5% level or lower. Other test statistics show a similar pattern: the Weibull density does not generally fit well from a statistical point of view.

These departures are more informatively displayed in Figure 3.5, which shows the observed and fitted distributions of the data on a logarithmic scale. The three-hourly exposure Xe_{135} and $I_{131}(g)$ datasets in Figure 3.5(a) and (b) were chosen because their fit as judged by the Anderson- Darling statistics is poor, and those in 3.5(c) and (d) because their statistics give the opposite impression. The first two have some common features compared with their Weibull approximations: the lower tail of the Weibull distribution is too long; from about the 10% level to the median the Weibull exposure levels are too high by a factor of up to about 1.5; between the median and the 90% level the Weibull levels are too low by a factor at most 1.5; and then the Weibull exposure levels are too high in the top 5-10% of the data. Divergence is most marked in the lower tails, where below the 5% probability level the ratio of observed to fitted levels exceeds two. The upper 70% of the exposure levels lie within a factor 1.5 of each other, except in the extreme upper tail where the Weibull is a conservative approximation. This is typical of the worst-fitted datasets analyzed.

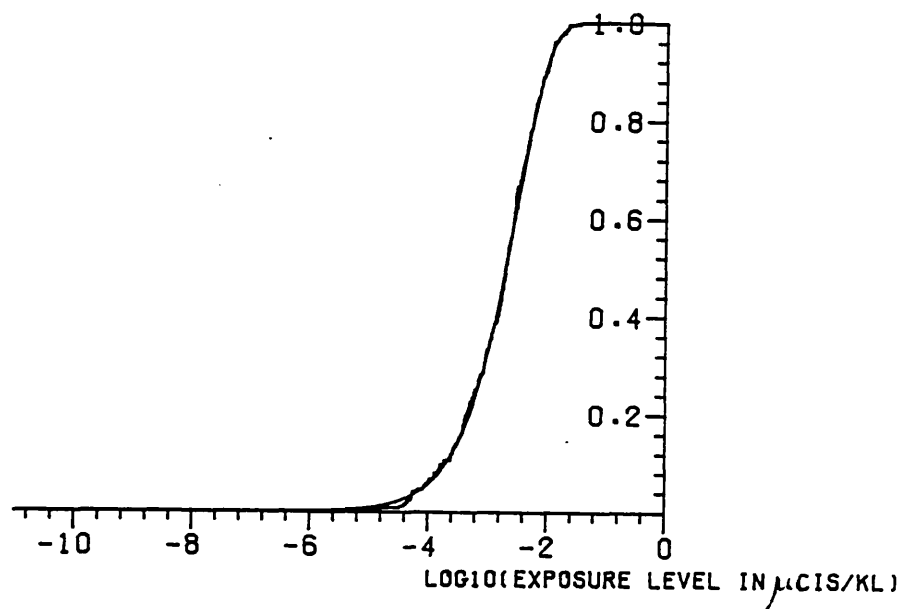


- (a) Xe_{135} air contamination Cadarache receptor 10, release duration 3 hours.

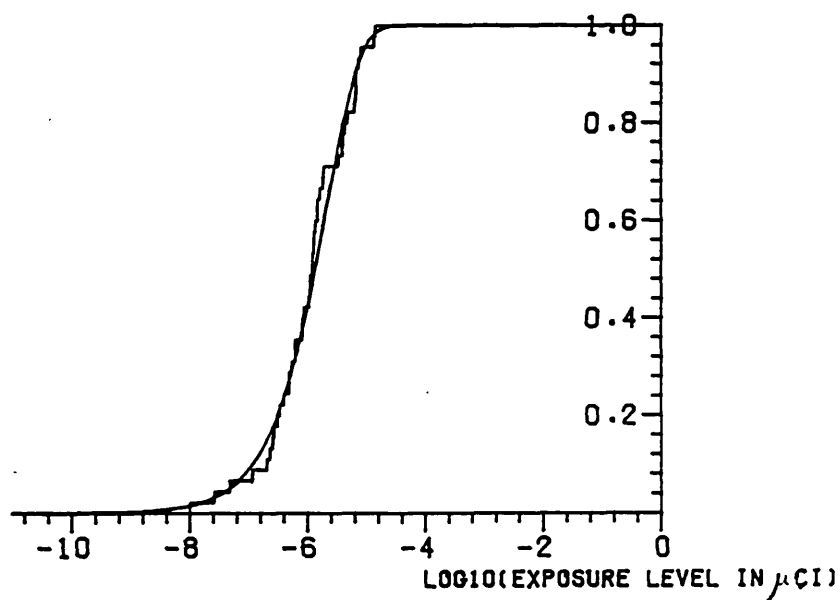


- (b) $\text{I}_{131}(\text{g})$ dry deposition Mol 1976 receptor 1, release duration 3 hours.

Figure 3.5 : Comparison of MESOS and fitted Weibull cumulative exposure distributions.

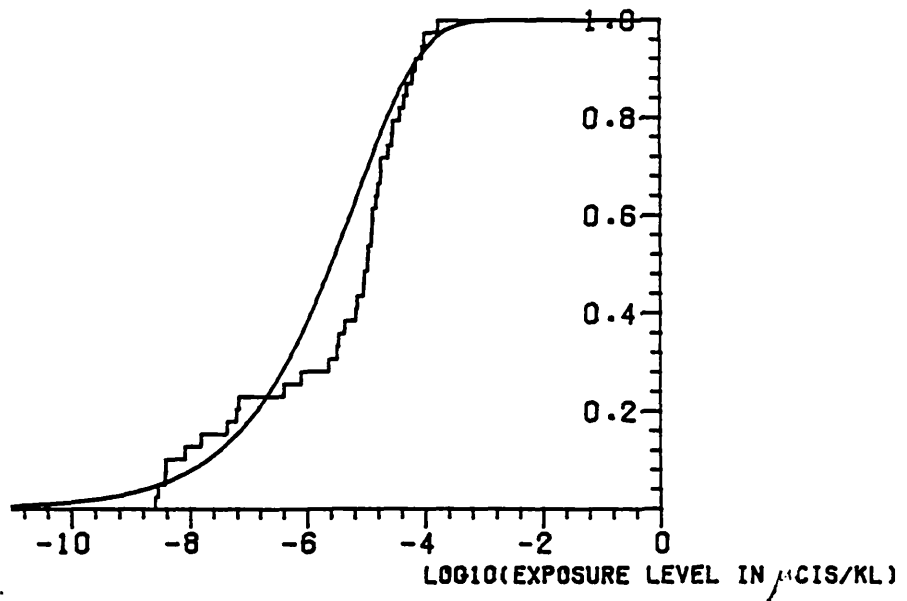


(c) Kr_{85} air contamination Heysham receptor 1, release duration 3 hours.

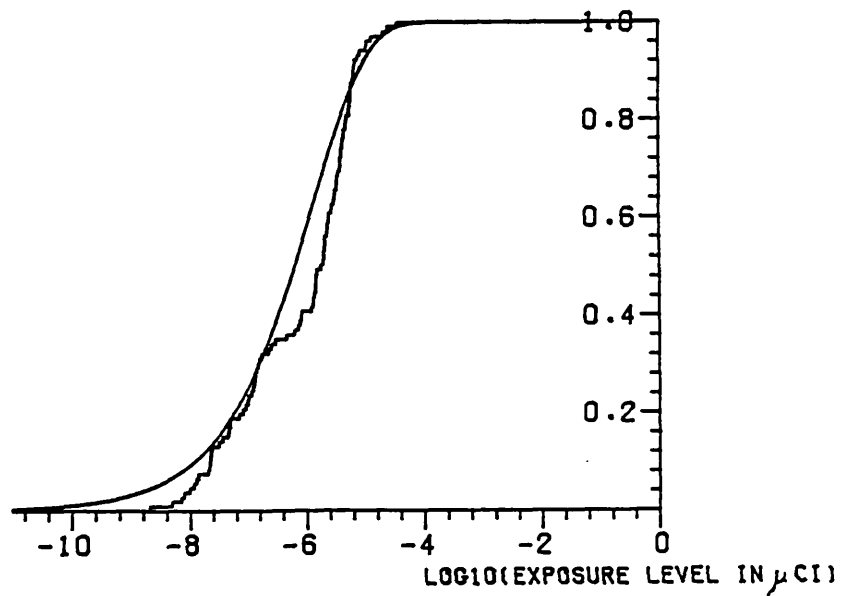


(d) $I_{131}(p)$ wet deposition Ispra receptor 8, release duration 3 hours.

Figure 3.5 : Comparison of MESOS and fitted Weibull cumulative exposure distributions.

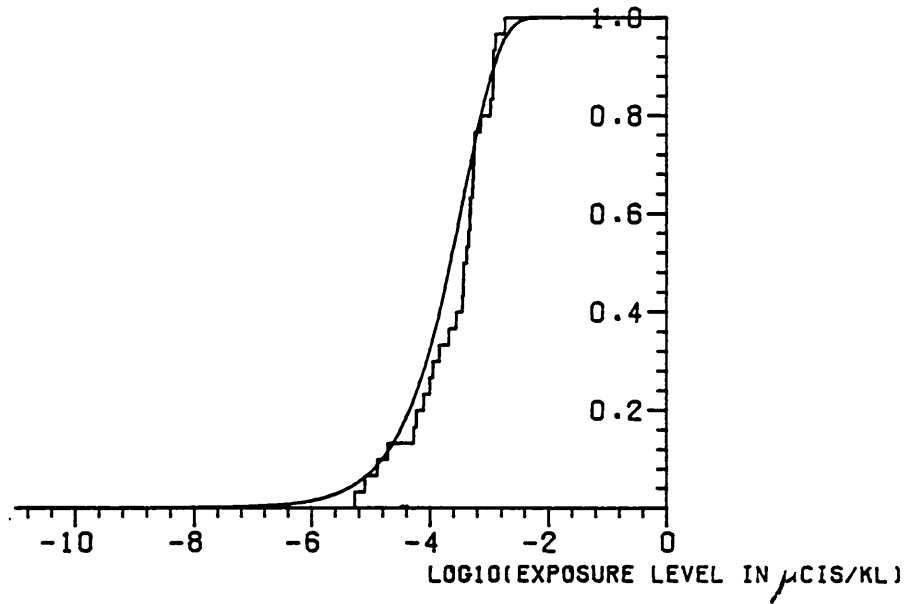


- (e) Xe_{135} air contamination Cadarache receptor 2, release duration 1 week.

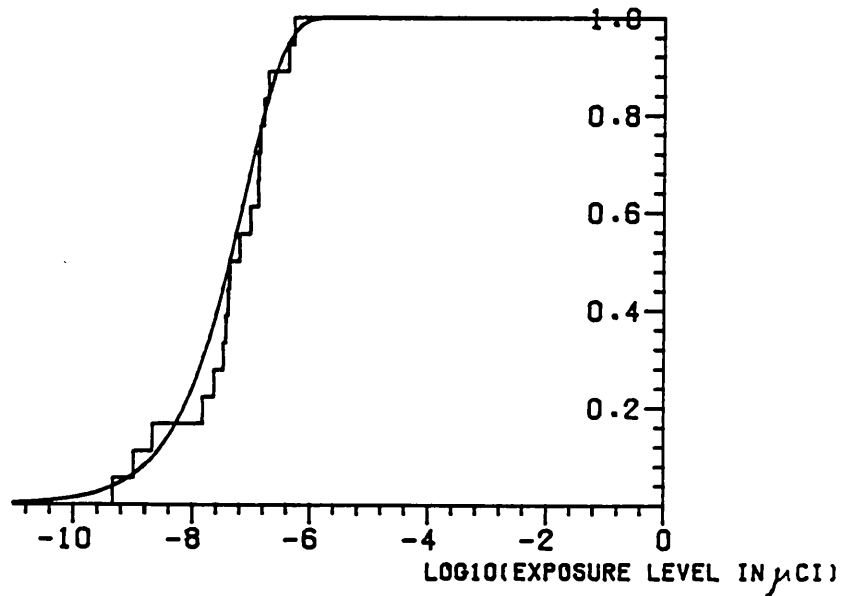


- (f) $I_{131}(g)$ dry deposition Mol 1976 receptor 1, release duration 1 day.

Figure 3.5 : Comparison of MESOS and fitted Weibull cumulative exposure distributions.



(g) Kr_{85} air contamination Heysham receptor 1, release duration 3 days.



(h) $I_{131}(p)$ wet deposition Ispra receptor 8, release duration 1 week.

Figure 3.5 : Comparison of MESOS and fitted Weibull cumulative exposure distributions.

Two three-hourly exposure distributions which the Weibull fits well are shown in Figure 3.5(c) and (d). The fit is good at all levels, even in the distribution tails, especially for the Kr_{85} exposures.

Comparisons for longer release durations are shown in Figure 3.5(e)-(h). Discrepancies between MESOS and fitted Weibull levels which are apparently much greater are not statistically significant because of smaller sample sizes. Figure 3.5(e) shows differences approaching an order of magnitude for Xe_{135} exposures due to releases of duration one week. This, together with Figure 3.5(f), represents the worst discrepancies found; the fit observed in Figure 3.5(g, h) is more typical. Note that fit in the upper tails is fairly good even in Figure 3.5(e, f).

The Weibull distribution fits most exposure distributions appreciably better than those in Figure 3.5(a, b, e, f). Recalling the uncertainties in the MESOS data themselves, the further approximations introduced by representing them as Weibull are not unacceptable, even in the worst cases.

The possibility of fitting to the data the gamma rather than the Weibull distribution was mentioned in the previous section. A general way to test the usefulness of this is to embed the two distributions in some larger family, thus enabling a direct comparison to be made. One appropriate family is the generalized gamma family whose probability density is

$$f(y) = \frac{\alpha(y/\mu)^{\alpha/\beta-1}}{\mu\Gamma(1/\beta)} \exp\{ -(y/\mu)^\alpha \},$$

$$(\alpha, \beta, \mu > 0; y > 0)$$

which is Weibull when $\beta=1$ and gamma when $\alpha=1$. The idea is then to fit the more general distribution and see if it provides a better

representation of the data than the Weibull. Since the gamma distribution is being considered as an alternative the possibility that $\alpha \approx 1$ is of interest. The statistic ℓ_{diff} , which is

$$2 \{ \ell(\hat{\alpha}, \hat{\beta}, \hat{\mu}) - \ell(\hat{\alpha}_1, 1, \hat{\mu}_1) \}$$

is displayed in Table 3.3 for the three-hourly exposure data. Here $\ell(\alpha, \beta, \mu)$ - the loglikelihood of an exposure dataset evaluated at α , β , μ - is a measure of the plausibility of the generalized gamma density in accounting for the observed data. The maximum likelihood estimates of (α, β, μ) when all three parameters may vary are $(\hat{\alpha}, \hat{\beta}, \hat{\mu})$; and those when $\beta=1$, which corresponds to the Weibull distribution, are $(\hat{\alpha}_1, 1, \hat{\mu}_1)$. The statistics in Table 3.3 would be chi-squared on one degree of freedom were the data truly Weibull; large values of them indicate that the generalized gamma distribution fits the data better.

Some of the values of the loglikelihood ratio statistic are large; they generally correspond to datasets which Table 3.2(a) shows are poorly fitted by the Weibull distribution. However none of their values of $\hat{\alpha}$, shown in Table 3.4 for the Xe₁₃₅ data, are close to one. They are usually smaller than the corresponding values of $\hat{\alpha}_1$, showing that departures from the Weibull distribution are not in the direction of the gamma distribution. Although on some occasions the generalized gamma distribution does appreciably better, it does not do so consistently enough to supplant the Weibull.

The conclusion drawn is that the Weibull function is flexible enough to model the distribution of the MESOS exposure levels sufficiently well for the present purpose, for a wide range of distances, types of exposure, and release durations.

Receptor	Heysham Kr ₈₅ air	Karlsruhe Xe ₁₃₃ air	Mo1 1973 I ₁₃₁ (p) air	Mo1 1976 I ₁₃₁ (g) dry	Cadarache Xe ₁₃₅ air	Ispra I ₁₃₁ (p) wet
1	.62	.76	1.71	.33	.08	.23
2	.07	2.67	.27	1.02	8.91°	.02
3	5.82*	4.24*	.12	1.36	.63	.15
4	.08	.07	1.34	.20	1.14	1.49
5	7.49°	1.39	9.56°	3.01	11.15!	2.86
6	7.13°	3.83	17.88!	.36	3.66	.48
7	.08	14.16!	24.12!	.04	3.47	1.10
8	1.67	8.19°	.13	2.81	6.40*	.41
9	2.36	10.54°	12.00!	1.26	12.98!	.26
10	1.88	22.00!	.32	.27	15.35!	.95
11	.58	5.12*	.12	3.07	21.78!	.12
12	2.56	2.02	1.76	.10	.58	2.16
13	.47	1.27	1.41	5.15*	5.70*	.34
14	18.83!	1.18	.17	2.01	4.00*	.82
15	17.01!	1.74	.01	2.00	4.19*	.30
16	3.74	6.45*	5.49*	.93	4.32*	1.34

Table 3.3: Loglikelihood ratio statistics for Weibull distribution
within generalized gamma family

* significance level between .05 and .01

° significance level between .01 and .001

! significance level less than .001

Receptor	No of Exposures	Weibull		generalized gamma			ℓ_{diff}
		α_1	$*\mu_1 \times 10^5$	α	β	$*\mu$	
1	504	0.39	46.33	0.36	0.89	3.12×10^{-4}	0.08
2	317	0.45	6.22	0.18	0.19	2.31×10^{-7}	8.91
3	186	0.50	1.98	0.37	0.62	4.23×10^{-6}	0.63
4	111	0.56	0.76	0.34	0.42	4.02×10^{-7}	1.14
5	328	0.41	9.87	0.14	0.14	2.98×10^{-11}	11.15
6	194	0.54	2.61	0.28	0.31	2.28×10^{-7}	3.66
7	457	0.35	36.98	0.21	0.40	2.64×10^{-6}	3.47
8	324	0.43	7.02	0.20	0.26	4.05×10^{-8}	6.40
9	126	0.55	1.10	0.10	0.04	1.18×10^{-19}	12.98
10	613	0.38	73.29	0.92	3.47	7.92×10^{-3}	15.35
11	225	0.55	1.54	0.10	0.04	2.08×10^{-20}	21.78
12	589	0.36	65.05	0.31	0.77	2.33×10^{-4}	0.58
13	413	0.45	11.03	0.25	0.38	1.49×10^{-6}	5.70
14	345	0.45	6.94	0.26	0.41	1.47×10^{-6}	4.00
15	140	0.48	1.86	0.15	0.12	8.18×10^{-12}	4.19
16	96	0.54	0.87	0.19	0.14	1.67×10^{-10}	4.32

* units for μ_1
and μ are
 $\text{Cism}^{-3} \times 10^{-6}$

Table 3.4 Weibull and generalized gamma parameter estimates : Cadarache Xe₁₃₅ three-hourly time-integrated air concentration data.

3.3 The effect of covariates

Summarization of MESOS exposure distributions in terms of two parameters is a big reduction of the data, but does not cut them down to size enough - the effect of nuclide characteristics and other variables is not explicitly represented, and so cannot be found for other nuclides of interest. This dependence must be built into a simple model interpretable in terms of appropriate physical ideas. Such a model can be posited by considering the following chain of reasoning.

Suppose that a puff released from a source initially contains Q_0 Curies of a nuclide with decay constant λ (seconds⁻¹), deposition velocity v_d (metres/second), and washout coefficient $\lambda_w J^{0.8}$ (seconds⁻¹). Here J is the rainfall rate in millimetres/hour. After a travel time T the quantity of material still in the cloud is roughly

$$Q_0 \exp\{ -\lambda T - v_d(T/h) - \lambda_w T p(\text{rain}) \},$$

where h is the mean height of the puff over its trajectory and $p(\text{rain})$ is the proportion of travel time during which rain has fallen on the puff. Here it is assumed that wet and dry deposition do not interact. If the source-receptor distance in metres is d , then $T=d/u$, where u is the mean speed of the puff, at least for fairly direct exposures. If the puff is thought of as roughly brick-shaped, it has approximate volume

$$T^* u h d^\omega,$$

where d^ω is its lateral dispersion and T^* is the time it takes to cross the receptor - which will be small compared with its travel time T . Many simple trajectory models use a power-law representation of plume broadening; see Doury(1976), who argues that $\omega \approx 1$.

Then the mean time-integrated air concentration at a receptor

should be proportional to

$$\exp\{ - d\lambda/u - dv_d/uh - d\lambda_w p(\text{rain})/u - \omega \log(d) \},$$

which suggests fitting a linear model of form

$$\theta_1 + \theta_2 d\lambda + \theta_3 dv_d + \theta_4 d\lambda_w + \theta_5 \log(d)$$

to the log-mean of the Weibull exposure distributions. Here the unknown parameters θ are to be estimated from the data. All but θ_1 - which represents an overall mean - are expected to be negative; and they may be simply interpreted as functions of u , h and so on:

$\theta_2 \approx -1/u$, $\theta_3 \approx -1/uh$, $\theta_4 \approx -P(\text{rain})/u$, and $\theta_5 \approx -\omega$. A similar equation

$$\phi_1 + \phi_2 d\lambda + \phi_3 dv_d + \phi_4 d\lambda_w + \phi_5 \log(d)$$

may be fitted to the log-standard deviation of the Weibull distributions.

Although α and μ parametrize the Weibull density conveniently for maximum likelihood estimation, it must be parametrized in terms of its mean and standard deviation for the regression modelling. The mean and standard deviation of a Weibull distribution with parameters α and μ are

$$M = \mu \Gamma(1+1/\alpha)$$

and

$$S = \mu \{ \Gamma(1+2/\alpha) - \Gamma(1+1/\alpha)^2 \}^{1/2}$$

where $\Gamma(x)$ ($x > 0$) is the gamma function $\int_0^\infty u^{x-1} e^{-u} du$. Given values of M and S , α and μ can be recovered by numerical or graphical solution of these equations.

The regression equations were fitted in GLIM using ordinary least squares. Regression was based on the mean and standard deviation of the time-integrated air contamination exposure

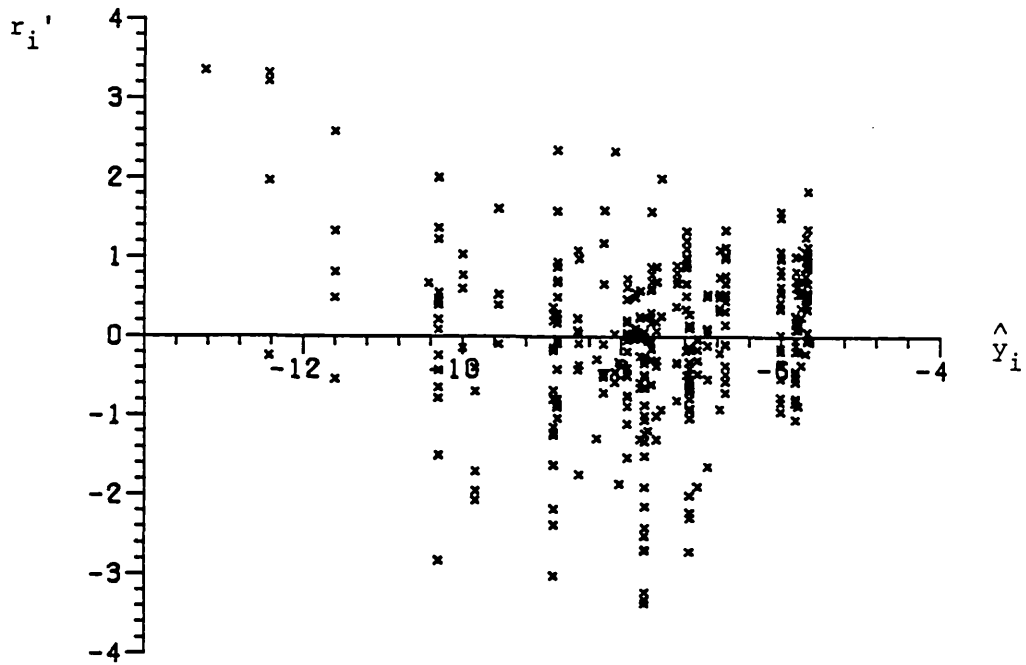
distributions at the sixteen receptors for $I_{131}(p)$, $I_{131}(g)$, Cs_{137} , Kr_{85} , Xe_{133} , and Xe_{135} for Mol in both 1973 and 1976; and $I_{131}(p)$, Kr_{85} , and Xe_{135} results for the other sources: 368 observations in all for each regression.

The fitted equation for the distribution log-mean is

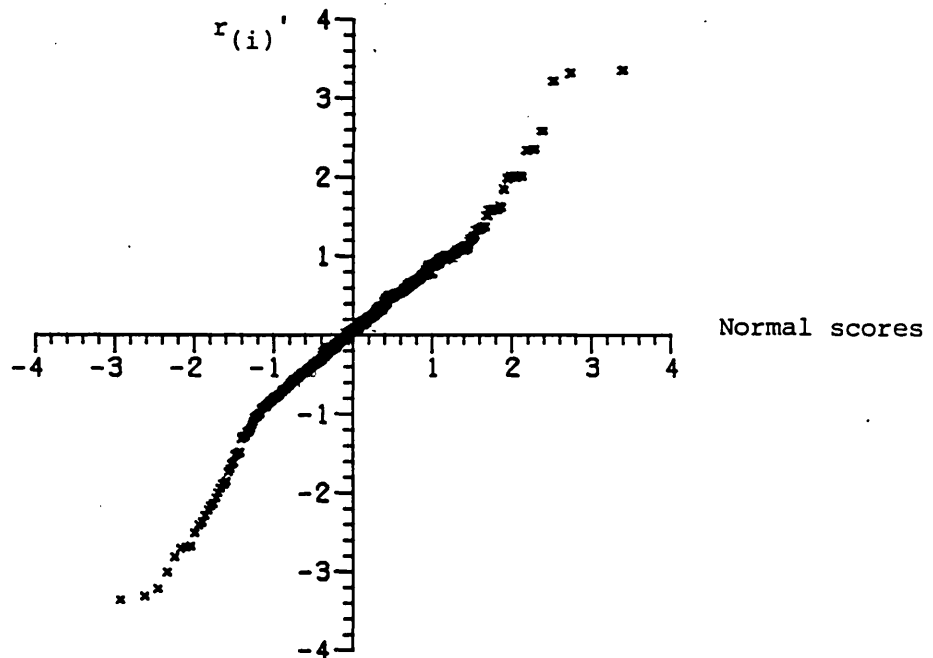
$$\begin{aligned} 5.155 & - 0.1583d\lambda - 2.869 \times 10^{-4} dv_d - 9.263 \times 10^{-3} d\lambda \\ (0.831) & (4.188 \times 10^{-3}) (5.132 \times 10^{-5})^d (3.081 \times 10^{-3})^w \\ & - 0.9399 \log(d) \dots 3.3.1, \\ & (3.105 \times 10^{-2}) \end{aligned}$$

where the estimated standard errors of the parameter estimates are displayed beneath the estimates themselves. All the estimates are statistically significant. The effect of fitting the four extra parameters in addition to the overall mean to the data is to reduce the sum of squares for regression from 863.3 on 367 degrees of freedom to 56.3 on 363 degrees of freedom, a drop whose size indicates that the fitted equation explains the variation in the log-Weibull means well. Inspection of the residuals from the fitted equation reveals no particular dependence on nuclide half-life, deposition velocity, or washout coefficient, but shows that the equation tends to underestimate the mean by a little at 100 kilometres and by slightly more at distances of 1100 kilometres or greater. However the largest numerical difference between a log-mean and its fitted value is ≈ 1.3 , and only 31 of the 368 residuals correspond to error factors of two or more between fitted and Weibull means. Of the 31, 21 are for Xe_{135} , which the Weibull distribution does not fit well because of the bimodality of the exposure distributions close to the source.

Residual plots for the regression are on display in Figure 3.6 (a)-(d). Dependence of the standardized residuals r'_i on distance is apparent in Figure 3.6(a), which shows the r'_i plotted against the

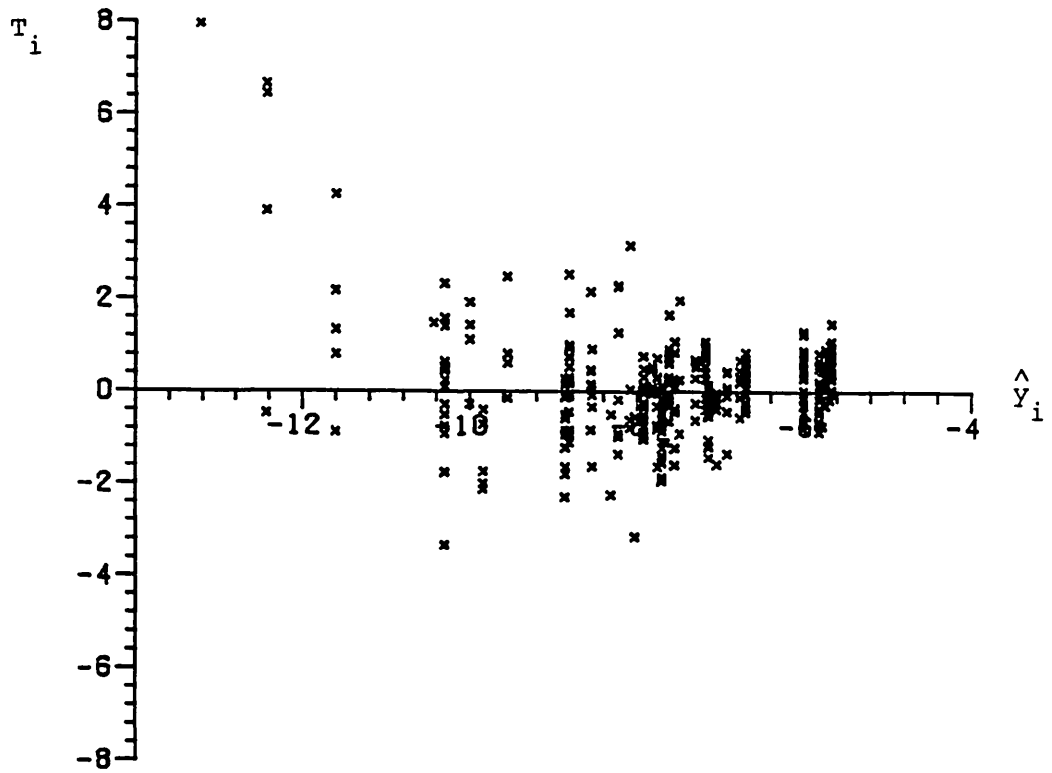


(a) standardized residuals r_i' vs. fitted values \hat{y}_i

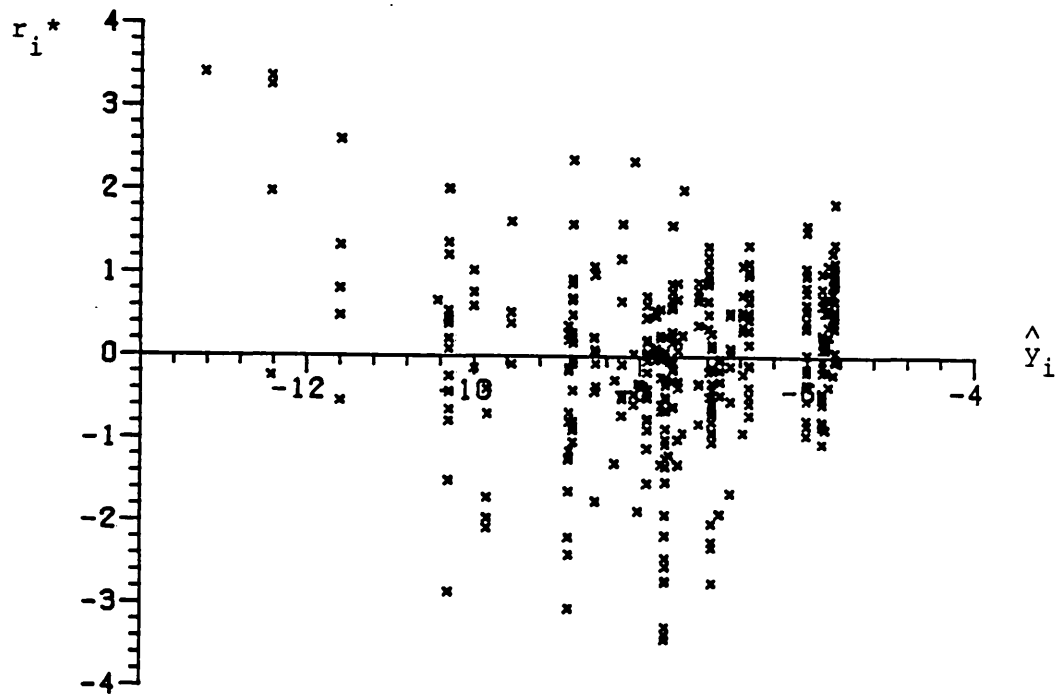


(b) ordered standardized residuals $r(i)'$ vs. Normal order statistics.

Figure 3.6 : Residual plots for regression of log-Weibull means; air contamination data.



(c) modified Cook statistic T_i vs. fitted values \hat{Y}_i



(d) jack-knifed residuals r_i^* vs. fitted values \hat{Y}_i

Figure 3.6 : Residual plots for regression of log-Weibull means; air contamination data.

fitted values \hat{y}_i for the regression. That the r'_i are roughly Normally distributed is clear when their ordered values are plotted against Normal order statistics as in Figure 3.6(b). The plot of the Cook statistics T_i against the \hat{y}_i in Figure 3.6(c) shows three rather large values for small \hat{y}_i , but when the corresponding y_i are omitted and the parameters re-estimated the regression equation changes little. Accordingly they have been retained. The plot in Figure 3.6(d) of jackknifed residuals r^*_i against the \hat{y}_i is little different from Figure 3.6(a). The regression is adequate.

Weather conditions in 1976 lead to generally higher exposures than in 1973. This is reflected in the very significant $F_{5, 358}$ statistic 10.75 for testing different regression equations for each year against the single equation. The biggest difference of parameters is between $\theta_2(1973)$ and $\theta_2(1976)$, which reflect the influence of λd on exposure levels. Their values are .14 and .18 respectively, and correspond to mean puff windspeeds of 7.1 m/s in 1973 and 5.6 m/s in 1976. This suggests that the overall effect of a summer with many blocking anticyclones on mean puff windspeeds is to decrease them by roughly 1-2 m/s.

The equation fitted to the log-Weibull standard deviations of the distributions is

$$\begin{aligned}
 10.41 & - 0.1153d\lambda - 2.892 \times 10^{-4} d v_d + 4.209 \times 10^{-4} d \lambda \\
 (0.377) & \quad (4.122 \times 10^{-3}) \quad (5.042 \times 10^{-5})^d \quad (3.032 \times 10^{-3})^w \\
 & \quad - 1.328 \log(d) \quad \dots \quad 3.3.2, \\
 & \quad (3.056 \times 10^{-2})
 \end{aligned}$$

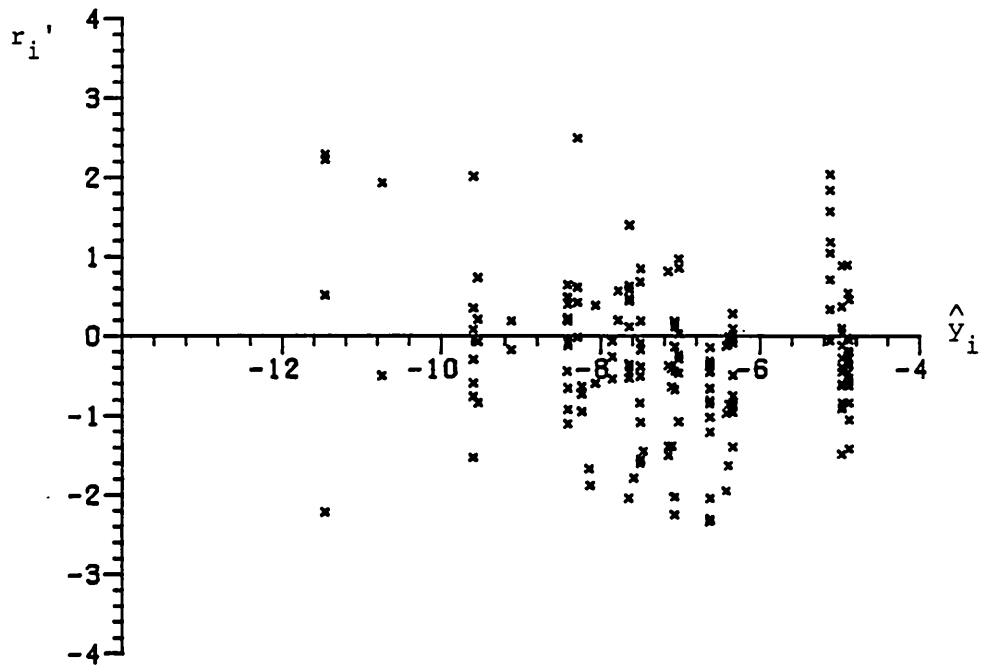
which reduces the sum of squares for the regression from 938.9 on 367 degrees of freedom to 54.5 on 363 degrees of freedom. Details of both regression equations are shown in Table 3.5, and Figure 3.7 gives residual plots for the standard deviation regression. Broadly the same comments as above apply to them as did to those for the

Parameter	Estimate	s.e.	
φ_1	5.155	0.3831	
φ_2	-0.1583	4.188×10^{-3}	
φ_3	-2.869×10^{-4}	5.123×10^{-5}	
φ_4	-9.263×10^{-3}	3.081×10^{-3}	
φ_5	-0.9399	3.105×10^{-2}	$\hat{\sigma}^2 = 0.1550$
φ_1	10.41	0.3770	
φ_2	-0.1153	4.122×10^{-3}	
φ_3	-2.892×10^{-4}	5.042×10^{-5}	
φ_4	4.209×10^{-4}	3.032×10^{-3}	
φ_5	-1.328	3.056×10^{-2}	$\hat{\sigma}^2 = 0.1502$

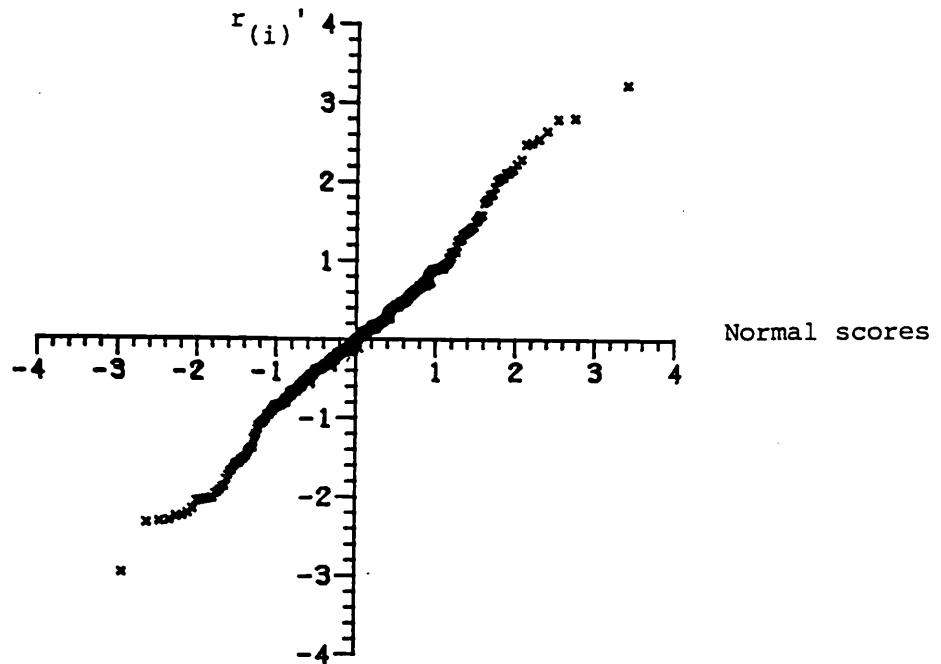
Correlation matrix of estimates

	1	2	3	4	5
1	1.0	0.4719	0.1367	-0.1422	-0.9974
2		1.0	0.1127	0.1381	-0.5040
3			1.0	-0.3403	-0.1468
4				1.0	-0.1523
5					1.0

Table 3.5 Details of regression equations for dependence of Weibull air concentration distributions on covariates.

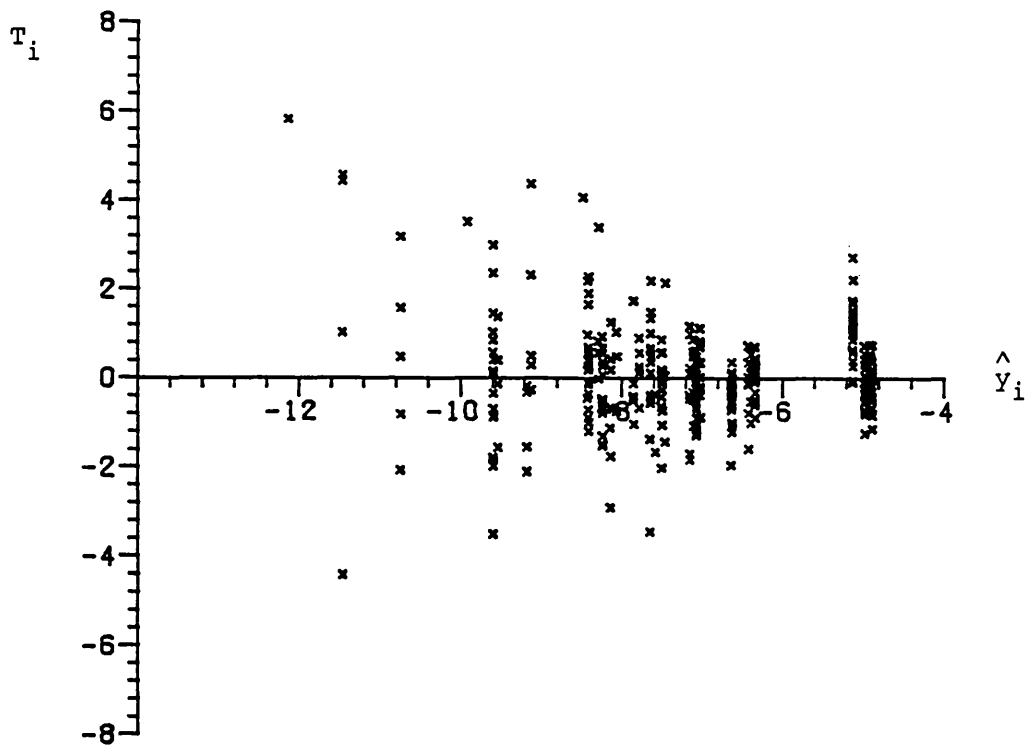


(a) standardized residuals r_i' vs. fitted values \hat{y}_i

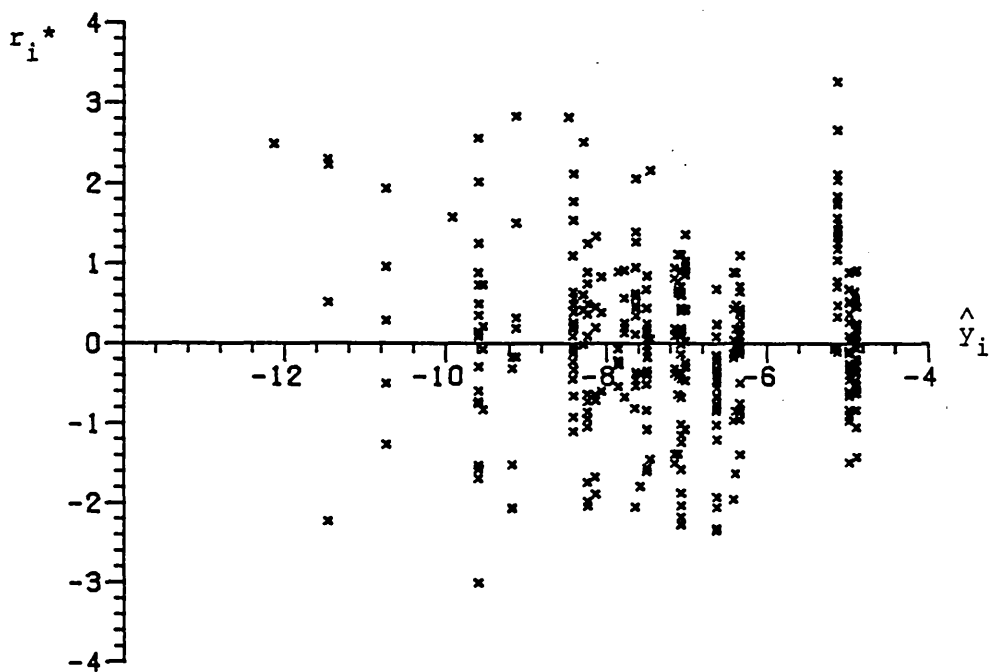


(b) ordered standardized residuals $r(i)'$ vs. Normal order statistics.

Figure 3.7 : Residual plots for regression of log-Weibull standard deviations; air contamination data.



(c) modified Cook statistics T_i vs. fitted values \hat{Y}_i



(d) jack-knifed residuals r_i^* vs. fitted values \hat{Y}_i .

Figure 3.7 : Residual plots for regression of log-Weibull standard deviations; air contamination data.

mean regression. The dependence of the residuals on distance apparent in Figure 3.6 is less obvious in Figure 3.7, but the big residuals for both equations generally correspond to each other, and once again it is the Xe_{135} residuals which tend to be worst. Although retained in the equation, $\hat{\phi}_4$ is smaller than its standard error, indicating that washout has little or no detectable effect on the spread of exposures to air contamination and consequently dry deposition.

As already explained, the equations may be interpreted in terms of physical quantities. The mean equation implies that $u \approx 6.3$ m/s, $h \approx 550$ m, $p(\text{rain}) \approx 0.06$, and $\omega \approx 0.94$. The value of h seems rather a small mean height for the puff; and lateral broadening of a puff almost proportional to source-receptor distance seems large for the distances being considered here, although presumably some allowance must be made for the effect of synoptic divergence of trajectories, which otherwise is not built into this simple model. Lateral dispersion is not always well-described by a power-law, and it seems likely that the simplistic use of puff broadening proportional to d^ω is responsible for the high value of ω and the low mean puff height. Despite this the values of u and $p(\text{rain})$ are consistent with the work of Wrigley(1982).

The equation for the standard deviation is best interpreted by considering the coefficient of variation of the fitted exposure distribution, its standard deviation divided by its mean, a measure of the relative variation of a distribution about its mean. This is roughly

$$\exp\{ 5.3 + 4.3 \times 10^{-2} d \lambda - 2.3 \times 10^{-6} d v_d + 9.7 \times 10^{-4} d \lambda_w - 0.39 \log(d) \}.$$

The approximate ranges over which the nuclide parameters and distance vary are: source-receptor distance $d \approx 10^5 - 10^6$ m; nuclide decay

rate $\lambda \approx 10^{-9} - 3.0 \times 10^{-5} \text{ s}^{-1}$; deposition velocity $v_d \approx 0 - 3.0 \times 10^{-3} \text{ ms}^{-1}$; and washout coefficient $\lambda_w \approx 0 - 5.0 \times 10^{-5} \text{ J}^{0.8} \text{ s}^{-1}$. Clearly the values of v_d and λ_w have little or no discernable impact on the spread of the air contamination exposure distributions relative to their means, at least for the ranges of d of interest here. However the value of λ is important: nuclides such as Xe_{135} whose decay constants correspond to short half-lives of the order of a few hours have exposure distributions whose relative spread changes little as source-receptor distance increases, but nuclides with half-lives greater than the time taken for material to cross the map have exposure distributions whose relative spread drops by a factor two or so as d increases from 10^5 to 10^6 metres.

A qualitative explanation for this may be deduced from Figure 3.8 - amalgamated from Figures 5.10 and 5.11 of Wrigley(1982) - which shows scatter diagrams for time-integrated air concentrations experienced 100 and 800 kilometres north of Mol in 1973, as functions of travel-time. Note that both scales are logarithmic. The Kr_{85} exposures indicate for inert nuclides the range of variability due solely to meteorological conditions, whereas the effect of decay is manifest from the Xe_{135} exposures. The $\text{I}_{131}(\text{p})$ results, more complicated because of the effect of wet and dry deposition, are not directly relevant just now; but nevertheless it is informative to note that the overall effect of deposition seems to be to increase the spread of the distribution for a given travel-time by a factor which grows from about one to about ten as travel-time increases from about two hours to its maximum possible value of four days, and to reduce its mean by similar factors.

To return to the effect of decay constant on the spread of an air contamination distribution relative to its mean, compare the Kr_{85} and Xe_{135} exposures in Figure 3.8. The overall spread of both Kr_{85}

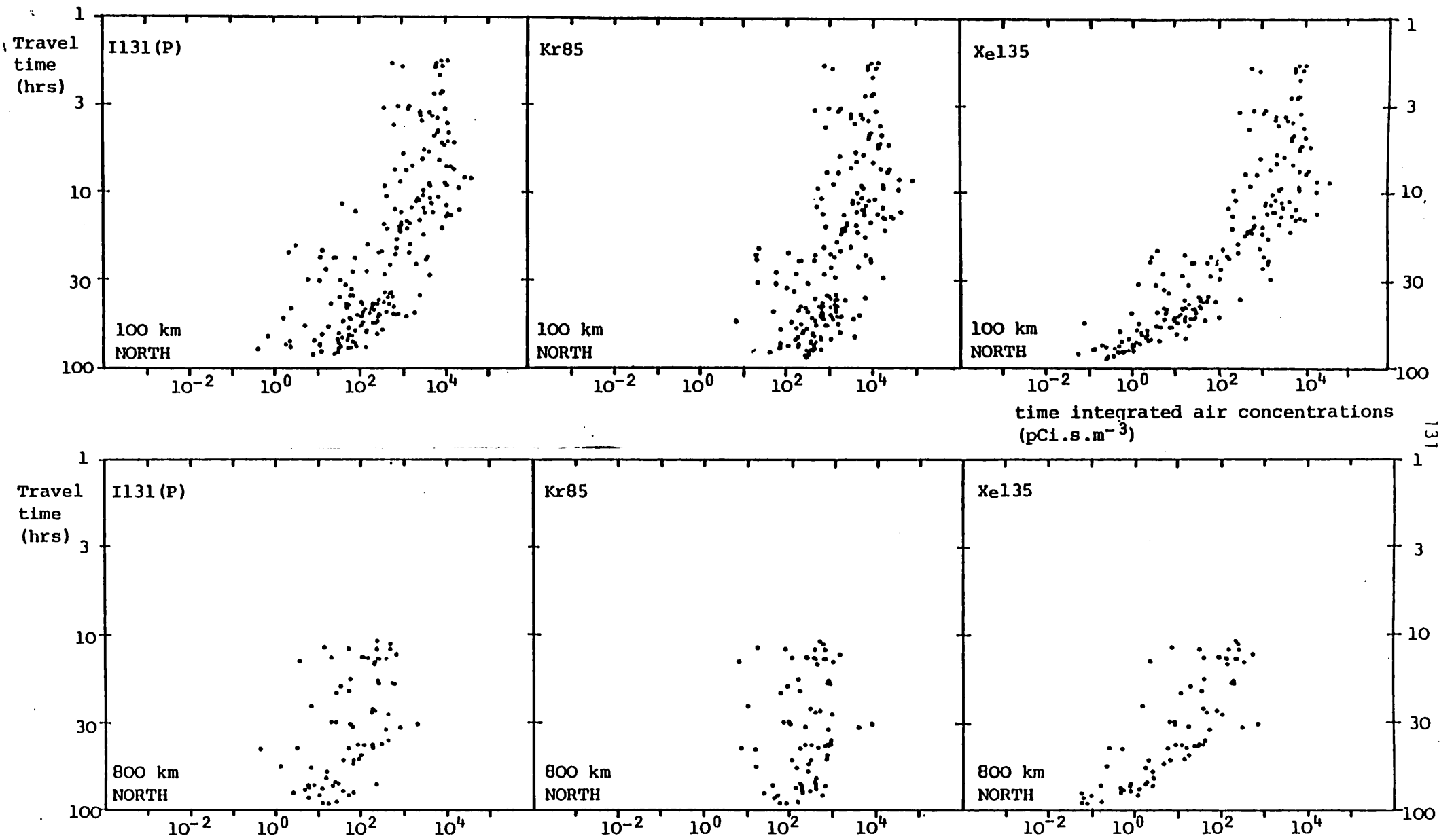


Figure 3.8 : Scatter plots for nuclide exposures north of Mol through 1973.

distributions is about 10^3 - 10^4 , but the mean exposure drops by a factor 10 or so with the extra 700 kilometres - from 3,500 to 460 $\mu\text{Ci m}^{-3}$. However the overall spread of the Xe_{135} distributions drops from a factor about 10^6 to about 10^4 , and the mean from about 10^4 to about 10^2 as distance increases, because of the effect of the large decay constant. The spread of the Xe_{135} data diminishes with distance by a factor comparable with the decrease in mean exposure, so the coefficient of variation changes little with distance.

The exposure distributions for dry deposition are very simply related to those for air contamination, since in almost all cases the level of dry deposition is just v_d times the time-integrated air concentration. This is not the case when prolonged stability of the lower part of the boundary layer leads to a non-uniform vertical concentration profile of pollutant due to preferential depletion of its lowest few metres. This is rare and leads to relatively low exposures so does not noticeably alter the shape of the exposure distribution. Table 3.6 gives an idea of the degree of consistency observed between the air contamination and dry deposition exposure distributions due to unit three-hourly releases from Mol of Cs_{137} through 1973 and $\text{I}_{131}(\text{p})$ through 1976. Corresponding estimates of the Weibull shape parameter α are very close - almost always within one standard error of each other, which implies that any differences may be ascribed to estimation errors. The ratio of dry deposition Weibull scale parameters μ to their air contamination counterparts is usually slightly less than the appropriate value of v_d , and moreover the same is true of the actual means and standard deviations of the distributions. This is a consequence of the argument given above for the distributions to be virtually identical apart from scale except that occasionally dry deposition is inhibited by the stability of the boundary layer. It implies that finding the dry deposition

receptor	TIC $\hat{\alpha}$	dry dep. $\hat{\alpha}$	$\hat{\mu}_{dry}/\hat{\mu}_{TIC} \times 10^{-3}$	$M_{3dry}/M_{3TIC} \times 10^{-3}$	$Q_{3dry}/Q_{3TIC} \times 10^{-3}$
1	0.528	0.525	0.95	0.96	0.96
2	0.673	0.676	0.93	0.92	0.93
3	0.702	0.697	0.96	0.96	0.99
4	0.703	0.694	0.97	0.99	1.01
5	0.589	0.594	0.97	0.95	0.91
6	0.723	0.724	0.98	0.98	0.98
7	0.779	0.779	0.99	0.99	0.99
8	0.775	0.712	0.98	0.99	1.00
9	0.474	0.475	0.98	0.96	0.94
10	0.689	0.691	0.99	0.99	0.98
11	0.729	0.729	0.98	0.99	0.99
12	0.744	0.743	0.99	0.99	0.99
13	0.507	0.509	0.95	0.95	0.95
14	0.679	0.668	0.92	0.94	0.93
15	0.706	0.702	0.95	0.95	0.98
16	0.990	0.989	0.86	0.86	0.91

Table 3.6 Comparison of Weibull and actual means and standard deviations of three-hourly exposure distributions

(a) Ca_{10T} TIC and dry deposition, Mol 1973

receptor	TIC $\hat{\alpha}$	dry dep. $\hat{\alpha}$	$\hat{\mu}_{dry}/\hat{\mu}_{TIC} \times 10^{-3}$	$M_{3dry}/M_{3TIC} \times 10^{-3}$	$Q_{3dry}/Q_{3TIC} \times 10^{-3}$
1	0.438	0.439	2.82	2.81	2.66
2	0.548	0.545	2.63	2.79	2.80
3	0.560	0.553	2.76	2.76	2.81
4	0.594	0.593	2.84	2.85	2.89
5	0.534	0.537	2.87	2.83	2.79
6	0.638	0.640	2.94	2.93	2.89
7	0.633	0.633	2.95	2.95	2.95
8	0.634	0.634	2.96	2.95	2.96
9	0.487	0.489	2.88	2.85	2.82
10	0.604	0.606	2.95	2.94	2.94
11	0.760	0.760	2.95	2.95	2.94
12	0.779	0.777	2.99	2.98	2.99
13	0.453	0.452	2.77	2.77	2.74
14	0.524	0.521	2.72	2.78	2.86
15	0.599	0.587	2.78	2.87	2.94
16	0.707	0.707	2.81	2.84	2.91

Table 3.6 Comparison of Weibull and actual means and standard deviations of three-hourly exposure distributions;

(b) $I_{131(p)}$ TIC and dry deposition, Mol 1976

distribution of dry deposition simply by multiplying the distribution of air contamination by a factor v_d would tend to overestimate dry deposition were the Weibull a perfect fit to the air contamination data. Given the known discrepancies between the Weibull and MESOS exposure distributions, this procedure is totally adequate.

This simple idea will not work for wet deposition, which depends also on the incidence of rain at the receptor. When it is raining there, however, wet deposition is roughly proportional to the time-integrated air concentration, and thus is

$$\lambda_w \exp\left\{ -d\lambda/u - dv_d/uh - d\lambda_w p(\text{rain})/u - \omega \log(d) \right\}.$$

However washout coefficient λ_w depends on rainfall rate J - whose mean value varies in different parts of Europe due to the influence of orographic effects and so on - so this is unlikely to be so good a fit as the corresponding equation for air contamination. Throughout the rest of this thesis the various values of λ_w are taken to be the numerical values of the constants in Tables 1.1 and 1.4: dependence on the values of the rainfall rate J is generally ignored. Because of the restriction to wet weather at the receptor the values of u , h , $p(\text{rain})$, and ω may not be the same as those above, but it seems likely that a model for wet deposition will emerge from following a procedure akin to that used above for air contamination: in this case fitting model equations

$$\log(\lambda_w) + \eta_1 + \eta_2 d\lambda + \eta_3 dv_d + \eta_4 d\lambda_w + \eta_5 \log(d)$$

and

$$\log(\lambda_w) + \nu_1 + \nu_2 d\lambda + \nu_3 dv_d + \nu_4 d\lambda_w + \nu_5 \log(d)$$

to the log-means and log-standard deviations of the Weibull wet deposition distributions. The unknown parameters η and ν are to be estimated from the data.

However a difficulty arises with estimating the effect of decay constant on wet deposition. The depositing isotopes I_{131} and Cs_{137} have half-lives 8 days and 30 years - long compared with the mean travel-time of a puff across the map - which make it hard to assess the effect of decay on wet deposition exposures simply by using their values. The model suggested by the arguments above fits the data quite well, but does not have any simple interpretation because of the impossibility of estimating the effect of λ with adequate precision. The consequence of this uncertainty is that parameter estimates are statistically significant but have the wrong signs. 'Pseudo-nuclide' exposures can be constructed to overcome this.

A pseudo-nuclide is one whose half-life is artificial: it does not belong to a radioisotope posing potentially grave risks to the public as a result of leaks from nuclear power stations, but is considered solely to tackle the problem described above. Exposures to a pseudo-nuclide whose deposition velocity and washout coefficient correspond to those of I_{131} but whose decay constant is the specified value λ may be constructed as follows.

For a given receptor all the exposure time-series are archived together in the MESOS database. By taking a specific wet deposition incident and comparing the Xe_{133} and Xe_{135} exposures at the time, the travel-time associated with the incident can be found; it is

$$T' = \frac{\log(Q_{133}) - \log(Q_{135})}{\lambda_{135} - \lambda_{133}},$$

where Q_{133} and Q_{135} and λ_{133} and λ_{135} are the exposure and decay constants for Xe_{133} and Xe_{135} respectively. The wet deposition exposure which would have occurred had the puff experienced exactly the same weather conditions along its path, but contained a nuclide with the same deposition parameters as those ascribed to I_{131} and decay rate λ rather than λ_{131} is

$$Q_{131} \exp\{ -T'(\lambda - \lambda_{131}) \}$$

in an obvious notation. This is a useful trick for varying the effect of decay constant on exposure; but since the effects of depletion and deposition depend on the exact history of the puff - not just its travel-time - they cannot be varied in the same way.

Three-hourly wet deposition exposures at sixteen receptors were created with deposition parameters the same as those of $I_{131}(p)$, $I_{131}(g)$ and Cs_{137} for releases from: Cadarache and Mol through 1973 with half-lives of one day; Karlsruhe and Ispra with half-lives of two days; and Heysham and Mol through 1976 with half-lives of four days. In all, 288 three-hourly wet exposure distributions corresponding to these short-lived imaginary isotopes were created for the regression. The means and standard deviations of 288 original MESOS three-hourly wet deposition distributions are available, for $I_{131}(p)$, $I_{131}(g)$ and Cs_{137} exposures for releases from all sources. Use of the extra data enabled the effect of half-life on wet deposition exposures to be estimated and resulted in a reasonable model.

The equations eventually fitted to the log-means and standard deviations for the wet exposure data are

$$\begin{aligned} \log(\lambda_w) &+ 8.703 &- 0.1753d\lambda && && \\ &(0.4381) &(1.265 \times 10^{-2}) && && \\ &- 1.057 \times 10^{-2} d\lambda_w &- 0.8148 \log(d) & \dots & 3.3.3 && \\ &(1.755 \times 10^{-3}) &(3.640 \times 10^{-2}) && && \end{aligned}$$

and

$$\begin{aligned} \log(\lambda_w) &+ 13.79 &- 0.1269d\lambda &- 1.186 \log(d) & \dots & 3.3.4 & \\ &(0.3518) &(1.376 \times 10^{-2}) &(2.808 \times 10^{-2}) && & \end{aligned}$$

respectively. The corresponding drops in residual sums of squares due to the inclusion of the sets of four parameters η_2 - η_5 and ν_2 - ν_5 are 722.1 to 125.4, and 895.2 to 148.8, with drops of from 575 to 572

and 573 degrees of freedom. These are not such big proportional drops as seen in the air contamination data, because of the added variability introduced by rainfall.

In the equations fitted initially, the parameter estimates $\hat{\eta}_3$, \hat{v}_3 , and \hat{v}_4 , were not statistically significant. These represent the effect of the dry deposition velocity on the mean and spread of the exposures, and the effect of the washout coefficient λ_w on the spread of the wet deposition distribution. The sign of $\hat{\eta}_3$ was positive, whereas in fact increasing v_d ought to decrease the flux surviving to remote receptors. On the other hand it is not statistically significant and there is no way to use pseudo-nuclides to estimate it more accurately. Accordingly, and in view of the difficulty of interpreting positive estimates in physical terms, these particular parameters were not eventually fitted to the data. Had a larger range of deposition velocities been used in the original MESOS calculations, the parameters could have been estimated and this difficulty would not have arisen. Details of the equations are given in Table 3.7, and the residuals are plotted in Figures 3.9 and 3.10.

The fit of the equations seems good; in particular the residuals display no strong dependence on half-life. Although there are some lower outliers in the standard deviation plots they do not have high leverage so have not been excluded from the fit. They correspond to the same outlier included several times in the pseudo-nuclide calculations.

The parameter estimates give the following approximate values for physical quantities: $u \approx 5.7$ m/s and $p(\text{rain}) \approx 0.06$, both quite plausible. The value of mean puff windspeed u is slightly smaller than the one for time-integrated air concentration, and the value of $p(\text{rain})$ is a little low. The predicted coefficient of variation of the wet deposition exposure distributions is

Parameter	Estimate	s.e.	
η_1	8.703	0.4381	
η_2	-0.1753	1.265×10^{-2}	
η_4	-1.057×10^{-2}	1.755×10^{-3}	
η_5	-0.8148	3.640×10^{-2}	$\hat{\sigma}^2 = 0.2192$
ν_1	13.79	0.3518	
ν_2	-0.1296	1.376×10^{-2}	
ν_5	-1.186	2.808×10^{-2}	$\hat{\sigma}^2 = 0.2597$

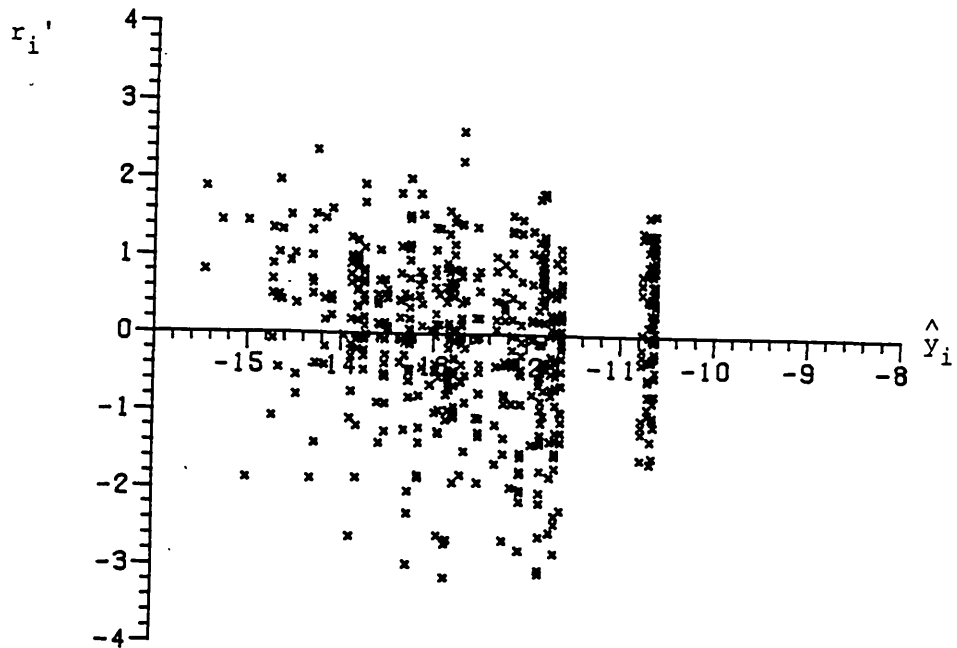
Correlation matrix of estimates of η

	1	2	4	5
1	1.0	0.3025	0.6750	-0.9976
2		1.0	-0.0190	-0.3181
4			1.0	-0.7055
5				1.0

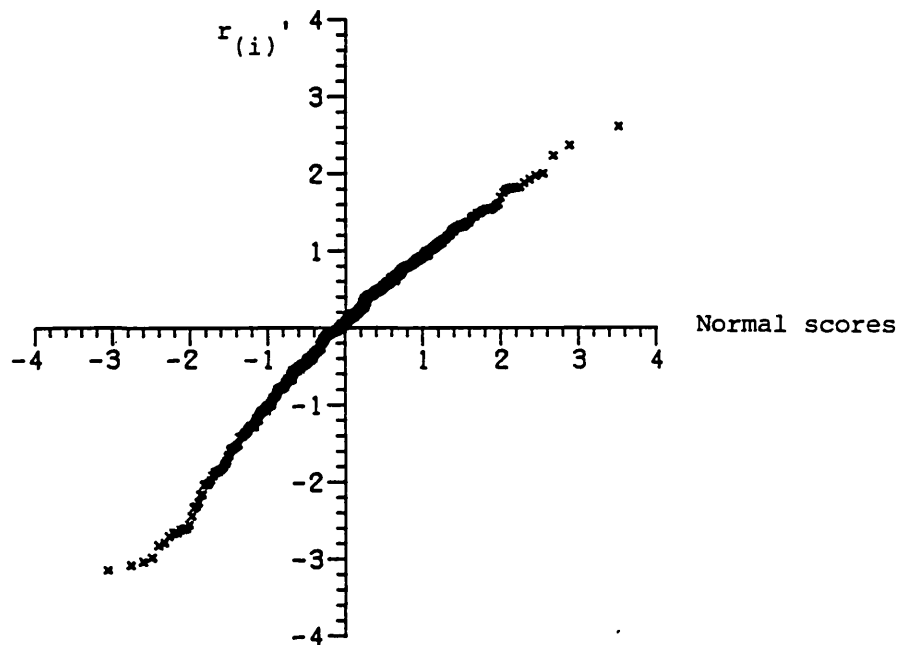
Correlation matrix of estimates of ν

	1	2	5
1	1.0	0.4274	-0.9972
2		1.0	-0.4678
5			1.0

Table 3.7 Details of regression equations for Weibull wet deposition distributions.

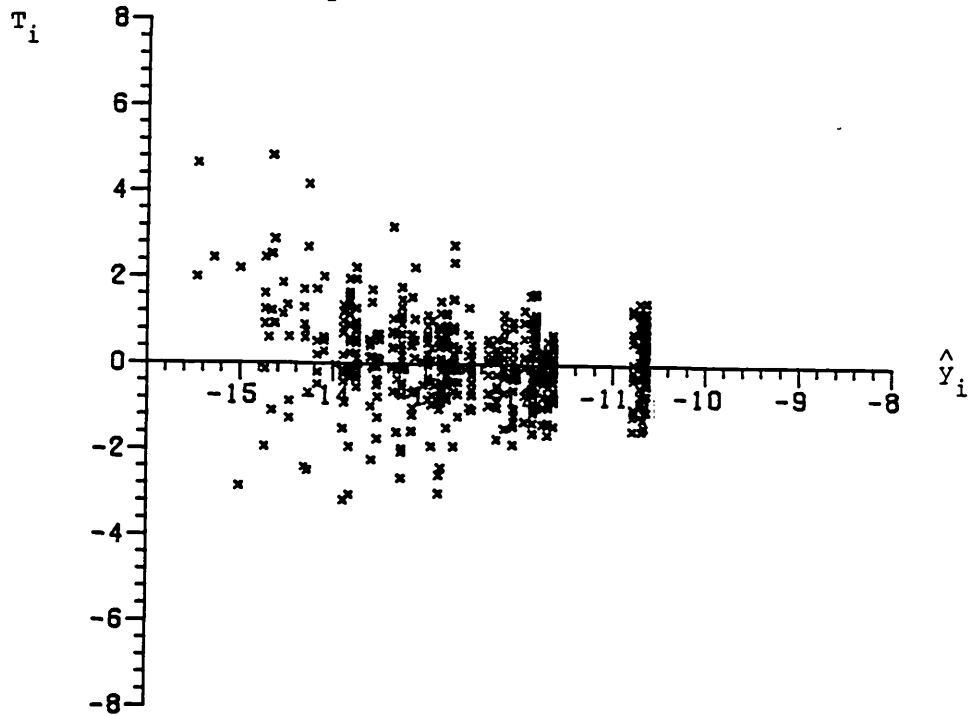


(a) standardized residuals r_i' vs. fitted values \hat{y}_i .

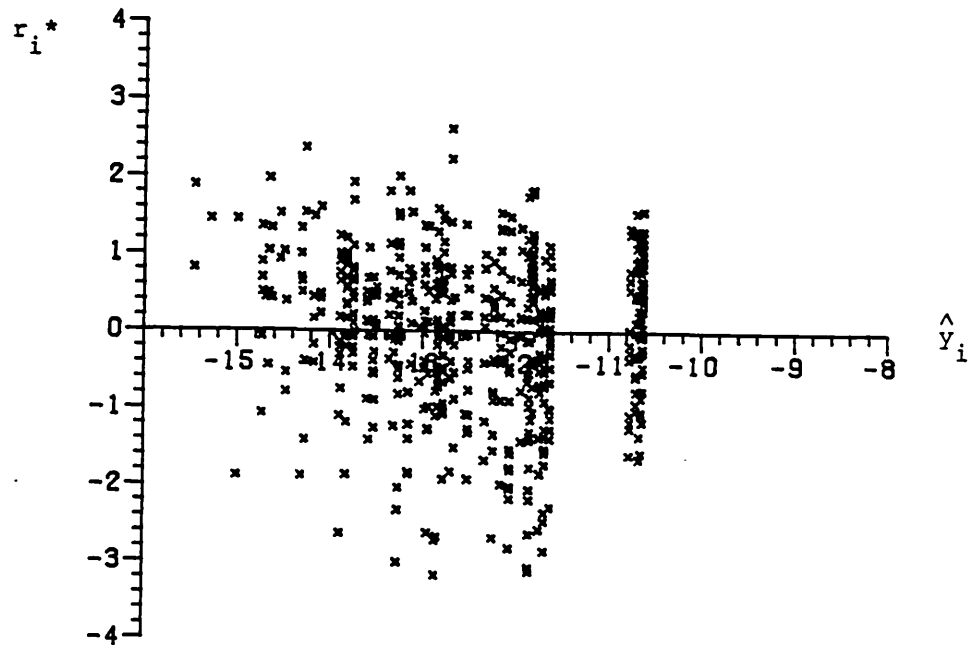


(b) ordered standardized residuals $r(i)'$ vs. Normal order statistics.

Figure 3.9 : Residual plots for regression of log-Weibull means; wet deposition data.

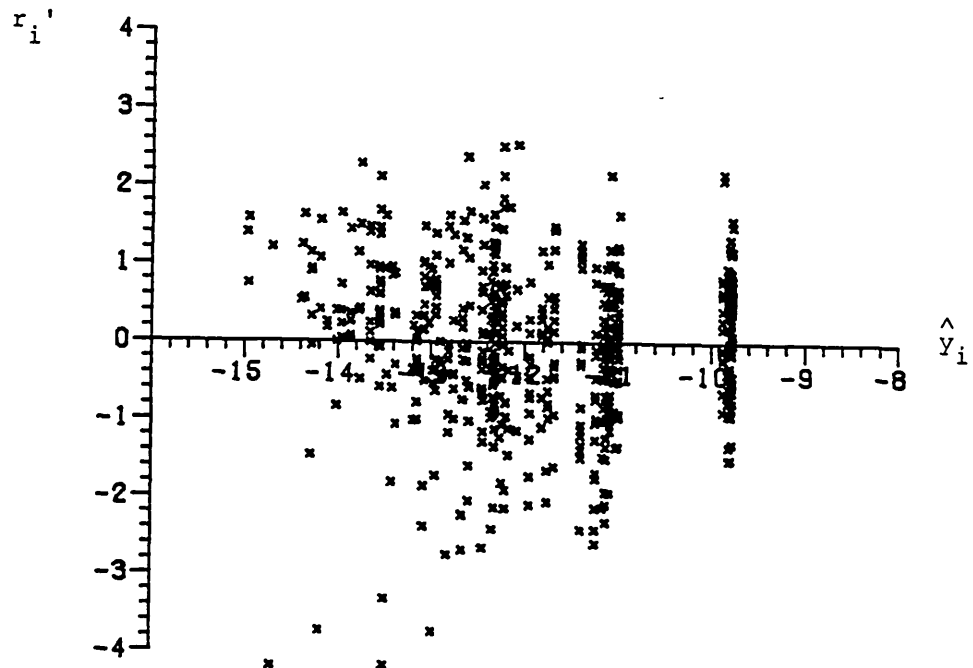


(c) modified Cook statistic T_i vs. fitted values \hat{y}_i .

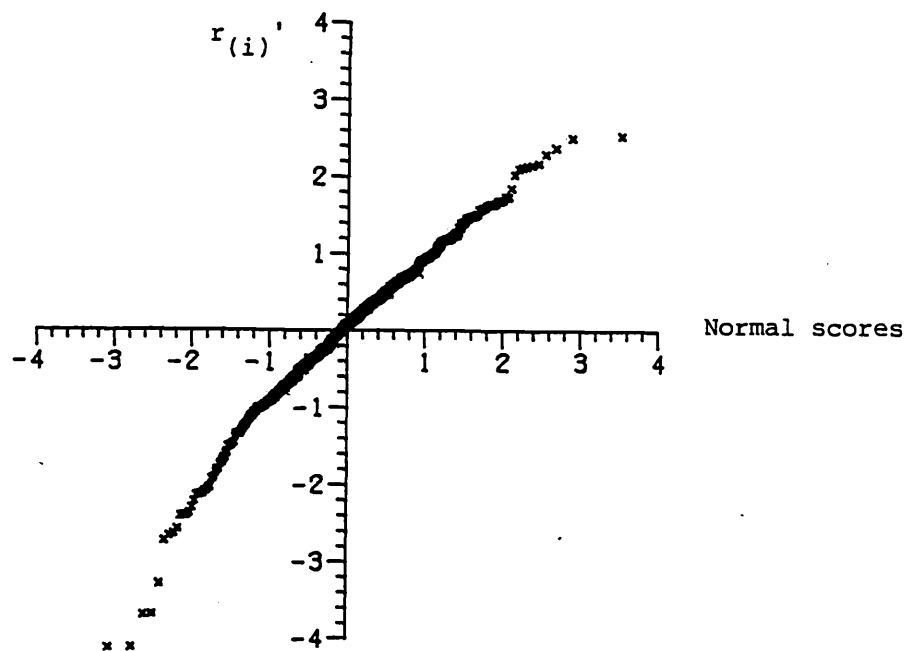


(d) jack-knifed residuals r_i^* vs. fitted values \hat{y}_i .

Figure 3.9 : Residual plots for regression of log-Weibull means; wet deposition data.

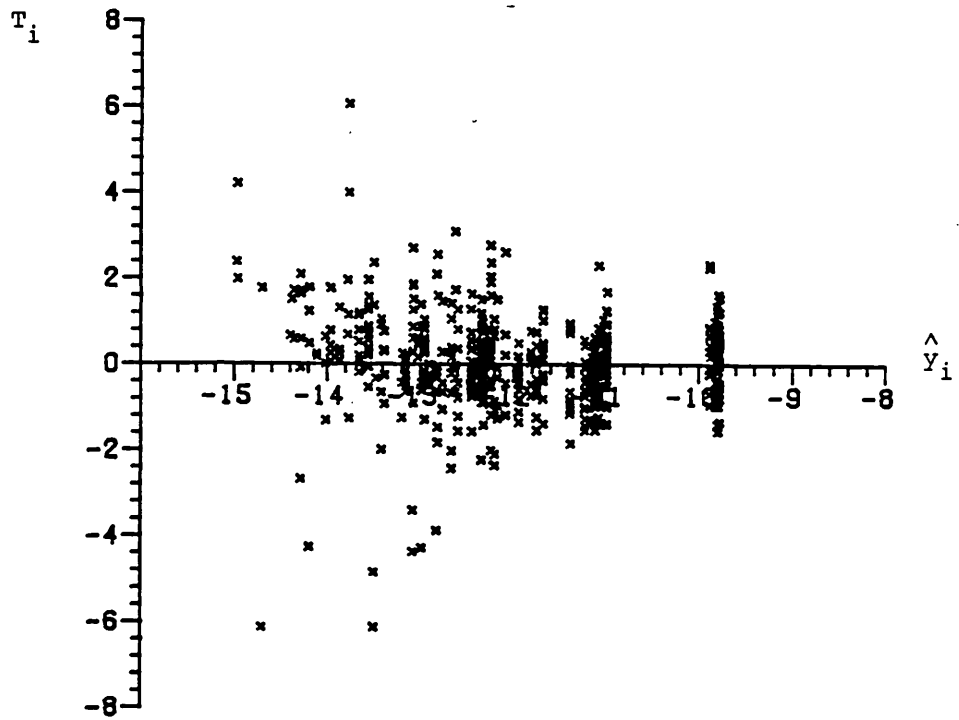


(a) standardized residuals r_i' vs. fitted values \hat{y}_i .

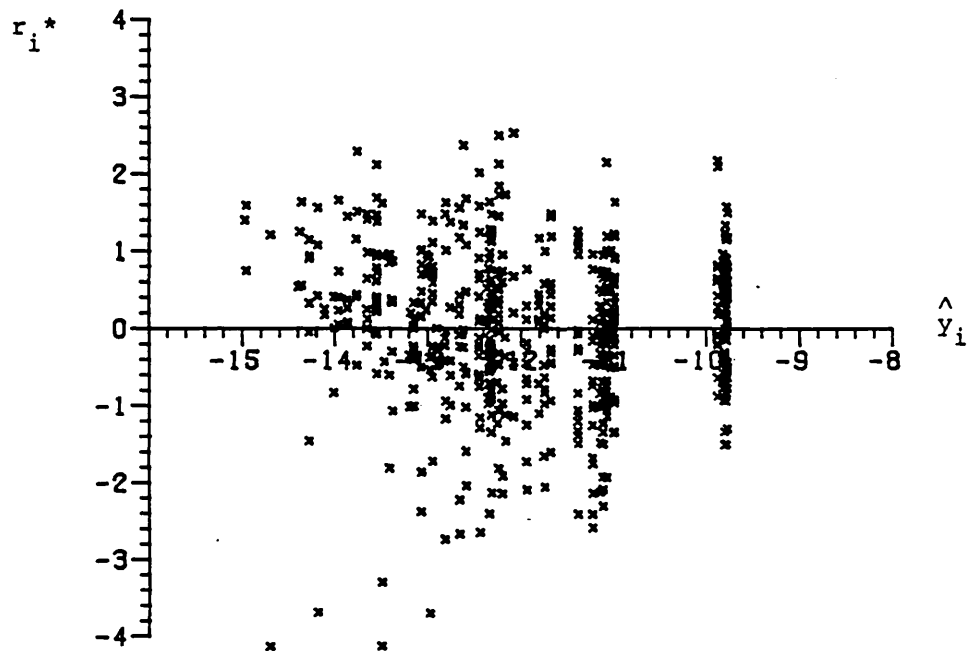


(b) ordered standardized residuals $r(i)'$ vs. Normal order statistics.

Figure 3.10 : Residual plots for regression of log-Weibull standard deviations; wet deposition data.



(c) modified Cook statistics T_i vs. fitted values \hat{y}_i .



(d) jack-knifed residuals r_i^* vs. fitted values \hat{y}_i .

Figure 3.10 : Residual plots for regression of log-Weibull standard deviations; wet deposition data.

$$\exp\{ 5.1 + 4.6 \times 10^{-2} d \lambda + 1.6 \times 10^{-2} d \lambda_w - 0.37 \log(d) \}.$$

Decay rate, largest for I_{131} with half-life eight days, has little impact on relative spread. Washout has a greater effect on wet than on air exposures: at short distances it has little effect on their spread, but at 1000 kilometres it may smear the distributions relative to their means by a factor of between 1.3 and 2.2, depending on the value of λ_w . Taken together with the effect of distance on relative spread, the coefficient of variation is halved over distances of 100 to 1000 kilometres if $\lambda_w \approx 1.5 \times 10^{-5} J^{0.8} \text{ s}^{-1}$, but it remains roughly constant if $\lambda_w \approx 5.0 \times 10^{-5} J^{0.8} \text{ s}^{-1}$.

The model described above enables exposure distributions for three-hour releases to be predicted - with an accuracy assessed in Section 3.4. However the effect of release duration must be incorporated if the model is to have general applicability.

Some dispersion models based directly on wind measurements parametrize lateral spread of a plume as a power-law function of release duration, with spread $\approx \sqrt{\text{release duration}}$. See Clarke (1976). This suggests that relationships of the form

$$\log(M_t) = \log(M_3) + \delta_1 \log(t/3) \quad \dots 3.3.5$$

$$\log(S_t) = \log(S_3) + \delta_2 \log(t/3) \quad \dots 3.3.6$$

may be useful to relate the mean M_t and standard deviation S_t of an exposure distribution for a unit nuclide release over t hours to the corresponding quantities M_3 and S_3 for unit releases over the baseline period of 3 hours. The estimates of δ_1 and δ_2 should be roughly $-\frac{1}{2}$, at least for air contamination distributions.

So far in this chapter, a fairly cavalier attitude has been adopted towards assumptions made in the course of the analysis. The intermittent positive exposures at each receptor have been treated as

independent and identically distributed, a gross oversimplification made for the sake of building as comprehensible a model as possible. However it seems overbold to treat the time-series of exposures for different release durations at the same receptor during the same period as independent. For this reason 208 (= 13 exposure datasets of 16 receptors each) time-series of exposures were taken, and a release duration $t = 3, 6, 12, 24, 72, 168$ hours was selected at random for each. The release durations were allocated to exposure datasets in such a way as to enable balanced estimation of the effects of different years and sources, but the only significant differences arise between the values of $\hat{\delta}_1$ and $\hat{\delta}_2$ for air concentration and wet deposition exposures. The parameter estimates are given in Table 3.8. The values of $\hat{\delta}_1$ and $\hat{\delta}_2$ for dry deposition and air contamination, both about -0.47 , are consistent with trajectory broadening with release duration at the rate indicated in the argument above, but the values -0.7 or so for wet deposition show that the mean and spread of wet exposure distributions drop more quickly as release duration increases. This is to be expected because wet deposition is a more sporadic phenomenon than dry exposure.

The fit of these equations is adequate: plots of the residuals show no big anomalies; there seem to be no systematic departures from them. Some large residuals from the equation for S_t for small release durations indicate that in some cases the fitted equations may slightly underestimate the spread of the exposure distributions for short release durations. Any inadequacies in the prediction equations due to this are discussed in Section 3.4.

To summarize, the mean and spread of the fitted Weibull distributions vary systematically with source-receptor distance, nuclide characteristics, and release duration; and moreover this

exposure type	parameter	estimate	s.e.
air contamination	δ_1	-0.4761	1.740×10^{-2}
	δ_2	-0.4617	1.993×10^{-2}
wet deposition	δ_1	-0.6483	2.015×10^{-2}
	δ_2	-0.6957	2.393×10^{-2}

Table 3.8 Details of regression equations for dependence of Weibull means and standard errors on release duration.

variation can be expressed succinctly in a few simple equations. The likely error introduced by use of these equations rather than the MESOS data itself is assessed in Section 3.4, where so far as possible they are verified.

3.4 Verification of the fitted equations

Exposure data for the notional nuclides whose parameters are given in Table 1.4, for releases from Hannover and Stuttgart through 1973, are here used to verify the regression equations detailed in Section 3.3. Recall that although these data were calculated using MESOS, they were not used to construct the statistical model and to that extent provide independent verification of it. The receptors at which data are available are given in Table 1.3. None of these notional isotopes are inert - non-depositing - so comparisons are also made for some inert nuclide data used in model fitting.

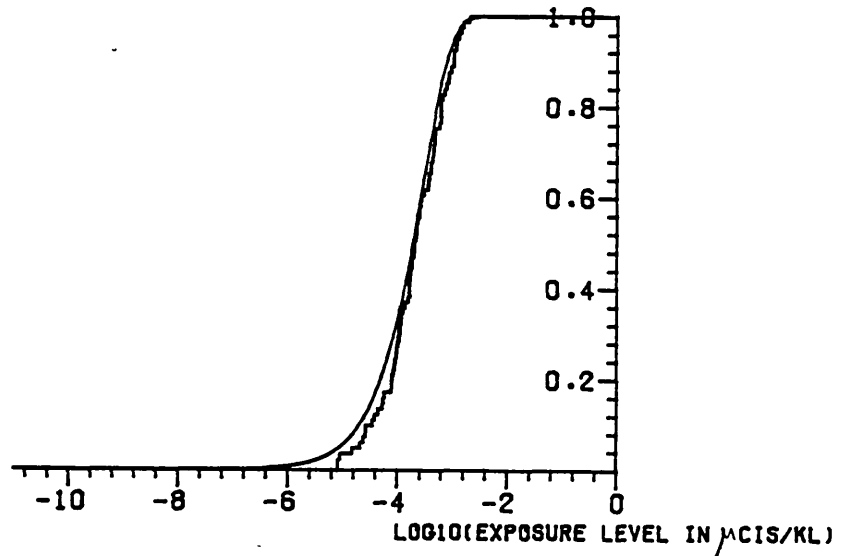
Of the imaginary nuclides in Table 1.4, only the nuclide labelled Case 3 decays - its decay constant λ corresponds to a half-life of about 3 days. The effect of decay can be assessed for depositing isotopes by the 'pseudo-nuclide' device: the generation of nuclides with imaginary half-lives using the technique explained in Section 3.3.

Only Case 1 has a deposition velocity in the same range as those used to build the model. Cases 1 and 2 have washout parameters λ_w in the same range as those used to derive the model, whereas those of Cases 3 and 4 are rather larger. Prediction of exposures for Case 1 therefore enables the model to be verified for depleting nuclides whose parameters are similar to those used to build it. Prediction for Case 2 enables the effect of extrapolating deposition velocity v_d alone to be found; and the effect on prediction of extrapolating both v_d and λ_w can be seen by considering Case 4.

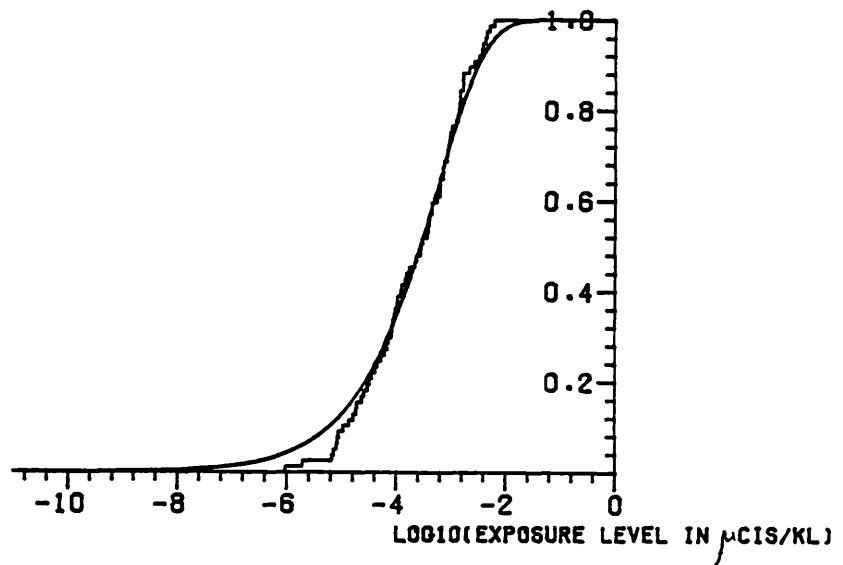
Figure 3.11 shows the comparisons for Case 1. The fit is very good in parts (a)-(c), which correspond to a nuclide with infinite half-life. It is not so good in parts (d)-(f). The biggest discrepancy occurs when in (d) the Weibull levels exceed the MESOS ones by a factor 2.5 over part of their range. Most Weibull levels are too low by a factor two or so in Figure 3.11(e). The fit in the tails of the distribution follows the same pattern as seen in Section 3.2. These results are typical for such a nuclide, and show that in the ranges of decay and deposition parameters used to fit the equations, the regression model gives air contamination - and hence dry deposition - predictions which disagree with MESOS calculations only to an unimportant extent.

The effect of using Case 1 deposition parameters but applying pseudo-nuclide calculations to the data to get an isotope with a short half-life of 9.1 hours is displayed in Figure 3.11(g, h). This extrapolation beyond the range of λ for depositing nuclides used to fit the model gives a poor fit to the lower 60% of the distribution at the receptor 100 km from its source, which suffers the joint effect of both high decay and deposition. The effect is especially marked in the lower exposures of the receptor, often due to meandering indirect puff trajectories, whose levels are overestimated by a factor at worst about 100. It is much less obvious in the comparison at 800 kilometres, whose MESOS levels are roughly a factor 6 higher than those statistically predicted. These comparisons illustrate the dangers of applying the statistical model inside the ranges of v_d and λ_w but outside that of λ for which it was derived.

In Figure 3.12 the MESOS and predicted cumulative exposure distributions for Case 2 air contamination are compared at two receptors. At short distances the statistical model tends to underestimate depletion by of dry deposition, and hence the distribution

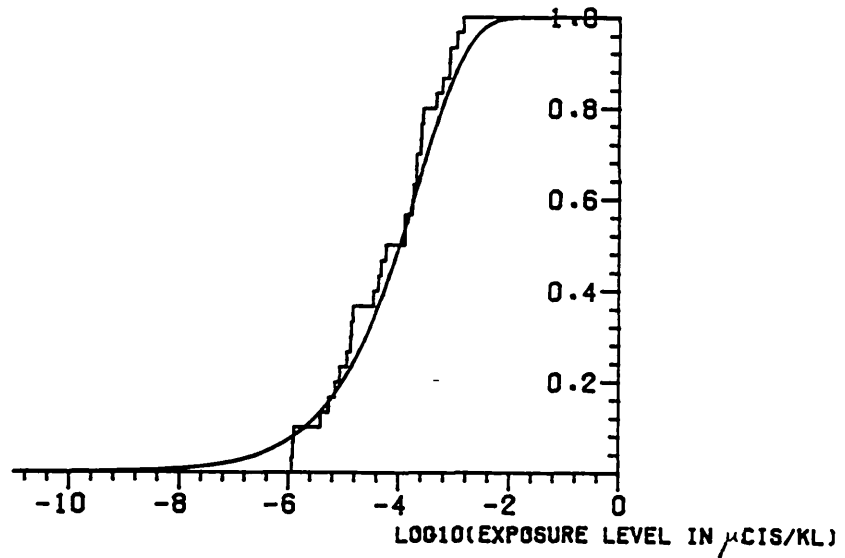


- (a) Case 1, infinite half-life, Hannover receptor 4, release duration 3 hours.

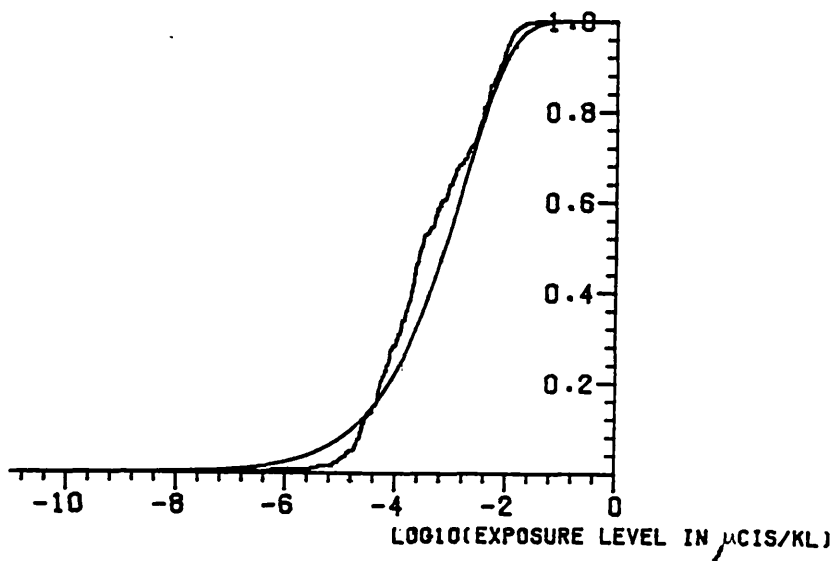


- (b) Case 1, infinite half-life, Hannover receptor 1, release duration 1 day.

Figure 3.11 : Comparison of MESOS and statistically predicted cumulative exposure distributions; Case 1 air contamination for releases from Hannover.

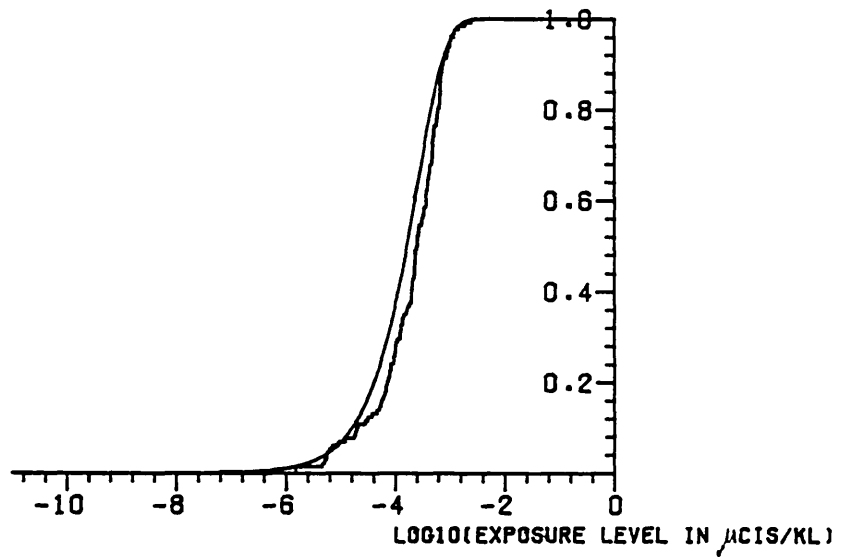


- (c) Case 1, infinite half-life, Hannover receptor 1, release duration 1 week.

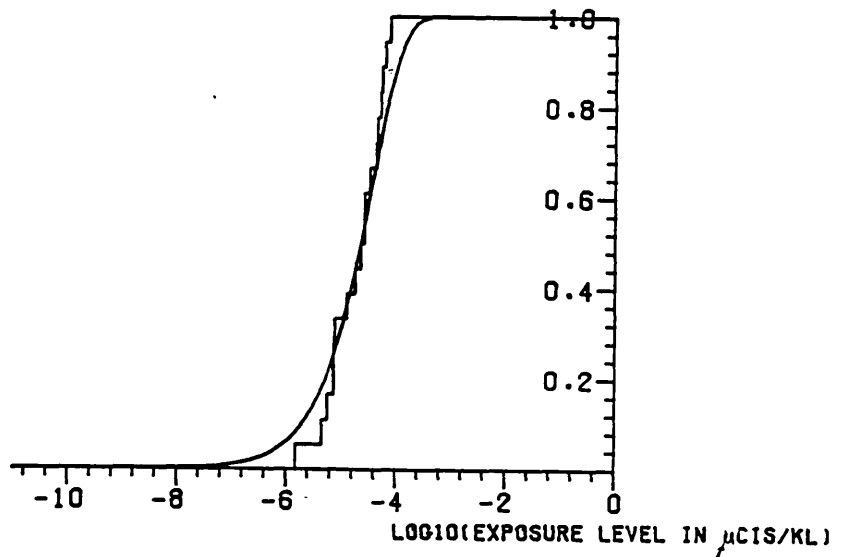


- (d) Case 1, half-life 8.1 days, Hannover receptor 13, release duration 3 hours.

Figure 3.11 : Comparison of MESOS and statistically predicted cumulative exposure distributions; Case 1 air contamination for releases from Hannover.

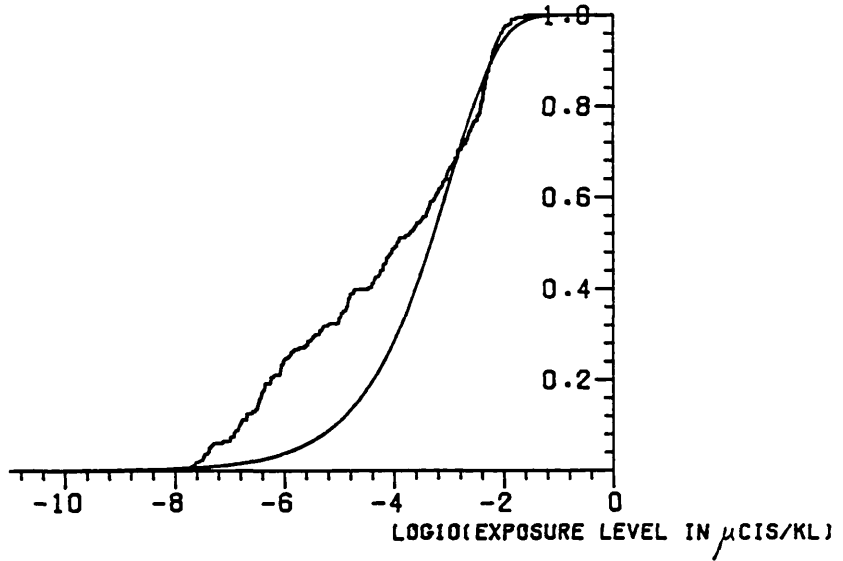


- (e) Case 1, half-life 8.1 days, Hannover receptor 16, release duration 3 hours.

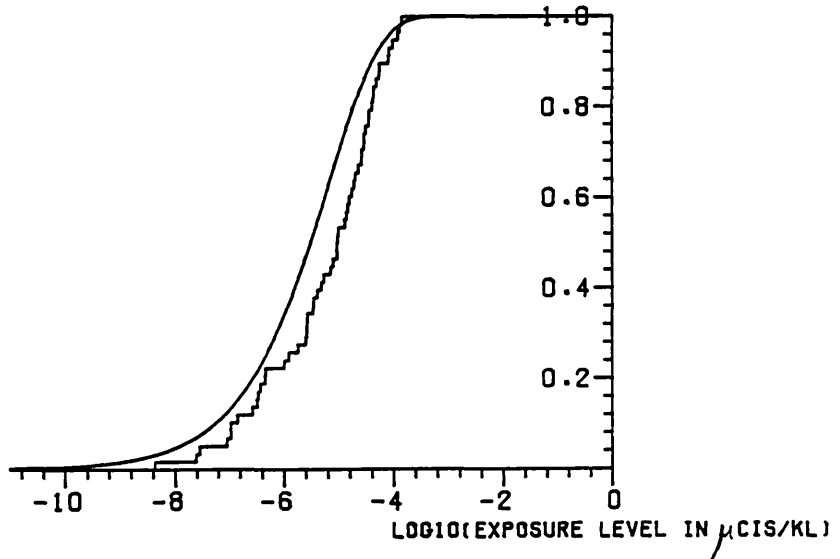


- (f) Case 1, half-life 8.1 days, Hannover receptor 16, release duration 1 week.

Figure 3.11 : Comparison of MESOS and statistically predicted cumulative exposure distributions; Case 1 air contamination for releases from Hannover.

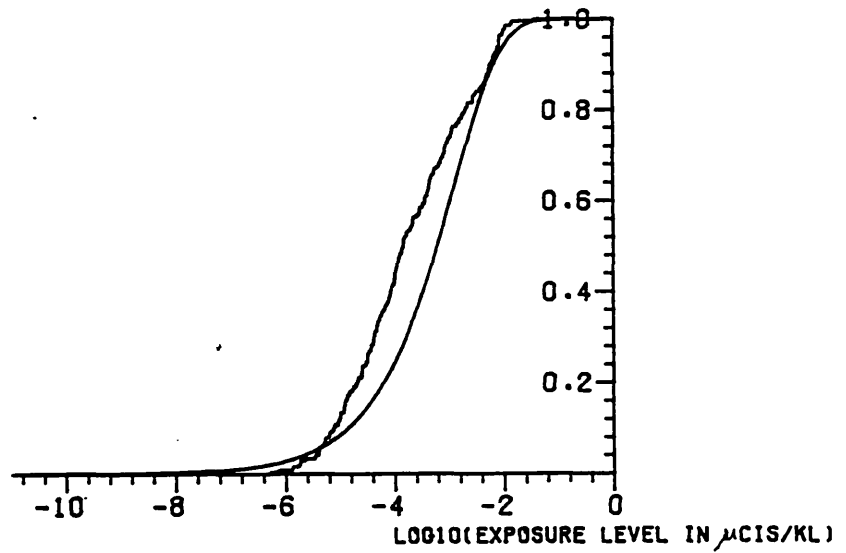


(g) Case 1, half-life 9.1 hours, Hannover receptor 1, release duration 3 hours.

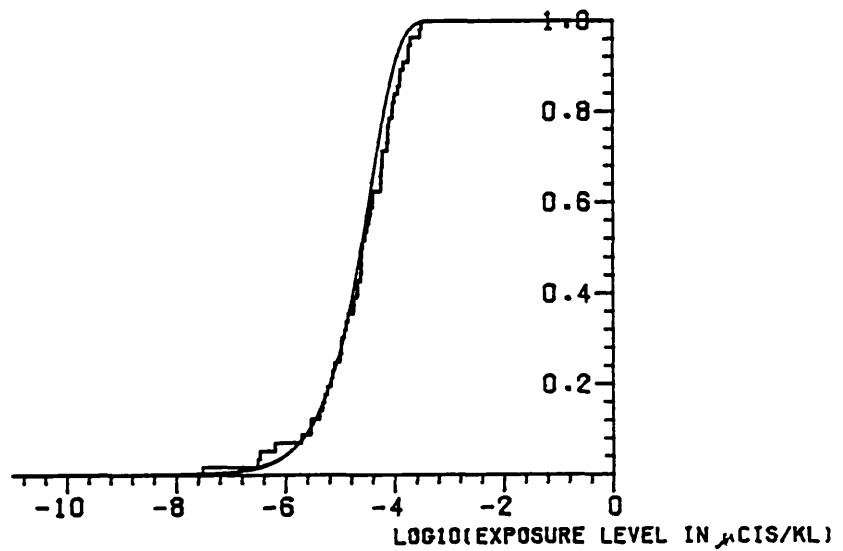


(h) Case 1, half-life 9.1 hours, Hannover receptor 4, release duration 6 hours.

Figure 3.11 : Comparison of MESOS and statistically predicted cumulative exposure distributions; Case 1 air contamination for releases from Hannover.

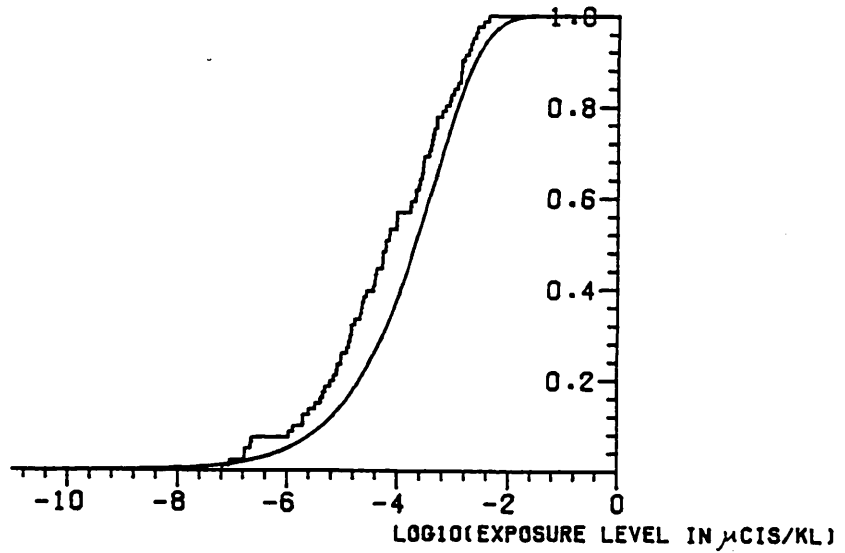


(a) infinite half-life, receptor 9, release duration 3 hours.

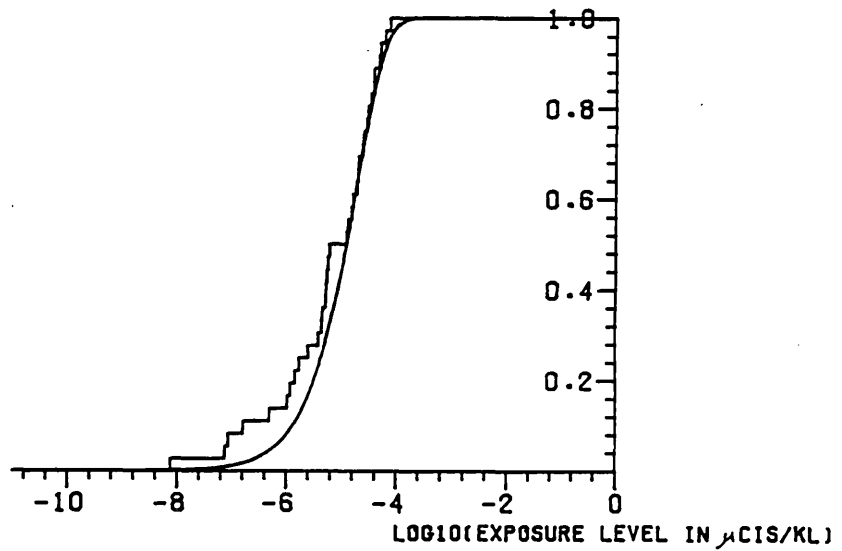


(b) infinite half-life, receptor 12, release duration 3 hours.

Figure 3.12 : Comparison of MESOS and predicted cumulative exposure distributions, Case 2 air contamination, Hannover.



(c) infinite half-life, receptor 9, release duration 1 day.



(d) infinite half-life, receptor 12, release duration 12 hours.

Figure 3.12 : Comparison of MESOS and predicted cumulative exposure distributions, case 2 air contamination, Hannover.

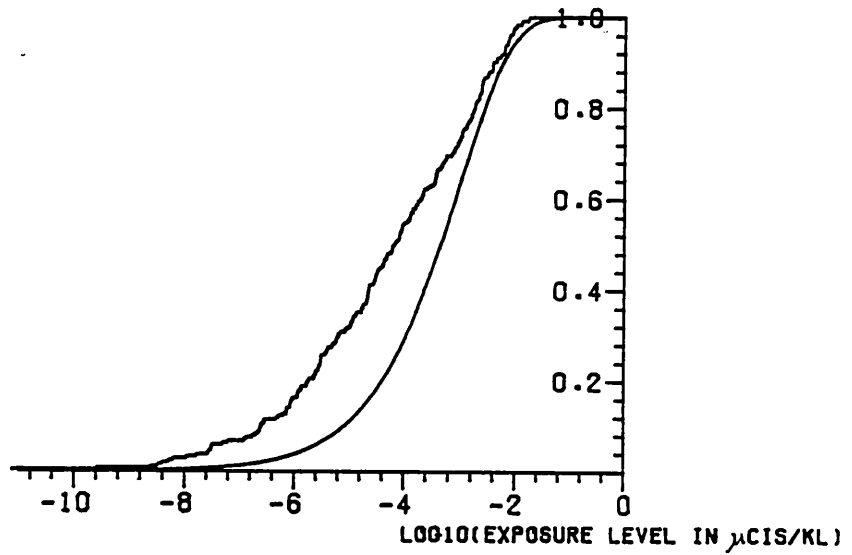
is overpredicted by a factor of about 6. Prediction is more accurate at longer distances, however. Figure 3.13 shows equivalent plots for Case 4 air contamination at two receptors, for which fit is very poor, because of the extrapolation of nuclide parameters beyond the range in which the model was derived.

The Hannover and Stuttgart MESOS computations did not include inert - non-depositing - nuclides, exposures to which cannot be obtained from existing data by using pseudo-nuclides. Therefore some of the inert nuclide data used to fit the regression model were taken and their MESOS and those predicted from the statistical model compared. Individual exposure datasets have little influence on the fitted equations, so this is close to direct verification of the model using Hannover and Stuttgart results.

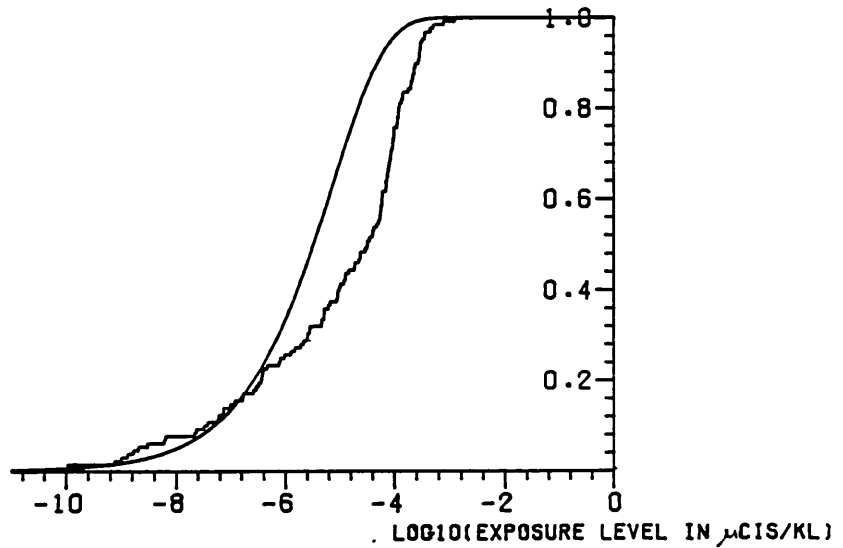
Comparisons are made in Figure 3.14 for Kr_{85} air contamination data due to releases from Heysham. The half-life of Kr_{85} is effectively infinite for the purpose of the MESOS calculations. The distribution in Figure 3.14(a) is underpredicted by a factor at most 2.5 in its top half, increasing to about six at its 10%-point. The top 70% of the data lie within a factor three of the predicted distribution. The fit is better further away from the source at receptors 3 and 10. These differences are repeated in the comparisons at longer release durations, with a large overprediction for receptor 10 data for exposures to one-day releases.

Figure 3.15 shows comparisons for Xe_{133} releases from Cadarache. They show a similar pattern; exposure levels are in general underpredicted by the statistical model by a factor three or so, but the fit is worse for long release durations.

Comparisons for Xe_{135} - whose half-life is 9.1 hours - released from Mol during 1973 are displayed in Figure 3.16. Near the source the levels are very similar over the top half of the distribution,

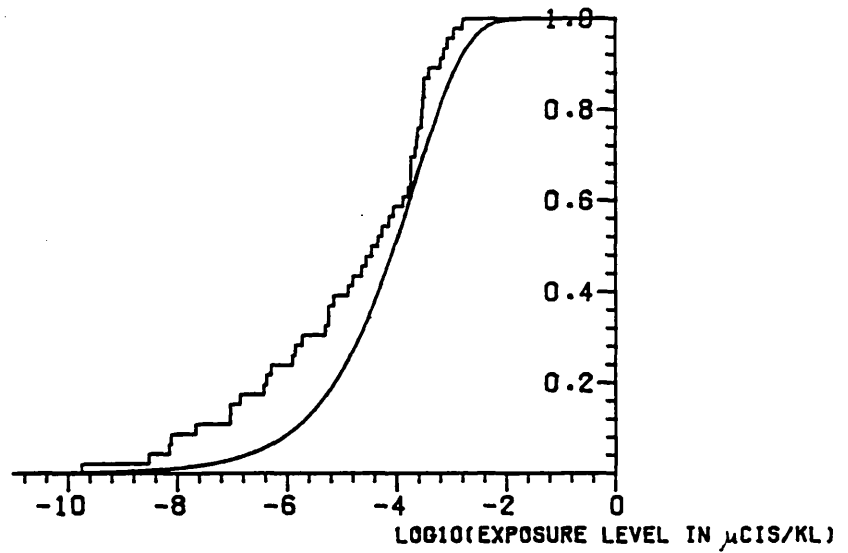


(a) infinite half-life, receptor 13, release duration 3 hours.

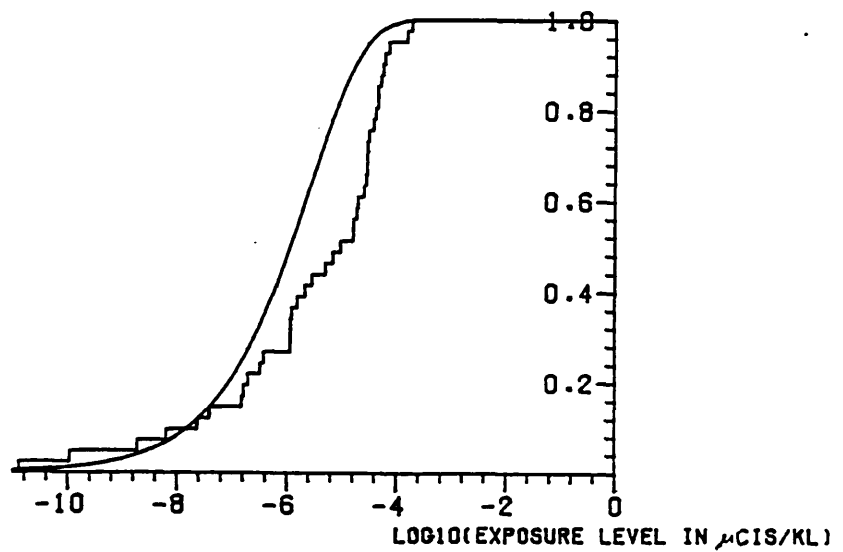


(b) infinite half-life, receptor 16, release duration 3 hours.

Figure 3.13 : Comparison of MESOS and predicted cumulative exposure distributions, Case 4 air contamination, Hannover.

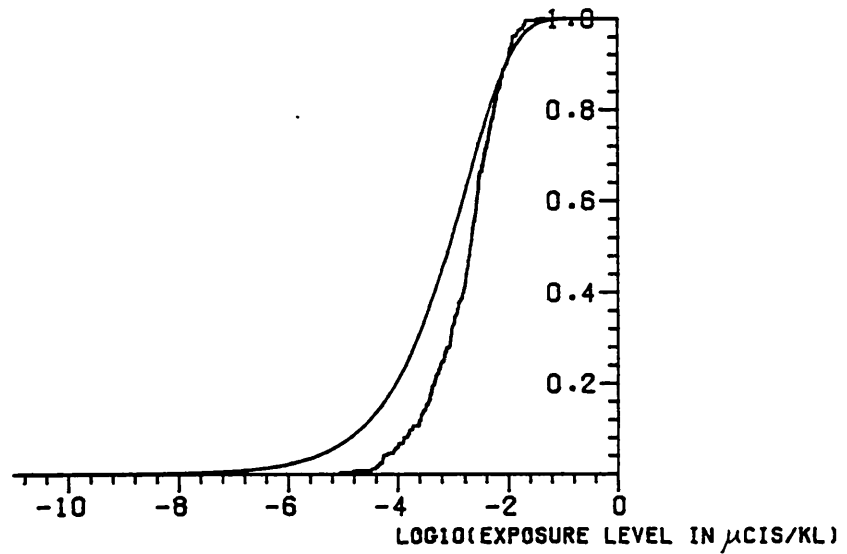


(c) infinite half-life, receptor 13, release duration 3 days.

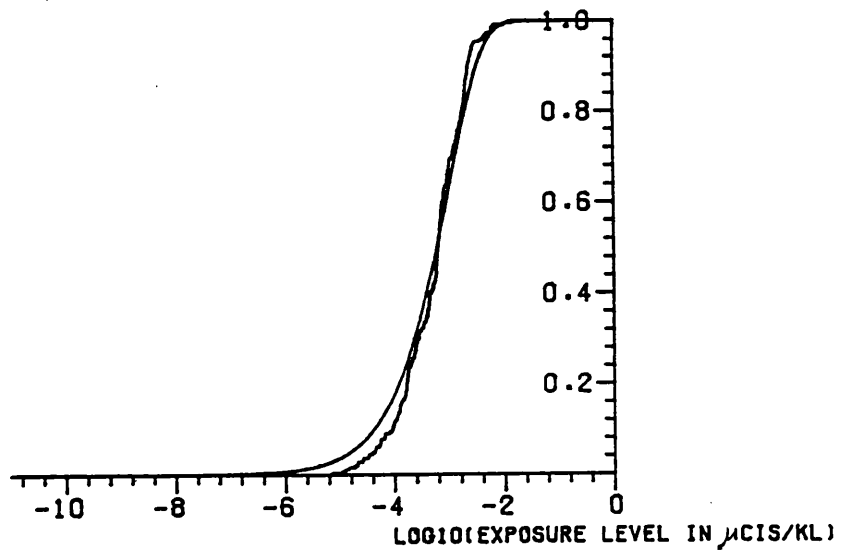


(d) infinite half-life, receptor 16, release duration 1 day.

Figure 3.13 : Comparison of MESOS and predicted cumulative exposure distributions, Case 4 air contamination, Hannover.

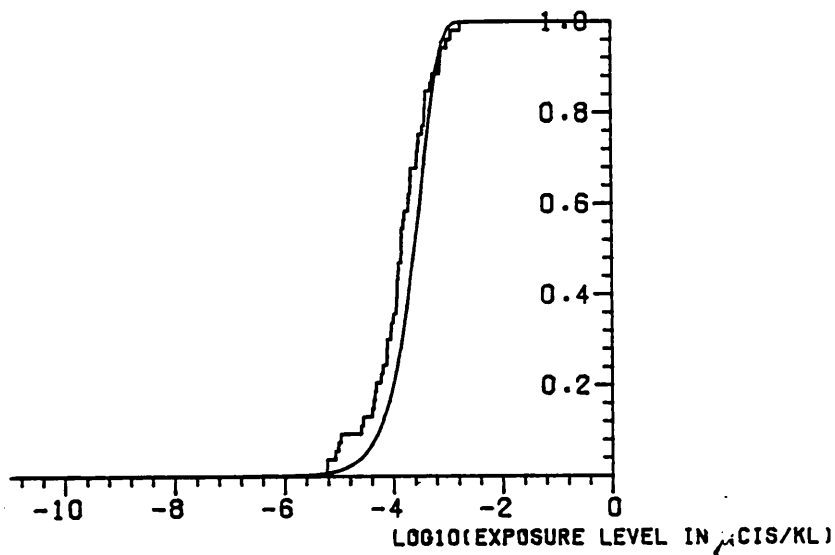


(a) receptor 1, release duration 3 hours.

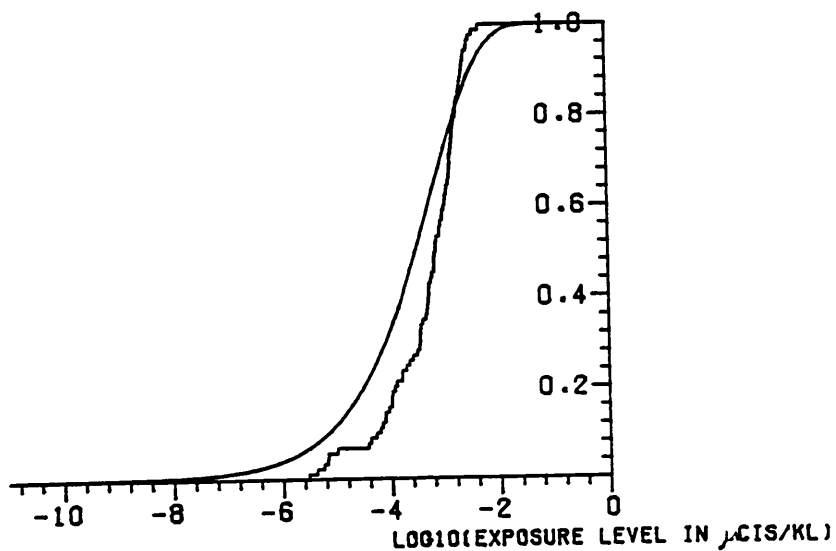


(b) receptor 3, release duration 3 hours.

Figure 3.14 : Comparison of MESOS and predicted cumulative exposure distributions, Kr_{85} air contamination, Heysham.

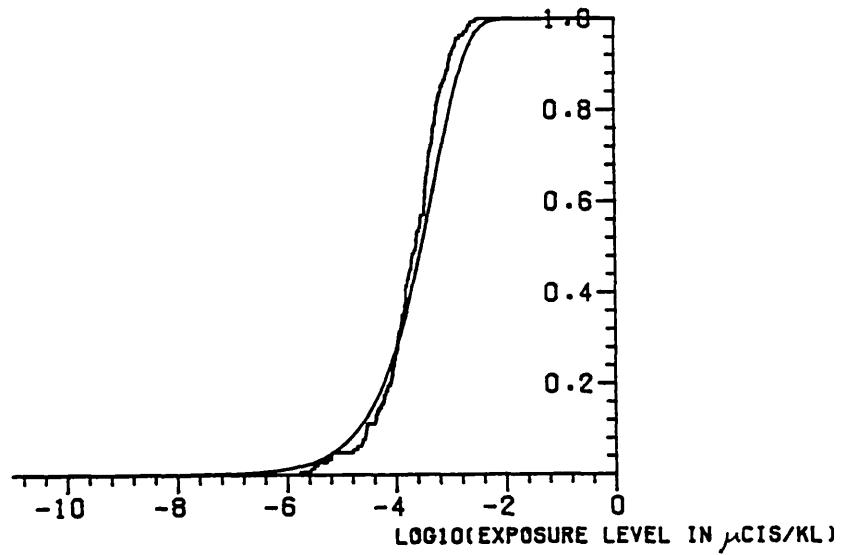


(c) receptor 10, release duration 3 hours.

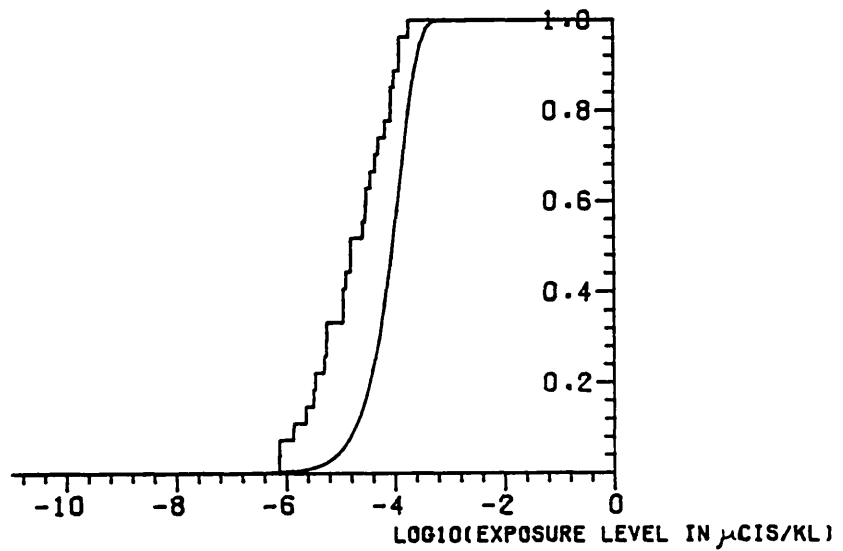


(d) receptor 1, release duration 1 day.

Figure 3.14 : Comparison of MESOS and predicted cumulative exposure distributions, Kr_{85} , air contamination, Heysham.

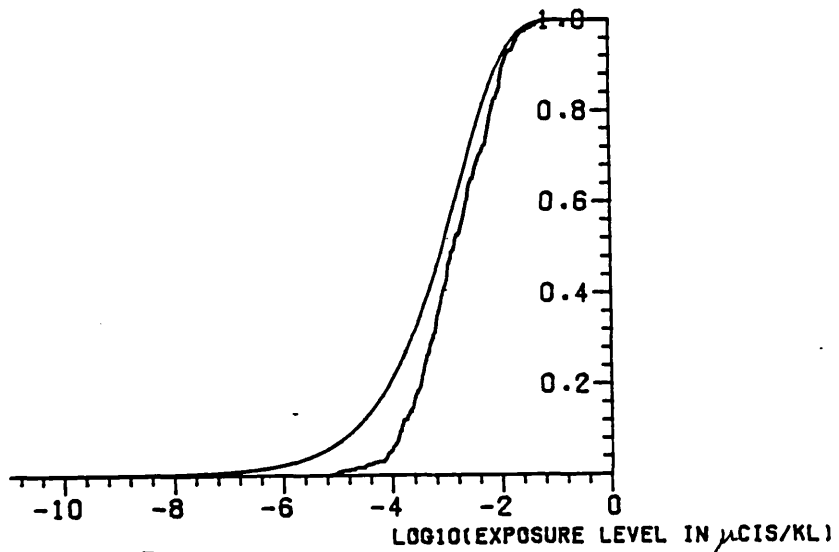


(e) receptor 3, release duration 12 hours.

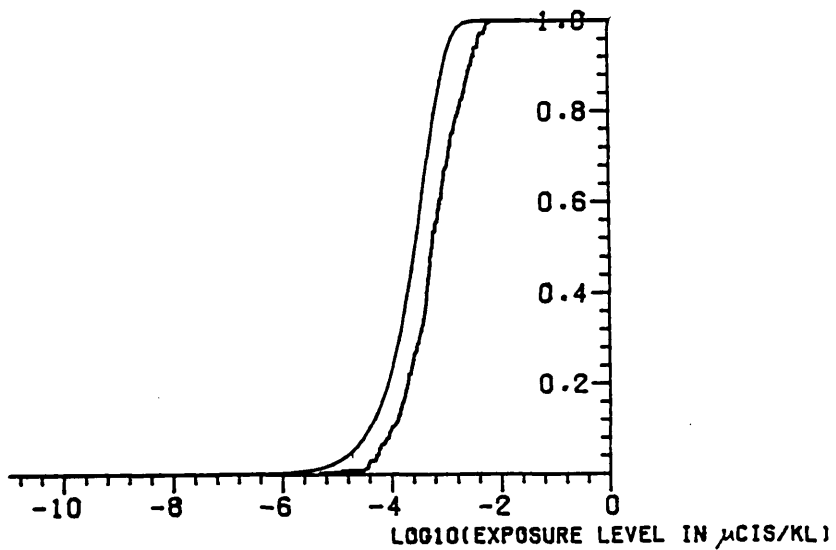


(f) receptor 10, release duration 1 day.

Figure 3.14 : Comparison of MESOS and predicted cumulative exposure distributions, Kr_{85} air contamination, Heysham.

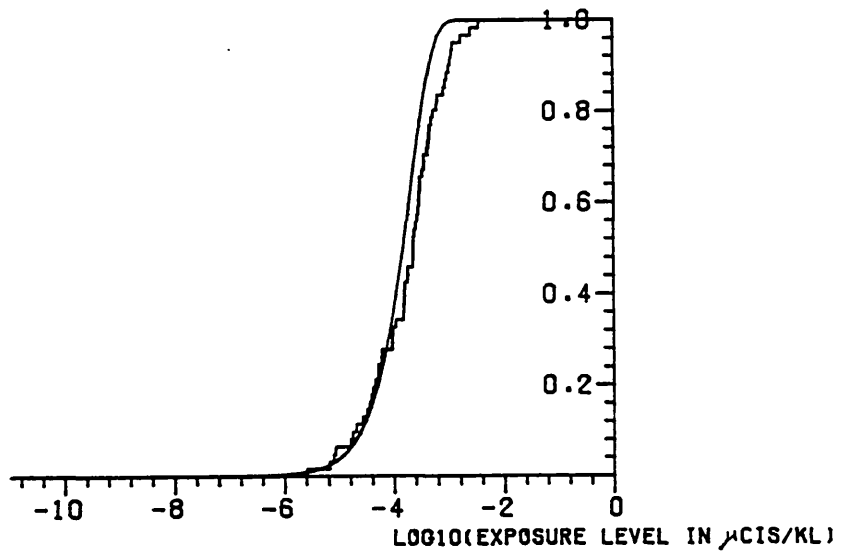


(a) receptor 7, release duration 3 hours.

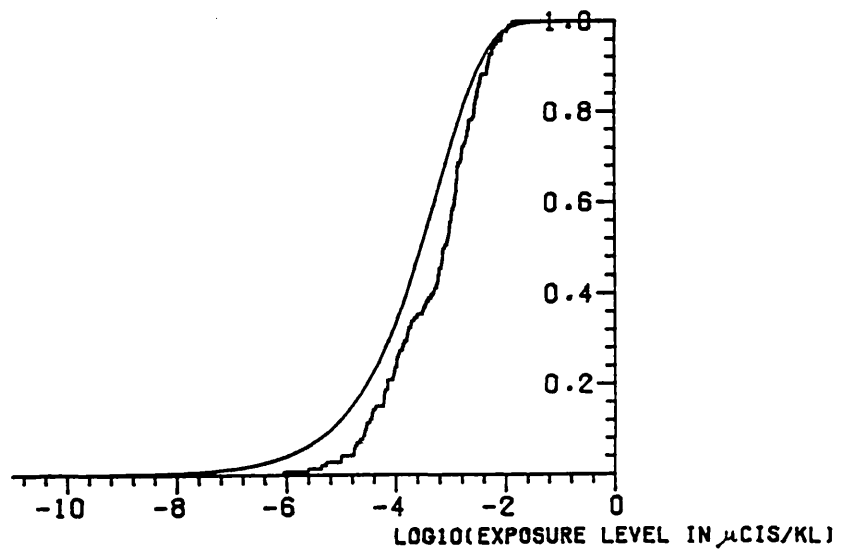


(b) receptor 11, release duration 3 hours.

Figure 3.15 : Comparison of MESOS and predicted cumulative exposure distributions, Xe_{133} air contamination, Cadarache.

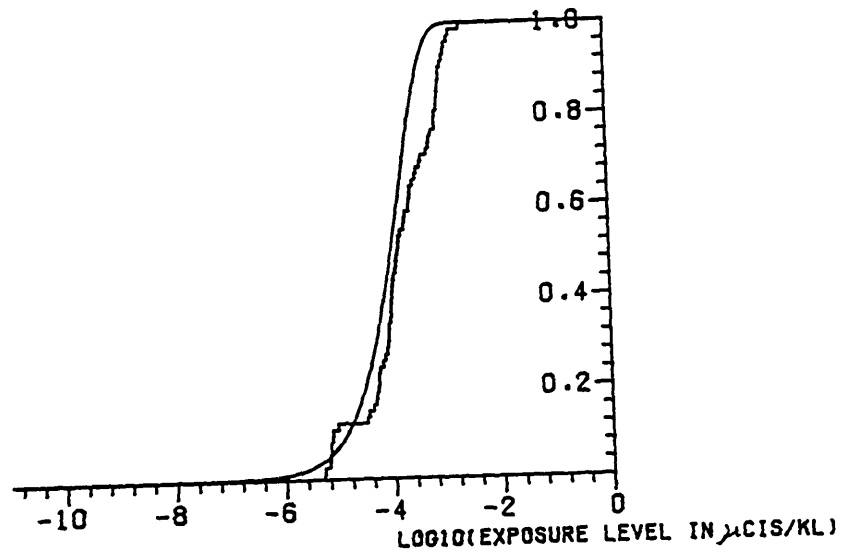


(c) receptor 16, release duration 3 hours.

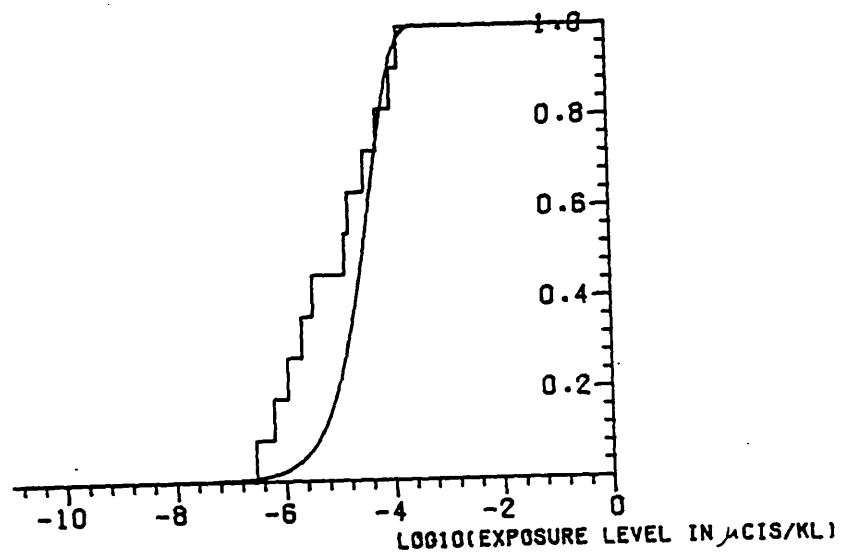


(d) receptor 7, release duration 1 day.

Figure 3.15 : Comparison of MESOS and predicted cumulative exposure distributions; Xe_{133} air contamination, Cadarache.

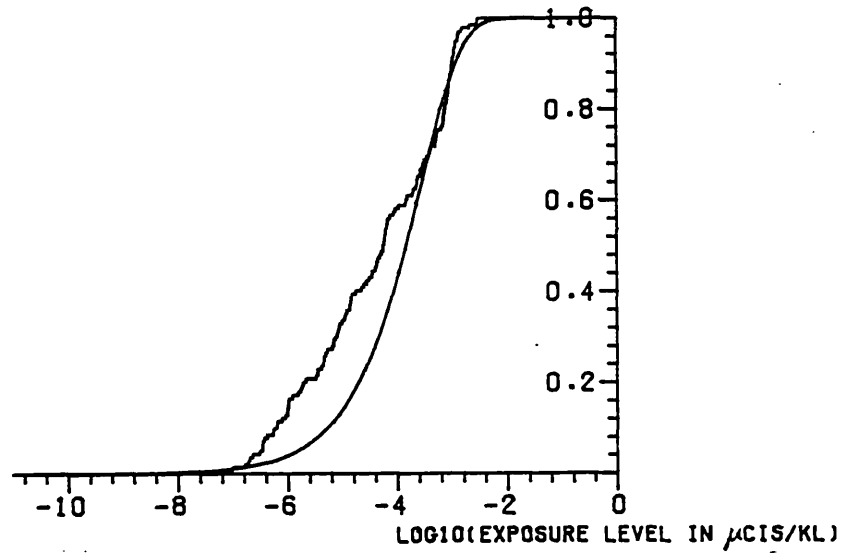


(e) receptor 11, release duration 1 day.

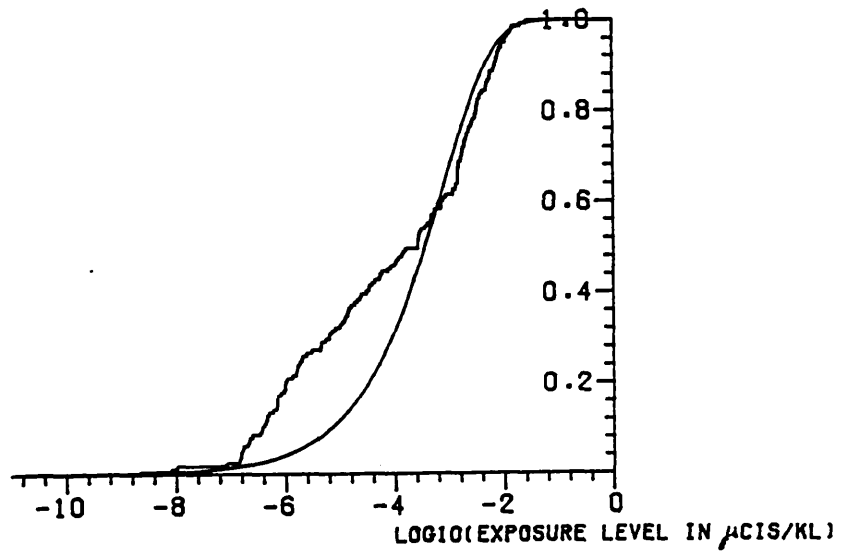


(f) receptor 16, release duration 1 week.

Figure 3.15 : Comparison of MESOS and predicted cumulative exposure distributions; Xe₁₃₃ air contamination, Cadarache.

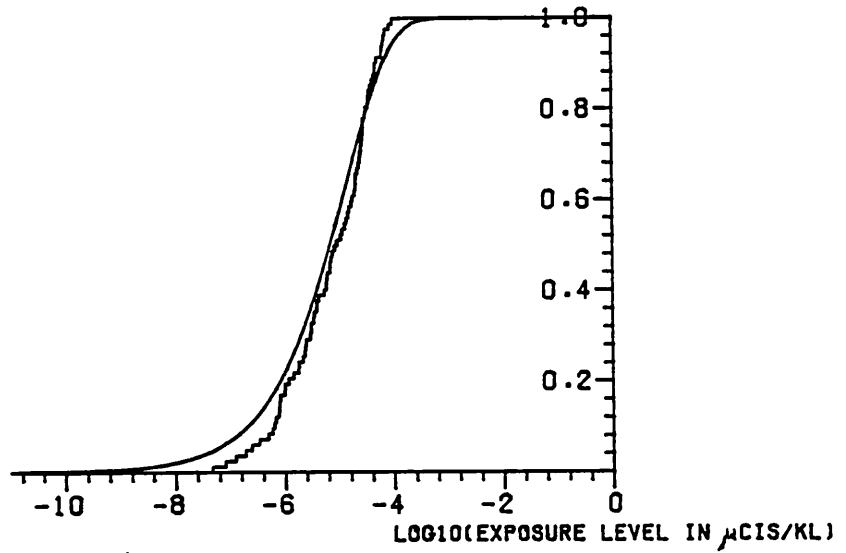


(a) receptor 2, release duration 3 hours.

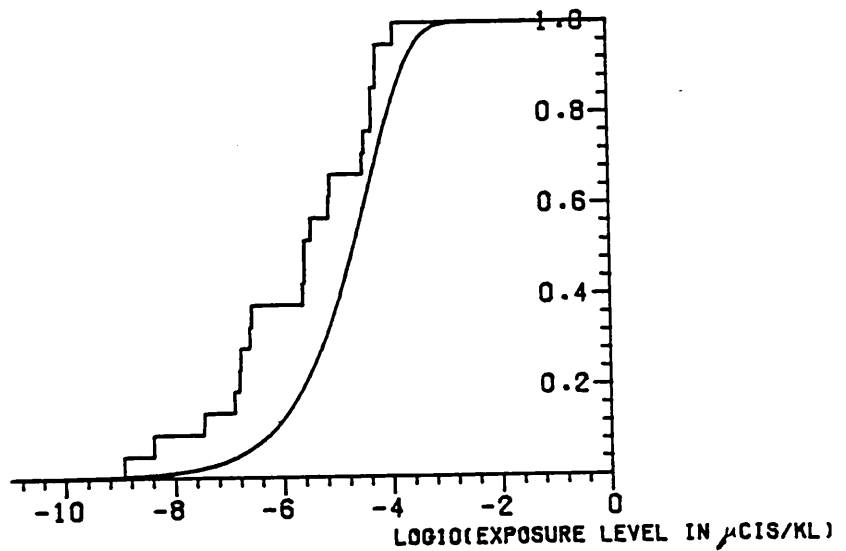


(b) receptor 13, release duration 6 hours.

Figure 3.16 : Comparison of MESOS and predicted cumulative exposure distributions; Xe_{135} air contamination, Mol 1973.

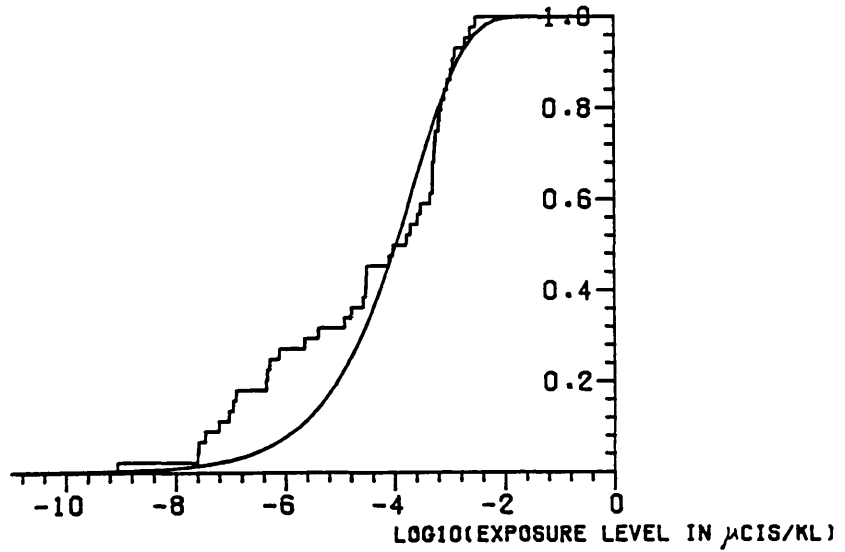


(c) receptor 16, release duration 6 hours.

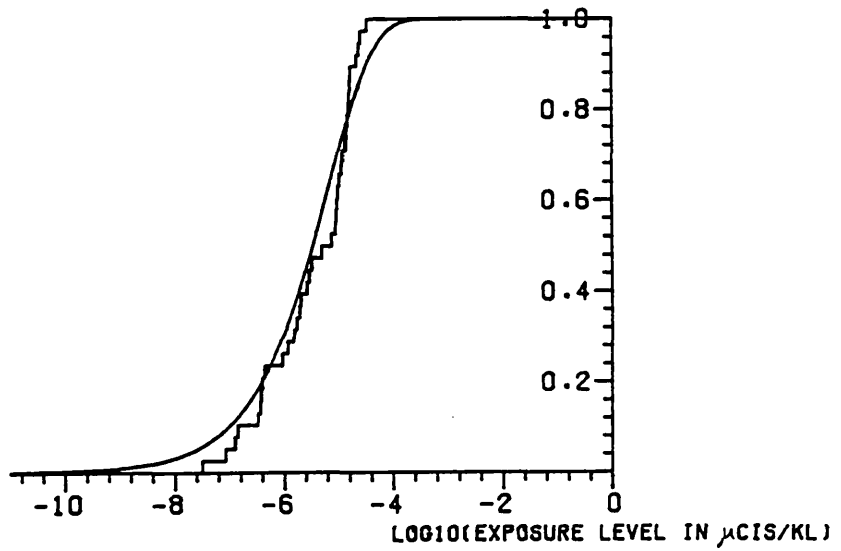


(d) receptor 2, release duration 1 week.

Figure 3.16 : Comparison of MESOS and predicted cumulative exposure distributions, Xe_{135} air contamination, Mol 1973.



(e) receptor 13, release duration 3 days.



(f) receptor 16, release duration 1 day.

Figure 3.16 : Comparison of MESOS and predicted cumulative exposure distributions; Xe_{135} air contamination, Mol 1973.

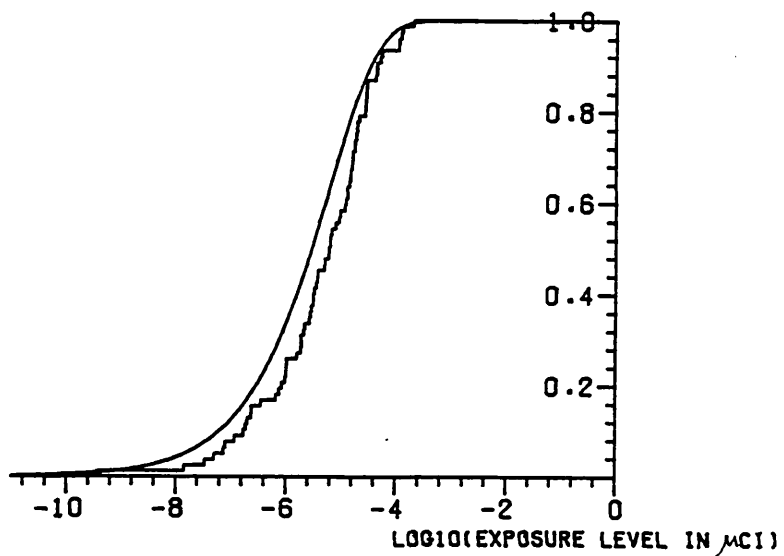
but they are overpredicted in its lower half. Further from the source the fit is better throughout the distribution. The predicted Weibull level is generally a conservative approximation and fit is best towards the upper tail of the data.

Figure 3.17 compares MESOS calculations and Weibull predictions for exposures to Case 1 wet deposition due to releases from Stuttgart. The effect of decay for a short half-life of 8.1 days is almost imperceptible. The distribution levels generally lie within a factor three of each other, but there are bigger differences in parts (e) and (f) of the figure - where the Weibull distributions are conservative.

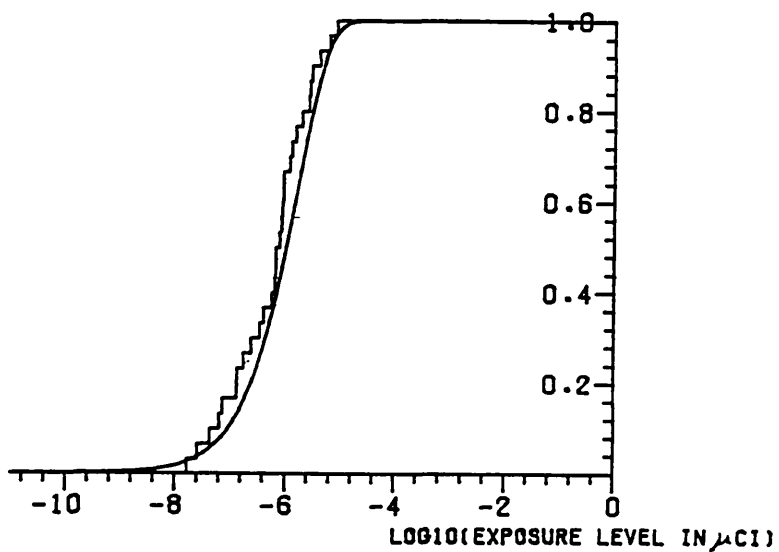
Figure 3.18 shows comparisons for Case 4 wet deposition. The effect of the high deposition parameters is to overpredict exposure levels, most notably in the lower half of the distributions. Again the statistical model tends to give high exposure levels.

To summarize: the likely error introduced by using statistically predicted rather than MESOS air contamination and dry deposition exposure distributions for nuclides whose parameters λ , v_d and λ_w lie in the range of those used in the MESOS data on which the statistical model is based is usually a factor of about two to three over the upper 70% of the exposure distribution. However it may be greater - up to a factor six or so - for exposures due to releases of duration one day or more, especially far from the source, where the predicted distributions are typically conservative. The correspondence in the top 25% of the exposure data is usually appreciably better: the Weibull approximation is generally conservative, yielding exposure levels within a factor two of the MESOS levels; however it may be in error by a factor four or so in the worst cases.

If nuclide parameters are extrapolated beyond the range used to develop the model, the differences are likely to be much worse:

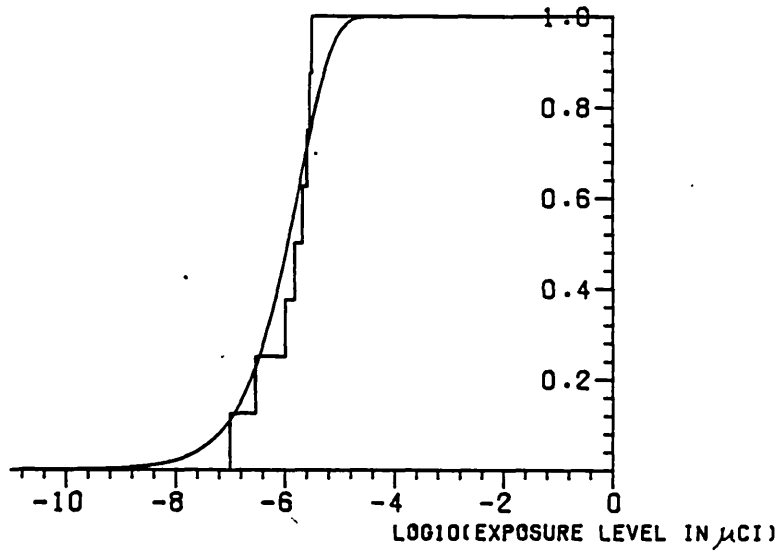


(a) infinite half-life, receptor 9, release duration 3 hours.

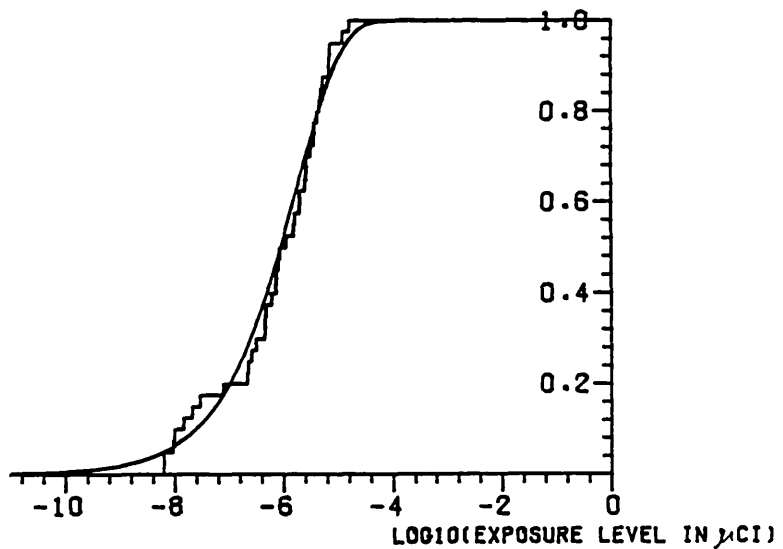


(b) infinite half-life, receptor 16, release duration 3 hours.

Figure 3.17 : Comparison of MESOS and predicted cumulative exposure distributions, Case 1 wet deposition, Stuttgart.

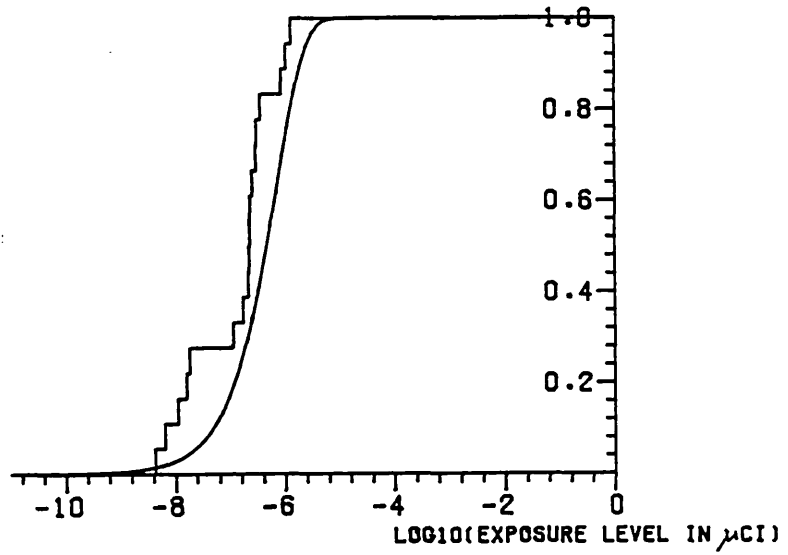


(c) half-life 8.1 days, receptor 8, release duration 3 hours.

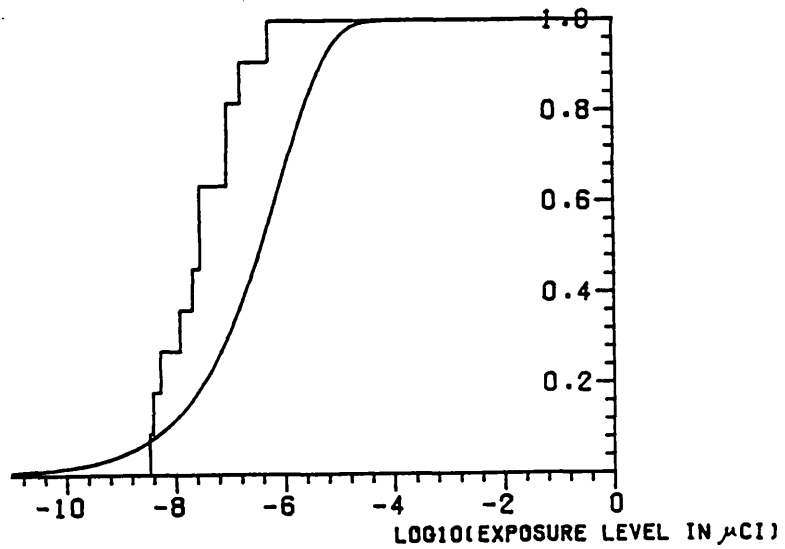


(d) infinite half-life, receptor 9, release duration 1 day.

Figure 3.17 : Comparison of MESOS and predicted cumulative exposure distributions, Case 1 wet deposition, Stuttgart.

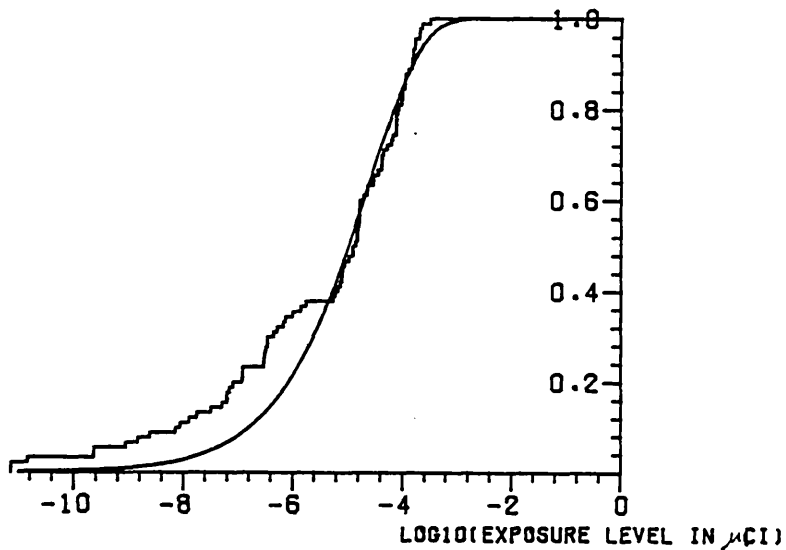


(e) infinite half-life, receptor 16, release duration 12 hours.

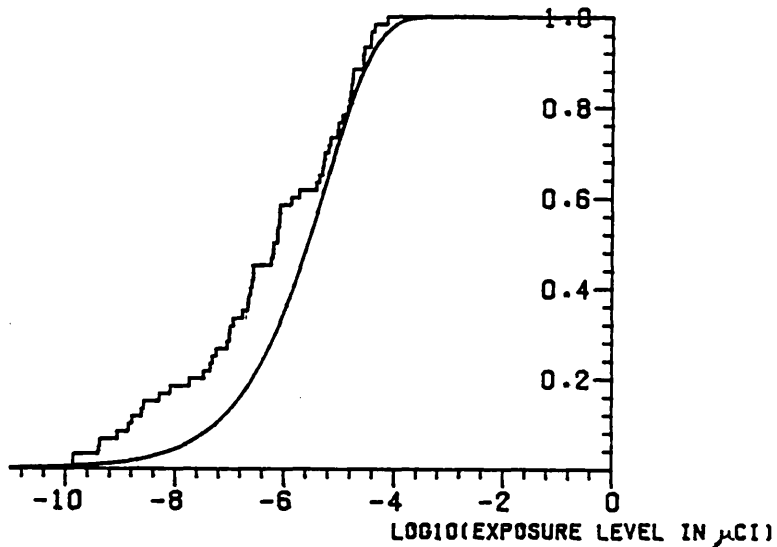


(f) half-life 8.1 days, receptor 13, release duration 1 week.

Figure 3.17 : Comparison of MESOS and predicted cumulative exposure distributions, Case 1 wet deposition, Stuttgart.

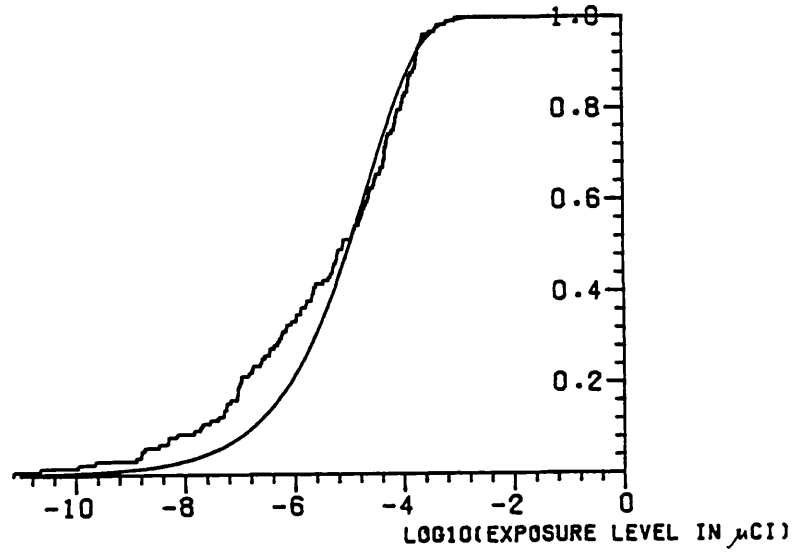


(a) infinite half-life, receptor 1, release duration 3 hours.

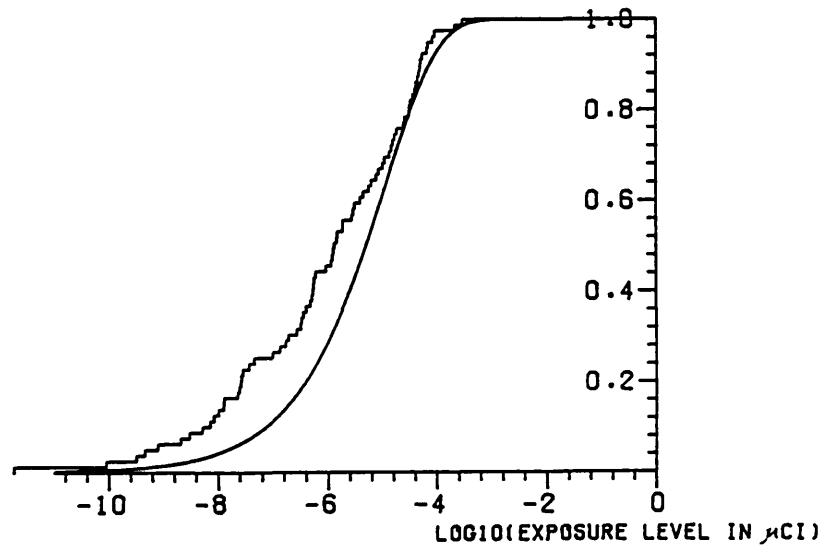


(b) infinite half-life, receptor 3, release duration 3 hours.

Figure 3.18 : Comparison of MESOS and predicted cumulative exposure distributions, Case 4 wet deposition, Stuttgart.

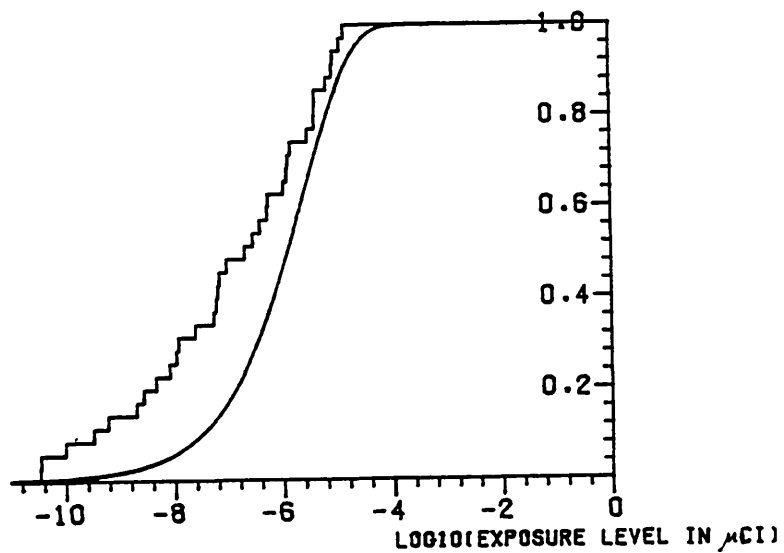


(c) half-life 8.1 hours, receptor 5, release duration 3 hours.

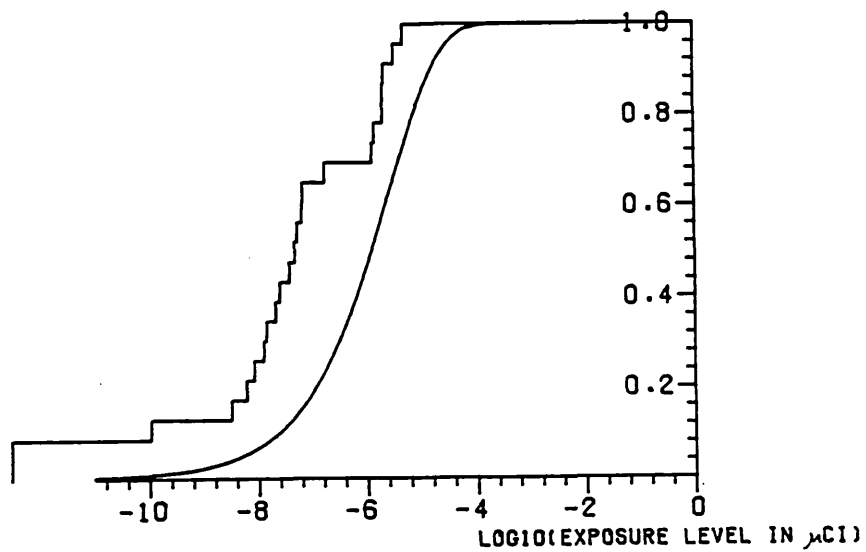


(d) half-life 8.1 hours, receptor 5, release duration 12 hours.

Figure 3.18 : Comparison of MESOS and predicted cumulative exposure distributions, Case 4 wet deposition, Stuttgart.



(e) infinite half-life, receptor 3, release duration 12 hours.



(f) half-life 8.1 hours, receptor 2, release duration 3 days.

Figure 3.18 : Comparison of MESOS and predicted cumulative exposure distributions, Case 4 wet deposition, Stuttgart.

exposure levels may then be over- or underestimated by a factor of six to ten, but possibly as great as two orders of magnitude.

The statistical model for wet deposition is generally accurate to within a factor three for releases of short duration and nuclide parameters in the range of those used to build the model. For longer durations and nuclide parameters outside the range of those in the original MESOS calculations the statistical equations overpredict exposure levels by a factor up to about ten, although fit is better in the upper tail of the data.

Fit may appear to be worse for both air concentration and wet deposition exposures due to longer release durations because fewer observations are then available to assess it.

The statistically predicted Weibull distributions do not fit the MESOS data exactly, because of differences between Weibull and MESOS exposure distributions and the gross simplification introduced by the model regression equations. To assess the likely maximum effect of the discrepancies, suppose that for some source and receptor, the probability distribution function of exposure of a given type to some nuclide is $F(\cdot)$. That is,

$$\text{Prob}(\text{ exposure } < x) = F(x).$$

Let $r(x)$ be the risk to some activity or individual at the receptor due to an exposure of size x . The function $r(\cdot)$ will invariably be increasing and will often be linear: $r(x) = ax$, for some positive constant a . It may be a dose-response curve for long- or short-term effects of ionising radiation on humans, for example. In some such cases it has been suggested that a quadratic curve $r(x) = ax+bx^2$ ($a, b > 0$) is more likely: see Hemming et al.(1983). The exact form of $r(\cdot)$ is immaterial here, but it will increase at least linearly.

The expected risk to the individual or activity due to a release

resulting in an exposure distribution $F(\cdot)$ at the receptor is

$$\begin{aligned} \int_0^{\infty} r(x) dF(x) &= \int_0^{x_{75}} r(x) dF(x) + \int_{x_{75}}^{\infty} r(x) dF(x) \\ &= I_1 + I_2, \end{aligned}$$

say, where x_{75} is the value of x for which

$$F(x_{75}) = 0.75.$$

The comparison of MESOS and statistically predicted exposure levels above suggests that levels above x_{75} may be in error by at very worst a factor four or so, and those below x_{75} by a factor at very worst ten, provided the equations are used within their range of validity. Making the unrealistic assumption that levels below and above x_{75} may vary independently, the biggest probable range of error for the expected risk is

$$aI_1'/10 + aI_2'/4 \leq \text{expected risk} \leq 10I_1' + 4I_2',$$

where $I_1' = \int_0^{x_{75}} x dF(x)$ and $I_2' = \int_{x_{75}}^{\infty} x dF(x)$.

Table 3.9 shows for two Kr_{85} exposure distributions - chosen as being relatively homogeneous - the percentage contribution to the distribution mean, $100\% \times \int_0^y x dF(x) / \int_0^{\infty} x dF(x)$, for several values of y . About three-quarters of the mean exposure is due to the top 25% of the exposure distribution. In distributions for exposures to decaying or depositing nuclides the relative contribution by the the highest exposure levels is even greater.

Thus at worst the likely range of error of the expected risk calculated using a statistically predicted distribution rather than a MESOS one is $(R/5, 5R)$, where R is the risk which would have been found were the MESOS data used directly. Usually the expected risks based on Weibull and MESOS distributions will be much closer.

Mol 1976 Kr₈₅ receptor 4

Mol 1976 Kr₈₅ receptor 6

level y $\mu\text{Cism}^{-3} \times 10^{-4}$	#obs \geq y	% contribution to mean Q(y)	level y $\mu\text{Cism}^{-3} \times 10^{-4}$	#obs \geq y	% contribution to mean Q(y)
0	258	0	0	431	0
1.22	143	2.5	5.02	208	5.26
2.45	111	7.2	10.04	146	13.8
3.67	88	12.8	15.06	101	24.6
4.89	72	18.3	20.08	80	31.8
6.11	59	24.2	25.1	48	45.9
7.34	50	29.2	30.12	38	51.2
8.56	40	35.7	35.14	29	56.8
9.78	31	42.5	40.17	26	59.1
11.00	28	45.0	45.19	23	61.6
12.23	26	46.9	50.21	20	64.4

TABLE 3.9: Percentage contributions to mean exposure by excesses over thresholds.
Kr₈₅ distributions

$$Q(y) = \int_0^y x dF(x) / \int_0^\infty x dF(x) \times 100\%$$

For a linear risk function $r(x) = ax$ the mean risk predicted using the statistical model is $a\hat{M}_t$, where \hat{M}_t is found using equations 3.3.1 and 3.3.5 with appropriate parameter values in Table 3.8. For exposures to wet deposition use must be made of equations 3.3.3 and 3.3.5.

Confidence intervals for the MESOS value of M_t may be based on the Normal distribution of $\log\{\hat{M}_t\}$ under the model. For exposures to time-integrated air contamination, conditionally upon the estimated parameter values, the variance of $\log\{\hat{M}_t\}$ is $S^2 = 56.3/363 = 0.1551$, and the unconditional variance of $\log\{\hat{M}_t\}$ is

$$\text{Var}(\log\{\hat{M}_t\}) = S^2 + \underline{x} D \begin{vmatrix} V & 0 \\ 0 & 1 \end{vmatrix} D \underline{x}^t.$$

Here \underline{x} is the row vector

$$(1, d\lambda, dv_d, d\lambda_w, \log\{d\}, \log\{t/3\}),$$

D is the 6×6 matrix whose diagonal elements are $s.e.(\hat{\theta}_1)$, $s.e.(\hat{\theta}_2), \dots, s.e.(\hat{\theta}_5), s.e.(\hat{\delta}_1)$, and V is the 5×5 matrix of correlations between the parameter estimates $\hat{\theta}_i$. Then a $(1-2\alpha) \times 100\%$ confidence interval for the MESOS value of M_t is

$$(\hat{M}_t \exp\{\tau z_\alpha\}, \hat{M}_t \exp\{-\tau z_\alpha\}),$$

where $\tau = \sqrt{\text{Var}(\log\{\hat{M}_t\})}$ and $\Phi(z_\alpha) = \alpha$ for $\alpha < 0.5$, and $\Phi(\cdot)$ is the standard Normal distribution. For wet deposition exposures the $\hat{\theta}_i$ are replaced by the $\hat{\eta}_i$, $S^2 = 125.4/572 = 0.2192$, and V is taken from Table 3.7. Confidence limits for standard deviations of MESOS exposure distributions may be obtained similarly.

In view of the discrepancies between predicted Weibull and MESOS exposure levels, use of standard statistical techniques to find confidence intervals for MESOS exposure levels based on their

predicted values would be spurious, giving a false impression of the accuracy of the statistical model. The rather qualitative assessment above seems preferable.

Table 3.9 explain the stress laid above on the fit in the upper tails of the exposure distributions and the relative indifference to their lower tails. It shows the importance of statistical methods for high exposures - discussed in the next few chapters and used for data-analysis in Chapter 8.

4. EXCESSES OVER HIGH THRESHOLDS

The work presented in the next few chapters was motivated by the need for statistical techniques for the analysis of the upper tails of the MESOS exposure data; in particular methods for studying their dependence on the distance of the receptor from the source, nuclide decay and depletion characteristics, release duration, and other potential explanatory variables. The approach taken is to consider only the events occurring when contamination exceeds some high threshold level.

The problem of statistical inference for such excesses may arise in any area of science where analysis of sequences of observations and their extremes is important: hydrology, metallurgy, meteorology, oceanography, medicine, air pollution, and many others. The sequences of observations may be independent or may exhibit trend, seasonality, and long- or short-term dependence, all of which will probably complicate analysis. They may be related, for example several sequences of water levels at different points along the same river, or pollution levels at a number of different locations relative to a common point source of contaminant. It may then be required to link the sequences using covariates, which might be hydrological variables in the first case, and meteorological ones in the second.

Only exceedances over upper thresholds are considered, since those below lower ones may be treated simply by negating the data.

A number of approaches to modelling upper extremes of such sequences may be possible, depending on the structure and complexity of the data. If the sequences are fairly long, the classical method treating annual maxima of consecutive periods of equal length, for example years, months or days, of the series as independently and identically distributed in one of the extreme value distributions is

often espoused, following the work of Gumbel(1958). The method is commonly used in environmental applications and has met with success. As at present used, it has some drawbacks:

(i) parameter estimation in the presence of covariates, formal tests of goodness of fit (as opposed to informal graphical checks), and studies of influence are rather under-developed for the method, except in special cases - see Stephens(1977);

(ii) more seriously, its use of data is rather uneconomical and inference based on short sequences is likely to be unreliable. This last difficulty is common to all methods of analysis for sample extremes; the point here is that if, say, k years data are available, then - other things being equal - inference based on the upper k order statistics ought to be at least as good as inference based on the annual maxima. This is because order statistics are at least as great as the annual maxima and so might be expected to be more informative about the upper tail of the distribution.

Weissman(1983) discusses and gives references to methods of analysis based on the use of a fixed number k of upper order statistics in simple random samples when the sample size $n \rightarrow \infty$ and $k/n \rightarrow 0$. Under these conditions and some weak assumptions about the tail behaviour of the distribution function of the original sample, the joint asymptotic distribution of the order statistics can be derived and inference based upon it. These techniques have been developed partly to perform analysis in life-testing situations where a very large number of similar components are in use and it is hoped to estimate their minimum life. In this situation dependence on covariates may not be important, but no doubt the ideas could be extended to allow for it if necessary. However the joint density of the k statistics is potentially complicated and the method seems likely to be unwieldy in all but fairly simple situations.

Other suggestions with less theoretical basis have been made. Berger et al.(1982) report the study of high levels of atmospheric sulphur dioxide by fitting two-parameter exponential distributions to concentrations exceeding high threshold levels; the fit obtained is apparently very good. The authors of the Flood Studies Report(1975), inter alia, advocate modelling extremes of river-flow series by fitting the two-parameter exponential distribution to fixed numbers of upper order statistics, a variant of the 'Peaks Over Threshold' methods developed by hydrologists.

This family of models was first explicitly proposed by Todorovic and his co-workers (Todorovic and Zelenhasic, 1970; Todorovic and Rouselle, 1971; Todorovic and Woolhiser, 1972), in order to analyse the extremes of river-flow series. See also the Flood Studies Report(1975, I, Section 2.7). Essentially the idea is this: impose a high threshold level on the data and ignore what goes on beneath it - typically giving clusters of exceedances and their times of occurrence; then model maxima Y_i (peaks) of different clusters as independent one-parameter exponential variables, and the times T_i at which they occur as a Poisson process. The most flexible version of this in the hydrological literature is probably that of North(1980), whose peak epochs T_i are a seasonally varying Poisson process, and whose peak sizes Y_i , conditionally upon $T_i=t_i$, have independent exponential distributions with seasonally varying parameter $\lambda(t_i)$. Smith(1983), who reviews these models from a statistical viewpoint, observes that the twin assumptions of Poissonness of peak times and conditional independence of peak values are supported by empirical evidence and theoretical arguments, as in Todorovic(1979). However there are two issues worth raising in this context.

The first is the clustering of exceedances which occurs in practice. Although very many commonly studied stochastic processes

do not exhibit clustering of extremes (Leadbetter, Lindgren, and Rootzén, 1983), it is an unfortunate fact that this does appear in data - caused by storms, tidal surges, and other natural events which persist over a few hours or days but not weeks or months.

Incidentally, one broad implication of this is that it may be dangerous to model a set of data by, for example, a Gaussian process or a time-series model with Normal margins and then to try to derive extremal behaviour from the estimated parameters, since the original model may not adequately describe the tails of the data, exactly where fit is hardest to assess. This point is also made by DeMouchel(1983), in the context of fitting stable laws to whole samples to estimate their behaviour in the tails: he concludes that models specifically for the tails are likely to be less misleading, and goes on to suggest ideas enlarged upon below. Smith(1983) propounds a model of clustering and goes on to fit it successfully to wave-height data, but does not allow for the much more complicated possibilities which may arise if exceedance sizes in the same cluster are allowed to be dependent - a hard and unsolved problem.

The second issue is the choice of distribution with which to model the excesses themselves. Although the exponential distribution seems to fit the sulphur dioxide data mentioned above and some hydrological and oceanographic data well, it lies in the max-domain of attraction of the type I (Gumbel) extreme-value distribution and so cannot possibly model all tail behaviour for exceedances, which could include Pareto-type tails and upper endpoints. Following Smith(1983), Davison(1983), and DeMouchel(1983), the generalized Pareto distribution

$$F(y) = \begin{cases} 1 - (1 - ky/\sigma)^{1/k} & (k \neq 0) \\ 1 - \exp(-y/\sigma) & (k=0), \end{cases}$$

$$(-\infty < k < \infty , \sigma > 0 ; 1-ky/\sigma > 0),$$

is proposed as the natural form for the purpose. The justification for this is Pickands'(1975) important result that a continuous distribution function $G(\cdot)$ with $y_1 = \sup\{ x:G(x)<1 \}$ satisfies the condition

$$\lim_{u \rightarrow y_1} \inf_{0 < \sigma < \infty} \sup_{0 < x < \infty} | \{ 1-G(u+x) \} / \{ 1-G(u) \} - \bar{F}(x;\sigma,k) | = 0,$$

where $\bar{F}(y;\sigma,k) = 1-F(y)$, if and only if G lies in the max-domain of attraction of an extreme-value distribution. The practical consequence of this is that the generalized Pareto distribution may be regarded as the natural parametric family for exceedances over high thresholds, since their distribution functions can be made arbitrarily uniformly close together for high enough thresholds u and a suitable choice of σ . The approximation has potentially wide applicability, since all common continuous statistical distributions lie in the domain of attraction of an extreme-value distribution.

As Smith(1983) points out, there is a close connection with the classical generalized extreme-value distribution

$$G(y) = \begin{cases} \exp\{ -(1-k(y-\alpha))/\sigma^{1/k} \} & (k \neq 0) \\ \exp\{ -\exp(-(y-\alpha)/\sigma) \} & (k=0), \end{cases}$$

$$(-\infty < k, \alpha < \infty , \sigma > 0 ; 1-k(y-\alpha)/\sigma > 0),$$

as follows. Consider a simple random sample of n variates Y_1 with upper endpoint $y_1 < \infty$, and suppose that an increasing sequence of thresholds $u_n \rightarrow y_1$ is imposed. Take some value $x > 0$ and consider the distribution of the maximum M_n of the sample, which may be written

$$\text{Prob}(M_n \leq u_n + xv_n) = \sum_{i=0}^n \binom{n}{i} F(u_n)^{n-i} \{ 1-F(u_n) \}^i \text{Prob}(Y \leq u_n + xv_n)^i.$$

The right-hand side of this is

$$\sum_{i=0}^n \binom{n}{i} \{ 1 - \bar{F}(u_n) \}^{n-i} \bar{F}(u_n)^i \{ 1 - \bar{F}(u_n + xv_n) / \bar{F}(u_n) \}^i,$$

where $\bar{F}(y) = 1 - F(y)$ is the survivor function of the variables Y_i . Provided that $F(\cdot)$ is smooth enough for the sequences $\{u_n\}$ and $\{v_n\}$ to be chosen so that $n \bar{F}(u_n) \rightarrow \lambda > 0$ and to make $\bar{F}(u_n + xv_n) / \bar{F}(u_n)$ tend to the generalized Pareto limit as $n \rightarrow \infty$, then by the usual Poisson limit to the binomial distribution we have

$$\begin{aligned} \text{Prob}\{ (M_n - u_n) / v_n < x \} &\rightarrow \sum_{i=0}^{\infty} e^{-\lambda} \{ \lambda - \lambda(1 - kx/\sigma)^{1/k} \}^i / i! \\ &= \exp\{ -\lambda(1 - ky/\sigma)^{1/k} \} \end{aligned}$$

for some $\sigma > 0$ and real k - the generalized extreme-value distribution. This heuristic argument shows the connection between the limiting distributions clearly: for a given underlying $F(\cdot)$ the values of k for both the limit distribution of the maximum and the limiting conditional exceedance distribution are exactly the same.

The next few chapters of this thesis are laid out as follows.

In Chapter 5 some of the basic statistical properties of the generalized Pareto distribution, and their uses, are described, and it is characterized by its 'threshold-stability'.

In Chapter 6 maximum likelihood estimation of the distribution is studied in detail, both for complex covariate-dependent data and for simple random samples, for which results on bias, censoring, and influence for maximum likelihood estimates are given. Two other methods of estimation are compared - fairly unfavourably - with maximum likelihood.

Chapter 7 gives some diagnostic tools for assessing fit of the distribution; these include residuals, a score test for fit, and ideas for looking at influence of observations on estimators.

5. THE GENERALIZED PARETO DISTRIBUTION

This chapter describes some basic properties of the generalized Pareto distribution, with comments about their statistical implications. Then a theorem characterising the distribution precisely as the only distribution which is 'threshold-stable' - in a sense later defined - is stated and proved.

For $k < 0$ the distribution

$$F(y) = \begin{cases} 1 - (1 - ky/\sigma)^{1/k} & (k \neq 0) \\ 1 - \exp(-y/\sigma) & (k = 0) \end{cases}$$

$(k \in \mathbb{R}, \sigma > 0, 1 - ky/\sigma > 0),$

was one of three proposed by Pareto(1897), and consequently is known as a Pareto type II distribution. Karl Pearson derived it as type VI of the family of distributions which bears his name, and later Macguire, Pearson and Wynn(1952) found that it was the compound of exponential variates with gamma-distributed random hazard. In a different context to that of extremes Davis and Feldstein(1979) call the Pareto type III a generalized Pareto distribution, but here the usage of Pickands(1975) - who seems to have coined the phrase 'generalized Pareto' - is followed, and the term is applied to the law $F(\cdot)$ above.

Its density is monotonic decreasing for $k < 1$ and increasing for $k > 1$: it is uniform when $k=1$, triangular when $k=1/2$, and exponential when $k=0$. The r^{th} moment of the distribution exists when $r+1/k$ is negative, and is then

$$E[X^r] = \sigma^r (-k)^{-r-1} \Gamma(1+r) \Gamma(-1/k-r) / \Gamma(1-1/k).$$

The first four central moments are

$$\begin{aligned}\mu_1 &= \sigma/(1+k), & \mu_3 &= 2\sigma^3(1-k)/[(1+k)^3(1+2k)(1+3k)], \\ \mu_2 &= \sigma^2/[(1+k)^2(1+2k)], & \mu_4 &= 3\sigma^4(2k^2-k+3)/[(1+k)^4(1+2k)(1+3k)(1+4k)],\end{aligned}$$

when they exist. The hazard function is $(\sigma-ky)^{-1}$ for y less than y_1 , the upper support point $\sup\{y : F(y) < 1\}$ of the density, which is the finite value σ/k for k positive.

The expected value of the r^{th} order statistic in a simple random sample of size n exists provided that $n+1-r > -k$, and is then

$$E[Y_{r,n}] = -k^{-1}\sigma \left[\prod_{i=1}^r (n+1-i)/(n+1-i+k) - 1 \right].$$

Provided the expectations exist, the difference between two successive order statistics is

$$E[Y_{r+i,n} - Y_{r,n}] = (n-r+k)^{-1} \prod_{i=1}^r (n+1-i)/(n+1-i+k).$$

The corresponding quantity for the exponential distribution is $(n-r)^{-1}$, motivating the idea of plotting ordered sample values against exponential order statistics. Such a graph should be concave for $k > 0$, a straight line for $k = 0$, and convex for $k < 0$, providing a means of assessing the weight of the sample upper tail. Estimation procedures based on sample order statistics can be developed which sometimes have high efficiency in other applications, and may be useful for this distribution when k is large and positive.

The conditional distribution of Y -s for a positive threshold $s < y_1$, given that $Y > s$, is generalized Pareto with parameters k and $\sigma-ks$. This property suggests a characterization of the distribution - to which I return below - but for now note that provided $k > -1$ is implies that

$$E[Y-s \mid Y > s] = (\sigma-ks)/(1+k),$$

and suggests the following graphical procedure for assessing the tail behaviour of simple random samples: for a succession of increasing levels s form the mean excess $\bar{y}(s)$ of the sample over s , and plot $\bar{y}(s)$ against s . As s increases the generalized Pareto distribution will in almost all cases approximate the tail of the data, where the graph should be a straight line with slope $-k/(1+k)$ and intercept $\sigma/(1+k)$. This has two practical uses: it provides rough estimates of k and σ ; and it suggests a minimum level at which the threshold should be drawn, namely above any non-linear lower portion of the graph.

Figure 5.1 shows this plot for the data in Figure 1.3, the set of exposures to Kr_{85} time-integrated air concentration at the receptor 800 km north of Mol in 1976, due to one Curie releases of the isotope from Mol every three hours. It shows a strong upward trend, indicating that k is negative and that the data have a Pareto-type tail. Graphical estimation gives $\tilde{\sigma} \sim 0.003$ and $\tilde{k} \sim 0.55$. In this instance the graph gives little information about where the threshold should be drawn, suggesting as it does that the generalized Pareto distribution should fit the entire dataset quite well.

Here is the characterisation result mentioned above, given as the first of only two theorems in this thesis:

Theorem 1:

If F is a non-degenerate d.f. with mass between the points $y_0 = \inf\{y : F(y) > 0\}$ and $y_1 = \sup\{y : F(y) < 1\}$, if $0 < y_0 < y_1 < \infty$, and if a random variable Y with d.f. F has the property that

$$\text{Prob}(Y > s + t\alpha(s) \mid Y > s) = \text{Prob}(Y > t) \quad \dots 5.1$$

for all $0 < s, t < y_1$ and some function $\alpha(s)$, then F is the generalized Pareto distribution. Conversely the generalized Pareto distribution

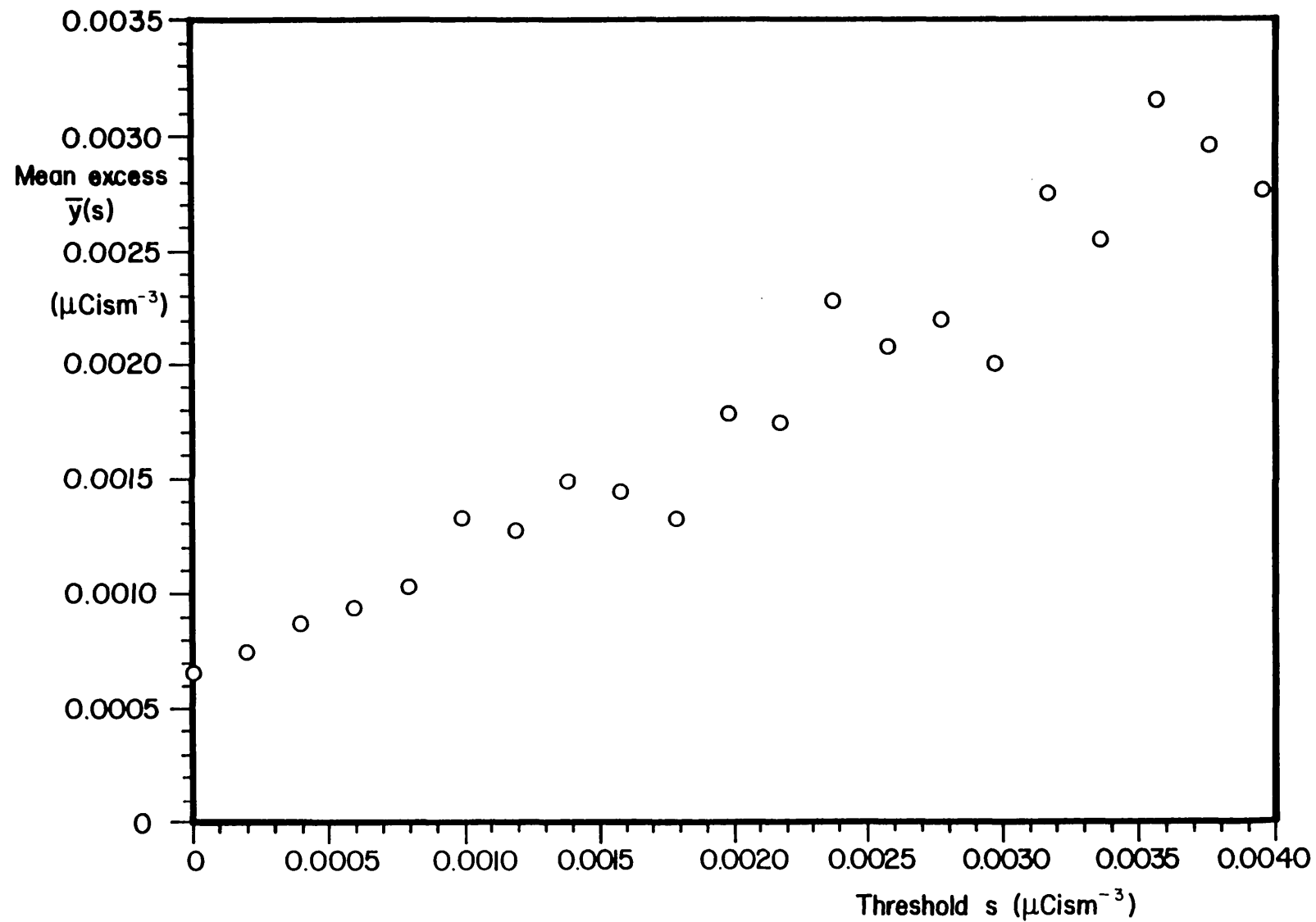


Figure 5.1 : Mean excesses over thresholds : Kr_{85} time-integrated air concentrations 800 km north of Mol due to unit releases over three-hourly periods through 1976.

satisfies 5.1.

Remark that if in the above expression we have $\alpha(s) \equiv 1$, we have the well-known 'loss of memory' characterisation of the exponential distribution. By analogy with the so-called 'max-stable' property which characterises the generalised extreme-value distribution, we call condition 5.1 'threshold stability', since it implies that the renormalised excess of Y over any level $s < y_1$ has precisely the distribution of Y .

The proof of the theorem is elementary but rather tedious. It proceeds in two halves: that for a d.f. satisfying 5.1 the only possible form of $\alpha(s)$ is $1+bs$ with b negative if y_1 is finite and b non-negative otherwise; and that for such an $\alpha(\cdot)$, $F(\cdot)$ must be generalized Pareto. The components of the proof are given as three lemmas. Throughout suppose that \bar{F} is the survivor function which corresponds to the distribution function F , that is, $\bar{F}(x) = 1-F(x)$.

Lemma 1:

If the d.f. of a random variable Y satisfies the conditions of the theorem above, then $y_0=0$ and $\alpha(s)=1+bs$, with $b < 0$ if y_1 is finite, and $b > 0$ otherwise.

- $\alpha(s)$ is positive if it is defined. For if $\alpha(s)=0$ for some s , then

$$1 = P[Y > s \mid Y > s] = P[Y > t] \quad (\forall t),$$

and F is degenerate, a contradiction. And if for some s , $\alpha(s) < 0$, then

$$1 = P[Y > s + t\alpha(s) \mid Y > s] = P[Y > t]$$

for all $y_1 > t > y_0$, and F is degenerate, again a contradiction.

When $s < y_1$, rewrite 5.1 as the functional equation

$$\bar{F}(t\alpha(s)+s) = \bar{F}(s) \bar{F}(t) \quad \dots 5.2.$$

with $0 < s, t < y_1$. Set $t=0$ to see that $\mathcal{F}(0)=1$, and set $s=0$ to see that $\alpha(0)=1$.

(i) Suppose that y_1 is finite, and choose any t such that $1 > \mathcal{F}(t) > 0$, letting $\alpha = \alpha(t)$ for short. Then

$$1 > \mathcal{F}(t)^n = \mathcal{F}(t(1+\alpha+\dots+\alpha^{n-1})) > 0$$

by induction on 5.2. Then $\alpha = \alpha(t) < 1$, otherwise $t(1+\alpha+\alpha^2+\dots+\alpha^n) > y_1$ for some n , leading to a contradiction. But as n increases, so $\mathcal{F}(t)^n \rightarrow 0$, and

$$t \left[\frac{1-\alpha^n}{1-\alpha} \right] \rightarrow y_1,$$

implying that $y_1 = t/(1-\alpha(t))$, whatever the value of $t \in [y_0, y_1]$. That is, $\alpha(t) = 1 - t/y_1$ for such t .

I now prove that $y_0 = 0$ by supposing otherwise and by deducing that $\mathcal{F}(\cdot)$ is a step-function obtaining yet another contradiction.

If $y_0 > 0$ then $\mathcal{F}(y_0) < 1$, so $\alpha(y_0) = 1 - y_0/y_1$. Choose s and t so that $y_0 < s < t < y_1$, so that

$$0 < (t-s)/\alpha(s) = y_1 \left[\frac{(t-s)/(y_1-s)}{1-\alpha(s)} \right] < y_1,$$

and thus

$$\mathcal{F}(t) = \mathcal{F}(s) \mathcal{F}\left(\frac{t-s}{\alpha(s)}\right) \quad \dots \quad 5.3,$$

from 5.2. Setting $s=y_0$, it follows that $\mathcal{F}(t) = \mathcal{F}(y_0)$ whenever $y_0 < t < y_0 + y_0(y_1 - y_0)/y_1$, that is, $\mathcal{F}(\cdot)$ is constant in the interval $[y_0, y_0 + y_0(y_1 - y_0)/y_1]$. It follows easily by induction that \mathcal{F} has constant value $\mathcal{F}(y_0)^k$ on each interval

$$\left[y_0 \sum_{i=0}^{k-1} (1-y_0/y_1)^i, y_0 \sum_{i=0}^k (1-y_0/y_1)^{i+1} \right).$$

Now choose s and t on opposite sides of $y_0(2-y_0/y_1)$, but close enough so that $(t-s)/\alpha(s) < y_0$. Then

$$\mathcal{F}(y_0)^2 = \mathcal{F}(t) = \mathcal{F}(s) \mathcal{F}\left(\frac{t-s}{\alpha(s)}\right) = \mathcal{F}(y_0),$$

a contradiction. Thus $y_0=0$, and $\alpha(s)=1-s/y_1$, as required.

(ii) Now suppose that $y_1 = \infty$. The sequence argument above at the start of (i) shows that $\alpha(t) > 1$ for $t > 0$ and that $\alpha(0)=1$. Another step-function argument shows that $y_0=0$. It only remains to show that $\alpha(s)=1+bs$ for some positive b .

The function \mathcal{F} is strictly monotonic. For if $\mathcal{F}(s) = \mathcal{F}(t)$, and $0 < s < t$, then

$$\mathcal{F}(s) = \mathcal{F}\left(\frac{t-s}{\alpha(s)}\right) = \mathcal{F}(t)$$

and thus $\mathcal{F}\left(\frac{t-s}{\alpha(s)}\right) = 1$. But $y_0=0$, so $s=t$. Therefore \mathcal{F}^{-1} exists,

and since

$$\mathcal{F}(s)\mathcal{F}(t) = \mathcal{F}(t)\mathcal{F}(s)$$

it follows that

$$\mathcal{F}(s+t\alpha(s)) = \mathcal{F}(t+s\alpha(t))$$

and so $s+t\alpha(s) = t+s\alpha(t)$ for all $s, t > 0$. Thus $(\alpha(s)-1)/s = (\alpha(t)-1)/t = b$, for some non-negative constant b , which establishes the lemma. •

Before proceeding to prove the theorem, I now make a definition, and state and prove two lemmas, one short, and one slightly longer.

Define the support $S(b)$ for $b \in \mathbb{R}$ thus:

$$S(b) = \begin{cases} [0, \infty) & (b > 0) \\ [0, -1/b) & (b < 0). \end{cases}$$

Henceforth assume that when b and \mathcal{F} appear together, the particular b chosen is such that 5.1 is satisfied with $\alpha(s)=1+bs$.

Lemma 2:

1. If $y \in S(b)$, then $\mathcal{F}(y)^n = \mathcal{F}\left(\frac{1+by)^{n-1}}{b}\right)$.
2. If $y, a \in S(b)$, and $a < y$, then $(y-a)/((1+ba)y) < 1$.

- 1. Proved using 5.3 by induction on n.
- 2. Since $y \in S(b)$, $1+by > 0$, so $y+bya > y-a$, so $1 > (y-a)/((1+ba)y)$, provided that $y > a$. •

Lemma 3:

If $b \in \mathcal{R}$ and $F(\cdot)$ is a d.f. with mass in $S(b)$ whose survivor function \bar{F} satisfies the equation

$$\bar{F}(s)\bar{F}(t) = \bar{F}(t+s+tsb)$$

whenever $s, t \in S(b)$, then:

- (i) if $b=0$, F is the exponential distribution function;
- (ii) if $b \neq 0$, $F(y) = 1 - (1+by)^{-1/c}$ for some real c with the same sign as b .

- This proof is in three stages.

(i) $b=0$. The result is well-known, and is given in Feller (1968, p359), under the weaker condition that \bar{F} be bounded in some interval.

The proof of part (ii) is a clumsy adaptation of Feller's proof.

(ii) $b > 0$. Now $\bar{F}(0)=1$, and $\bar{F}(x) > 0$ for all positive x . Choose the positive c for which $\bar{F}(1)(1+b)^{1/c}=1$, and define $v(x) = \bar{F}(x)(1+bx)^{1/c}$ for all positive x . $v(x)$ is bounded in bounded sets, in particular in the unit interval, and satisfies equations 5.2 and 5.3.

I aim to contradict the boundedness of $v(\cdot)$ if $v(\cdot)$ is not identically one in $S(b)$ by showing that otherwise a sequence $\{u_n\}$ in the unit interval exists with the property that $v(u_n) \rightarrow \infty$ as $n \rightarrow \infty$.

To begin, suppose that $u \in S(b)$ has $v(u)=q \neq 1$. If $u > 1$, note that

$$v(u) = v(1)v((u-1)/(1+b)) = v((u-1)/(1+b)) = q,$$

and repeatedly replace u with $(u-1)/(1+b)$ until $u < 1$. If $q < 1$, note that

$$1 = v(1-u + u) = v(u)v((1-u)/(1+bu)),$$

and therefore that by part 2 of Lemma 2, $u' = (1-u)/(1+bu) < 1$ and $v(u') > 1$. This establishes that if $v(\cdot)$ is not identically unity in $S(b)$, it exceeds it at some point of the unit interval.

Now

$$v(u)^n = v\left[\frac{(1+bu)^n - 1}{b}\right],$$

by part 1 of Lemma 2, and so the sequence u_n in $[0,1]$ defined as $\frac{(1+bu)^n - 1}{b}$ if this is less than one, and as any $x \in (0,1)$ for which $v(x) = v(u)^n$ otherwise, has $v(u_n)$ unbounded. Such a sequence can always be found because $v(u) = v\left(\frac{u-1}{1+b}\right)$ if $u > 1$.

But $v(\cdot)$ is bounded in bounded sets, a contradiction, and so if $b > 0$, $v(x)$ is identically one in $S(b)$.

$b < 0$. Choose $z \in S(b)$ and $c < 0$ such that $\mathcal{F}(z) = (1+bz)^{-1/c}$, and define $v(x)$ as above. $v(z) = 1$, and again $v(\cdot)$ satisfies equations 5.2 and 5.3 and is bounded on compact subsets of $S(b)$. The proof now follows much as before, except that z takes the place of the point 1.

This completes the proof of the lemma. ●

Theorem 1 may now be proved as follows.

● If for some function $\alpha(s)$, \mathcal{F} satisfies the conditions of Theorem 1, then by Lemma 1, $\alpha(s) = 1+bs$ for some b , and the distribution function $F = 1 - \mathcal{F}$ puts all its probability mass in $S(b)$. Apply Lemma 3 to establish the first half of the result. The converse is verified by direct substitution. ●

The objective of this part of the thesis is to develop and study methods for the statistical analysis of extremes, rather than to pursue the probabilistic ramifications of this theorem. But note that it provides one of the main building-blocks for a proof of the conjecture that if a non-degenerate d.f. $G(\cdot)$ and some function $\beta(\cdot)$

have the property that the limit

$$\lim_{s \rightarrow x_1} \frac{1-G(s+t\beta(s))}{1-G(s)} = \mathcal{F}(t),$$

where $x_1 = \sup\{x : G(x) < 1\}$, exists at all continuity points of the survivor function $\mathcal{F}(\cdot)$, then \mathcal{F} is the survivor function of the generalized Pareto distribution. In fact the theorem proved above is directly analogous to Theorem 1.4.1 of Leadbetter, Lindgren and Rootzen(1983)'s version of the 'extremal types theorem': the well-known result that the generalized extreme-value law is the only possible limiting form for the distribution of maxima sets of i.i.d. variates.

Part of the second half of the conjecture is contained in Theorem 2, which shows that a non-degenerate \mathcal{F} arising as above must of necessity be threshold-stable at its points of continuity.

Theorem 2:

Suppose that G is a distribution function with upper support point $x_1 = \sup\{x : G(x) < 1\}$, and that a function $\beta(s)$ exists such that the limit

$$\lim_{s \rightarrow x_1} \frac{1-G\{s+t\beta(s)\}}{1-G\{s\}} = \mathcal{F}(t)$$

exists at all continuity points of \mathcal{F} , which is the survivor function of a non-degenerate distribution. Then \mathcal{F} must be threshold-stable where it is continuous.

- Say that $y_1 = \sup\{x : \mathcal{F}(x) > 0\}$. First suppose that $0 < t, u < y_1$, and that t and u are continuity points of \mathcal{F} , and consider:

$$\begin{aligned} \mathcal{F}(t)\mathcal{F}(u) &= \lim_{s \rightarrow x_1} \frac{H\{s+t\beta(s)\}}{H\{s\}} \frac{H\{s+\beta(s)\{t+u\alpha_t(s)\}\}}{H\{s+t\beta(s)\}} \\ &= \lim_{s \rightarrow x_1} \frac{H\{s+\beta(s)\{t+u\alpha_t(s)\}\}}{H\{s\}}, \end{aligned}$$

where $\alpha_t(s) = \beta(s+t\beta(s))/\beta(s)$ and $H(s)=1-G(s)$.

The function $\alpha_t(s)$ must be bounded as s approaches x_1 . For if $y_1 < \infty$, then $t+u\alpha_t(s) < y_1$, so $\alpha_t(s) < 1-t/y_1$. And if $y_1 = \infty$, then if $\alpha_t(s)$ is not bounded, a sequence s_n exists such that $\alpha_t(s_n) > n$ for each natural number n . In which case $\mathcal{F}(t)\mathcal{F}(u) < \mathcal{F}(t+un)$ for each n , which is impossible.

Therefore a sequence s_n exists such that $\alpha_t(s_n)$ converges to a limit a_t , say, for each fixed t ; without loss of generality either $\alpha_t(s_n)$ increases or it decreases.

Suppose first that $\alpha_t(s_n) \uparrow a_t$. Then

$$\frac{H[s_n + \beta(s_n)\{t+ua_t\}]}{H[s_n]} > \frac{H[s_n + \beta(s_n)\{t+u\alpha_t(s_n)\}]}{H[s_n]} > \frac{H[s_n + \beta(s_n)\{t+u\alpha_t(s_m)\}]}{H[s_n]}$$

for any fixed $m < n$. Letting n tend to infinity,

$$\mathcal{F}(t+ua_t) > \mathcal{F}(t)\mathcal{F}(u) > \mathcal{F}(t+u\alpha_t(s_m)).$$

And letting m tend to infinity, since \mathcal{F} is right-continuous it must be threshold-stable.

Now suppose that $\alpha_t(s_n) \downarrow a_t$, and by a similar argument see that

$$\mathcal{F}(t+u\alpha_t(s_m)) > \mathcal{F}(t)\mathcal{F}(u) > \mathcal{F}(t+ua_t).$$

If $\mathcal{F}(t)\mathcal{F}(u) = \mathcal{F}(t+ua_t)$, the result is proven, so assume otherwise, i.e. $\mathcal{F}(t)\mathcal{F}(u) > \mathcal{F}(t+ua_t)$. \mathcal{F} is continuous at u , so a $\delta > 0$ exists such that

$$\mathcal{F}[t+(u+\delta)\alpha_t(s_m)] > \mathcal{F}(t)\mathcal{F}(u) > \mathcal{F}(t+ua_t).$$

But for large enough m , $(u+\delta)\alpha_t(s_m) > ua_t$, since $\alpha_t(s_m) \downarrow a_t$, and so

$$\mathcal{F}(t+ua_t) > \lim_{m \rightarrow \infty} \mathcal{F}[t+(u+\delta)\alpha_t(s_m)],$$

a contradiction.

Thus \mathcal{F} is threshold-stable at its points of continuity,

establishing the theorem. •

Theorems 1 and 2 go some way towards a direct proof of Pickands' (1975) result. What remains is some tidying-up, and some effort to find equivalent conditions for convergence of distributions of maxima and those of excesses. I turn instead to statistical inference for the generalized Pareto distribution.

6. ESTIMATION OF THE DISTRIBUTION

A major aspect of statistics is the description of systematic variation in response variables in terms of the behaviour of explanatory variables. Here there is the need to assess the extent to which exceedances for different but related series depend on external factors. There is a further point to make specifically in the context of extremes. The generalized Pareto distribution is an asymptotic approximation to the distribution of exceedances, an approximation which should improve as the level over which it is made increases. Use of covariates to synthesise tail information from several series allows higher thresholds to be imposed in each, so the overall approximation may be closer.

McCullagh and Nelder(1983) give an account of the analysis of complicated data through the use of the generalized linear model, which has proven to be a very successful technique. It provides a temptingly powerful tool for describing, unifying and assessing the dependence of random variables on known covariates. The temptation is not resisted much longer.

A great strength of the idea is the use it makes of maximum likelihood estimation, but in the present context this is not always possible: such estimation of the parameters k and σ of a generalized Pareto random variable is non-regular in the sense that the score statistic is not asymptotically Normal if $k > \frac{1}{2}$ (Smith, 1985); and if $k > 1$ it has infinite mean and so the usual Taylor expansions cannot be made. Smith studies maximum likelihood estimates in a number of non-regular cases, including the generalized Pareto distribution with $k > \frac{1}{2}$, and derives results about their rates of convergence. Experience with fitting the distribution to data indicates that such non-regularity is not a common problem, a view which seems to be shared by others working with the extreme-value distributions.

Typically the value of k is fairly close to zero. If the problem of non-regularity does arise it may be overcome by using Bayesian techniques, although the specification of a meaningful joint prior distribution for the parameters and programming and computation of the appropriate numerical integrations seem likely to be hard in all but the simplest cases.

Smith(1985) suggests in that in single-sample problems the upper endpoint of the distribution may be estimated by top sample order statistic, but it is not clear how to extend this to problems where the data involve dependence on covariates.

Cheng and Amin(1983) propose a class of estimators which allow endpoints to be estimated in such cases, but their results do not extend immediately to censored data.

Throughout this chapter the Y_i are regarded as generalized Pareto variates which are independent, conditionally on the values of their associated covariates. In view of the comments in Chapter 4 this assumption will only rarely be exactly right for excesses from the same cluster, but it seems an appropriate starting-point for what follows.

Section 6.1 gives some fairly detailed results about maximum likelihood estimation and hypothesis testing for a general parametrization of the distribution. Section 6.2 turns to more detailed results on bias, censoring, and influence for maximum likelihood estimates in the special case of a simple random sample. Then in Section 6.3 moments and least-squares estimation for the distribution are considered and their efficiency assessed. Finally Section 6.4 contains some results on discrimination between different types of tail behaviour for simple random samples.

6.1 Maximum likelihood estimation: generalities

Suppose that the Y_i ($i=1\dots n$) have generalized Pareto distributions with parameters k_i and σ_i , that conditionally on the known value of $p+1$ - and $q+1$ -vectors of covariates x_i and z_i they are stochastically independent, and that $k_i < 1/2$ for all i so that standard maximum likelihood theory applies. Suppose in addition that $k_i = k(\gamma^t z_i)$ and that $\sigma_i = \sigma(\beta^t x_i)$ in terms of parameters $\theta = (\gamma, \beta)$. The natural functions $k(\cdot)$ and $\sigma(\cdot)$ here are the identity and exponential functions respectively, in which case $k_i = \gamma^t z_i$ and $\sigma_i = \exp(\beta^t x_i)$, but other possibilities may be suggested by the nature of or examination of the data. Finally suppose that the Y_i may be upper-truncated by some mechanism independent of them, perhaps representing random censoring in a medical context, or loss of information due to a measuring instrument going out of its range of calibration or being blown down or washed away. Say that $\delta_i = 0$ if the observed value y_i of Y_i is the true one, and otherwise that $\delta_i = 1$ in which case the actual value of Y_i is known to be equal to or greater than y_i . Then the contribution to the loglikelihood made by the i^{th} observation is

$$\ell_i = \begin{cases} (1/k_i - 1) \log\{1 - k_i y_i / \sigma_i\} - \log \sigma_i + \delta_i [\log \sigma_i + \log\{1 - k_i y_i / \sigma_i\}] & (k_i \neq 0) \\ -y_i / \sigma_i - \log \sigma_i + \delta_i \log \sigma_i & (k_i = 0) \end{cases}$$

$$(\sigma_i > 0, y_i > 0, 1 - k_i y_i / \sigma_i > 0),$$

and the $p+q+2$ vector $\theta(\gamma, \beta)$ of unknown parameters is to be estimated from the data. Often, the k_i will have common value k for all observations and there is no censoring; the resulting formulae for information matrices and so on are then much simpler. Here are the elements of the score vector, the second derivatives of, and the Fisher information matrix i_θ for the loglikelihood $\ell = \sum_i \ell_i$ of the

whole data: the score vector

$$\partial \ell / \partial \gamma_s = \sum_i z_{is} k_i' \partial \ell_i(y_i; \sigma_i, k_i) / \partial k_i,$$

$$\partial \ell / \partial \beta_u = \sum_i x_{iu} \sigma_i' \partial \ell_i(y_i; \sigma_i, k_i) / \partial \sigma_i,$$

where

$$\partial \ell_i(y; \sigma, k) / \partial k_i = \begin{cases} -k^{-2} \log(1-ky/\sigma) + (1-1/k)y/(\sigma-ky) - \delta y/(\sigma-ky) & (k \neq 0) \\ y/\sigma - \frac{1}{2}(y/\sigma)^2 - \delta y/\sigma & (k=0) \end{cases}$$

and

$$\partial \ell_i(y; \sigma, k) / \partial \sigma_i = \begin{cases} (1-k)y/\{\sigma(\sigma-ky)\} - 1/\sigma + \delta[1/\sigma + ky/\{\sigma(\sigma-ky)\}] & (k \neq 0) \\ [y/\sigma - 1 + \delta] / \sigma & (k=0); \end{cases}$$

the matrix of second derivatives

$$\partial^2 \ell / \partial \gamma_s \partial \gamma_t = \sum_i z_{is} z_{it} \left[k_i'^2 \partial^2 \ell_i(y_i; \sigma_i, k_i) / \partial k_i^2 + k_i' \partial \ell(y_i; \sigma_i, k_i) / \partial k_i \right],$$

$$\partial^2 \ell / \partial \gamma_s \partial \beta_u = \sum_i z_{is} x_{iu} k_i' \sigma_i' \partial^2 \ell_i(y_i; \sigma_i, k_i) / \partial k_i \partial \sigma_i,$$

$$\partial^2 \ell / \partial \beta_u \partial \beta_v = \sum_i x_{iu} x_{iv} \left[\sigma_i'^2 \partial^2 \ell_i(y_i; \sigma_i, k_i) / \partial \sigma_i^2 + \sigma_i' \partial \ell(y_i; \sigma_i, k_i) / \partial \sigma_i \right],$$

where

$$\partial^2 \ell_i(y; \sigma, k) / \partial k_i^2 = \begin{cases} 2k^{-3} \log(1-ky/\sigma) + 2y/\{k^2(\sigma-ky)\} + (1-1/k)y^2/(\sigma-ky)^2 - \delta y^2/(\sigma-ky)^2 & (k \neq 0) \\ (y/\sigma)^2 [1 - 2y/3\sigma] - \delta (y/\sigma)^2 & (k=0), \end{cases}$$

$$\partial^2 \ell_i(y; \sigma, k) / \partial k_i \partial \sigma_i = \begin{cases} -y/\{\sigma(\sigma-ky)\} + (1-k)y^2/\sigma(\sigma-ky)^2 + \delta [y/\{\sigma(\sigma-ky)\} + ky^2/\{\sigma(\sigma-ky)^2\}] & (k \neq 0) \\ y(y/\sigma - 1)/\sigma^2 + \delta y/\sigma^2 & (k=0), \end{cases}$$

and

$$\partial^2 \ell_i(y; \sigma, k) / \partial \sigma_i^2 = \begin{cases} (k-1)y [1/\sigma + 1/(\sigma-ky)] / \{ \sigma(\sigma-ky) \} + 1/\sigma^2 & (k \neq 0) \\ -\delta [1/\sigma^2 + ky \{ 1/\sigma + 1/(\sigma-ky) \} / \{ \sigma(\sigma-ky) \}] & \\ (1-2y/\sigma) / \sigma^2 - \delta / \sigma^2 & (k=0); \end{cases}$$

the elements of the Fisher information matrix are

$$-E [\partial^2 \ell / \partial \gamma_s \partial \gamma_t] = \sum_i z_{is} z_{it} k_i'^2 j_{kk} (y_i; \sigma_i, k_i),$$

$$-E [\partial^2 \ell / \partial \gamma_s \partial \beta_u] = \sum_i z_{is} x_{iu} k_i' \sigma_i' j_{k\sigma} (y_i; \sigma_i, k_i),$$

$$-E [\partial^2 \ell / \partial \beta_u \partial \beta_v] = \sum_i x_{iu} x_{iv} \sigma_i'^2 j_{\sigma\sigma} (y_i; \sigma_i, k_i),$$

where the j 's are given by the expressions

$$j_{kk} (y; \sigma, k) = [2 - \delta (2 + 2w + w^2 - 4kw - 3kw^2 + 2k^2 w^2) (1 - ky/\sigma)^{1/k-2}] / \{ (1-k)(1-2k) \},$$

$$j_{k\sigma} (y; \sigma, k) = - [1 - \delta (1 + w + 2kw) (1 - ky/\sigma)^{1/k-2}] / \{ \sigma(1-k)(1-2k) \},$$

$$j_{\sigma\sigma} (y; \sigma, k) = [1 - \delta (1 - ky/\sigma)^{1/k-2}] / \{ (1-2k)\sigma^2 \}.$$

Throughout the equations, $0 \leq s, t \leq q$, $0 \leq u, v \leq p$, and $z_{i0} = x_{i0} = 1$; while σ_i', k_i' and so on are the derivatives of σ_i and k_i with respect to their arguments; and $w = y/\sigma$.

In the case when $k_i = k$ for all i , and there is no censoring, without otherwise losing generality set $\sum_i x_{iu} = 0$ for $1 \leq u \leq p$ so that the overall mean is orthogonal to all the other regressors. The Fisher information matrix is then

$$(1-2k)^{-1} \begin{vmatrix} X^t X & 0 & 0 \\ 0 & 2n/(1-k) & -n/(1-k) \\ 0 & -n/(1-k) & n \end{vmatrix},$$

of side $p+2$, and this is easily inverted if $X^t X$ is full rank. This

form of the matrix i_{θ} has obvious implications for the robust and optimal design of experiments for sample extremes; however such data only rarely arise from designed trials.

Details of computational procedures for the estimates are not studied here. Nevertheless note that if each $k_i \leq 0$, methods for unconstrained optimization of the likelihood may be used, but that if some or all of the k_i are positive, maximisation is subject to the constraints $1 > k_i y_i / \sigma_i$, and a suitable algorithm may be needed.

A consequence of the non-regular behaviour of the estimators for $k_i > 1/2$ is that when $0 < k_i < 1/2$ and especially when $1/3 < k_i$, confidence intervals for parameters - and computational procedures for their estimates - which depend on quadratic approximations to the loglikelihood may be poor, so that it may be desirable to find confidence regions based on the likelihood ratio statistic rather than the asymptotic Normal distribution of the estimates. Plotting the loglikelihood is a useful guide as to the need for this; the matter is studied in more detail in the next section.

The usual theory of hypothesis testing for the effect of particular covariates, based on difference of maximized loglikelihoods, may be used, but care should be taken if any k_i are positive. One hypothesis of interest when the data are not censored and are divided into homogeneous subsamples is that the shape parameter k_i is constant throughout the data or some collection of subsamples, the alternative being that it differs in each. More generally it may be required to test the hypothesis of constant k against the alternative that some explanatory factors influence the k_i .

Although the difference of the maximized loglikelihoods under the models can be found, it will usually be more economical to construct a score test for the null against the alternative

hypothesis.

Such a test is based on partitioning the vector θ into parameters ζ of interest and nuisance parameters ξ , and testing the possibility that $\zeta = \zeta_0$, some particular null value. If $\theta_0 = (\hat{\xi}, \zeta_0)$ is the maximum likelihood estimate of θ on the hyperplane $\zeta = \zeta_0$, then the statistic $W_u = U(\theta_0) i(\theta_0)^{-1} U(\theta_0)$ is asymptotically $\chi^2_{\dim \zeta}$ when the true value of ζ is ζ_0 . Here $U(\theta_0)$ and $i(\theta_0)$ are respectively the score statistic and Fisher information evaluated at θ_0 . The test has power properties similar to those of the likelihood ratio test but the likelihood need only be maximized on the null hypothesis - an advantage in potentially non-regular situations. See Cox and Hinkley (1974, Chapter 9) for more details.

For testing $\gamma = 0$, $i(\theta_0)$ is the $p+q+2$ matrix

$$(1-2k)^{-1} \begin{vmatrix} 2(1-k)^{-1} Z^t Z & -(1-k)^{-1} Z^t X & 0 \\ -(1-k)^{-1} X^t Z & X^t X & 0 \\ 0 & 0 & M \end{vmatrix},$$

where M is $(1-2k)$ times the Fisher information matrix for (γ_0, β_0) .

The upper left $q \times q$ element of $i^{-1}(\theta_0)$ is

$$Q = (1-k)(Z^t Z)^{-1} Z^t \left[I + \frac{1}{2}(1-k)^{-1} X E^{-1} X^t \right] Z (Z^t Z)^{-1},$$

where E is the matrix

$$X^t \left\{ I - \frac{1}{2}(1-k)^{-1} Z (Z^t Z)^{-1} Z^t \right\} X.$$

The s^{th} ($1 \leq s \leq q$) element of $U(\theta_0)$ is

$$k^{-2} \sum_{i=1}^q z_{is} \left\{ k(k-1) y_i (\sigma_i - k y_i)^{-1} - \log(1 - k y_i / \sigma_i) \right\},$$

giving $W_u = U_q^t(\theta_0) Q U_q(\theta_0)$, where U_q is the vector of the q uppermost elements of the score U , to be compared with χ^2_q .

6.2 Maximum likelihood estimation: simple random samples

6.2.1 Existence of estimators

Smith(1983) points out that when the n variates Y_i are independent and identically distributed according to the generalized Pareto law with parameters σ and k it is possible to reduce the search for maximum likelihood estimates to a one-dimensional problem by means of a reparametrization. Letting $\tau = \sigma/k$, the sample loglikelihood is

$$\begin{aligned} \ell(\tau, k) &= n \log(\tau/k) + (1/k-1) \sum_1 \log(1-\tau y_i) \\ & \quad (k \in \mathbb{R}, \tau < 1/\max\{y_i\}), \end{aligned}$$

and its differential with respect to k is

$$\partial \ell(\tau, k) / \partial k = -n/k - 1/k^2 \sum_1 \log(1-\tau y_i),$$

so that for a given value of τ , $\partial \ell / \partial k = 0$ implies that

$$\hat{k}(\tau) = n^{-1} \sum_1 \log(1-\tau y_i).$$

The second derivative $\partial^2 \ell(\tau, k(\tau)) / \partial k^2$ is almost surely defined and negative, so the maximum likelihood estimates of k and τ - if they exist - may be found by maximizing $\ell_0(\tau) = \ell(\tau, \hat{k}(\tau))$ as a function of τ , either graphically or numerically.

The function ℓ_0 has a singularity as $\tau \rightarrow -1/\max\{Y_i\}$, and if one exists a local maximum of ℓ_0 should be chosen to provide maximum likelihood estimates. A local minimum must exist if a local maximum does, because of the singularity, so some care is needed to find the correct root of the equation $\ell_0'(\tau) = 0$. Table 6.1 shows the results of a small simulation study to investigate the existence of maximum likelihood estimates in single samples. Changes of sign of ℓ_0' were sought in the interval $(-1/\max\{Y_i\}, 20)$, for samples generated for

$\sigma=1$ and several values of k . The number of cases out 100, for several sample sizes, in which they could be found by evaluating the function at a mesh of points of separation ≈ 0.01 is recorded in the table. Clearly roots of the equation exist with high probability except for very small sample sizes and positive k . These results are in broad agreement with those of Smith and Weissman(1985) for the closely related case of estimating the parameters of the Weibull distribution in non-regular and almost non-regular cases.

sample size	value of k					
	-0.6	-0.4	-0.2	0.0	0.2	0.4
10	88	88	74	75	54	37
25	100	100	100	96	97	85
40	100	100	100	100	99	98
65	100	100	100	100	100	99

Table 6.1: Numbers out of 100 simulated simple random samples in which likelihood equation roots could be found.

6.2.2 Their small-sample properties

Although such estimators are asymptotically unbiased, they are usually biased to some extent in finite samples. In some cases expressions for the bias to order n^{-1} or higher may be obtained as follows (Bartlett, 1952; Haldane, 1953; Shenton and Bowman, 1977): make a second-order Taylor expansion of the likelihood equation $\partial \ell(Y; \hat{\theta}) / \partial \theta = 0$ about the true parameter-value θ , and solve the resulting equations to find the bias $E[\hat{\theta} - \theta]$ to order n^{-1} . For the case of a simple random sample of generalized Pareto variables with parameters σ and k , it emerges after appreciable calculation that the biases are

$$nE[\hat{k} - k] = 2(1-k)^3 / (1-2k)(1-3k) + O(n^{-1}),$$

$$n\sigma^{-1}E[\hat{\sigma}-\sigma] = -2k(4k^2-k-1)/(1-2k)(1-3k)+O(n^{-1}),$$

when standardized to be $O(1)$ and independent of the scale σ , provided that $k < 1/4$ so that the appropriate expansions can be made and expectations taken. Table 6.2 shows these expressions calculated for a few values of k , together with some simulated standardized biases.

The simulated results are based on 1000 samples each of sizes 25, 50 and 100 for each of the values of k indicated in the table. The one-dimensional line-search discussed above was used to find the estimates, if possible, until 1000 samples in which they could be found had been generated. The table also gives the total numbers of samples of size 25, 50 and 100 generated.

Comparison of the simulated and theoretical biases reveals that the $O(n^{-1})$ theoretical biases underestimate the actual bias of the estimators, except that of \hat{k} when $k=0.2$; furthermore none of the theoretical values lies inside the approximate 95% confidence intervals based on the simulated biases. In fact for samples of size about 25 the bias in \hat{k} will be about 0.1 over most of the range considered, and amount likely to be practically unimportant compared with the sampling variation of the estimates, but is it worth knowing that the value of k tends to be slightly overestimated and hence the tail weight slightly underestimated in samples of small and moderate size. The value of σ is slightly overestimated, and by a similar amount. In some sense this might be thought of as compensating for overestimating k .

The discrepancy between the theoretical and simulated values may be caused by the $O(n^{-1})$ terms in the expression for standardized bias, although were this the case it would seem odd that the differences between the simulated biases for $n=25$ and $n=100$ are not greater. Further terms in the expansion for bias could be found, but as the bias is not large enough to be dangerous except in very small

		shape parameter k					
		-0.6	-0.4	-0.2	0.0	0.2	0.4
standardized bias of \hat{k}	Theory	1.330	1.386	1.543	2.00	4.267	—
	n=25	2.438±0.539	2.504±0.525	2.656±0.437	3.013±0.394	2.770±0.365	2.218±0.305
	n=50	2.292±0.723	2.410±0.669	3.115±0.599	2.708±0.528	3.256±0.470	3.00 ±0.397
	n=100	2.330±1.013	1.956±0.939	2.954±0.780	2.607±0.673	3.680±0.631	3.818±0.560
standardized bias of $\hat{\sigma}$	Theory	0.303	0.008	0.114	0.0	1.733	—
	n=25	3.302±0.660	3.792±0.679	2.810±0.577	3.238±0.540	2.763±0.496	1.975±0.420
	n=50	2.773±0.839	3.231±0.815	3.344±0.808	2.746±0.705	3.321±0.656	2.76 ±0.557
	n=100	2.606±1.130	2.692±1.119	3.871±1.018	2.983±0.927	3.650±0.889	3.423±0.821

Table 6.2 Standardized bias of maximum likelihood estimates for generalized Pareto distribution
bias \pm 1.96 x s.e. (bias)

samples the result seems unlikely to justify the effort involved.

I now address the question of the speed of approach to Normality of the maximum likelihood estimators of σ and k , again by simulation. Mardia, Kent and Bibby(1979, pages 21 and 148) define multivariate measures of skewness and kurtosis, b_1 and b_2 respectively, for use in comparing populations of multivariate data with the multinormal distribution. The asymptotic distributions of b_1 and b_2 are known when the underlying population is multivariate Normal. The observed values of b_1 and b_2 for the joint distribution of $(\hat{\sigma}, \hat{k})$ in 1000 simulated samples of sizes 25, 50 and 100 are given in Table 6.3 for several underlying values of k , together with their asymptotic 95% confidence intervals based on the assumption that the estimators are bivariate Normal. The skewness is much too high even for $n=100$, as is the kurtosis for $n \leq 50$, but for $n=100$ the kurtosis is high but lies inside the interval for each value of k . Thus the estimates are not close to their asymptotic distribution even for $n=100$, which might in some contexts be regarded as a fairly large sample size.

These results tie in with those of Johnson and Haskell(1983), who studied properties of maximum likelihood estimators of the three-parameter Weibull distribution by simulation. For samples of size 70 they found that the estimates were not Normally distributed, being both biased and skewed.

This non-Normality raises the issue of the small-sample distortion of asymptotic confidence regions for the true parameter values, based on the sample loglikelihood. The usual three statistics upon which such confidence regions are based are Wilks' statistic, equal to twice the log-likelihood ratio evaluated at the estimates and null point respectively, and regions based on the asymptotic Normal distribution of the estimates using either the inverse expected or inverse observed information matrix, evaluated at

		shape parameter k					
		-0.6	-0.4	-0.2	0.0	0.2	0.4
b_1	n= 25	1.427	2.369	0.784	0.771	0.817	1.321
	n= 50	0.554	0.610	0.872	0.343	0.513	0.633
	n=100	0.203	0.305	0.100	0.232	0.271	0.364
b_2	n= 25	10.36	12.89	9.548	9.083	9.090	9.393
	n= 50	8.758	8.585	9.788	8.737	9.531	8.977
	n=100	8.382	8.479	8.341	8.383	8.252	8.382

Table 6.3 Observed multivariate skewness and kurtosis of mle $(\hat{\sigma}, \hat{k})$ for 1000 simulated samples.

Asymptotic 95% confidence intervals: $b_1 \in (0.003, 0.067)$

$b_2 \in (7.504, 8.496)$

(see Mardia, Kent, and Bibby (1979), pages 21, 148)

the estimates, as the covariance matrix of the estimators. All three sets of regions are based on the asymptotic χ^2_2 distribution of the test statistics. One difficulty with situations such as this in which the Fisher information matrix is only positive definite for a limited range of some parameter - in this case k - is that the matrix may not be positive definite at the estimate. In the example to hand, if $\hat{k} > 1/2$, then although the observed information is positive definite, the expected information is not. The only sensible policy in this circumstance is to use Wilks' statistic or the observed likelihood for confidence regions.

Table 6.4 shows the results of a small simulation study to compare the behaviour of the statistics relative to their asymptotic significance points in samples of sizes 25, 50, and 100. Wilks' statistic is close to its asymptotic points even for $n=25$. The statistics based on observed and expected information are roughly comparable but are overdispersed relative to the limiting distribution and are not close to it even for $n=100$. This point is brought out more strongly in Table 6.5, which shows some of the observed significance points of the statistics in the same simulations. The nominal .95 and .99 significance points are 5.99 and 9.21 respectively. Wilks' statistic behaves well even in the almost non-regular case $k=0.4$ for which only the second moment of the score is finite, but for $n=100$ and the totally regular case $k=-0.6$ the nominally chi-squared statistics based on observed and expected information are not close to their asymptotic distributions.

The immediate practical implication of these results is that confidence regions for the parameters of the generalized Pareto distribution should be based on the likelihood ratio statistic rather than the asymptotic Normality of the estimators, which is approached only rather slowly even in totally regular cases. They also suggest

		n=25		n=50		n=100	
		.95	.99	.95	.99	.95	.99
k=0.4	ℓ	.960	.987	.963	.995	.941	.985
	O	.770	.831	.827	.891	.819	.903
	E	.901	.926	.871	.910	.802	.854
k=0.2	ℓ	.952	.991	.937	.986	.941	.990
	O	.746	.818	.829	.892	.861	.926
	E	.802	.852	.762	.829	.795	.845
k=0.0	ℓ	.943	.991	.929	.987	.958	.991
	O	.774	.844	.834	.908	.913	.961
	E	.777	.829	.801	.880	.886	.941
k=-0.2	ℓ	.944	.988	.938	.989	.944	.983
	O	.802	.869	.859	.921	.914	.958
	E	.780	.836	.832	.888	.897	.945
k=-0.4	ℓ	.925	.987	.973	.991	.930	.991
	O	.815	.878	.917	.938	.902	.954
	E	.788	.852	.896	.926	.904	.954
k=-0.6	ℓ	.946	.992	.944	.986	.954	.990
	O	.838	.905	.882	.941	.927	.965
	E	.826	.887	.871	.938	.917	.964

Table 6.4 Observed proportion of likelihood confidence region statistics less than the 95- and 99-percentage points of their asymptotic distribution in 1000 samples of size n for different underlying values of k.

Key to statistics:

ℓ = twice loglikelihood ratio - Wilks' - statistic;
O = observed information matrix;
E = expected information matrix.

		n=25		n=50		n=100	
		.95	.99	.95	.99	.95	.99
k= 0.4	ℓ	5.53	10.07	5.52	7.90	6.18	9.87
	O	36.17	146.3	18.68	52.54	14.70	24.53
	E	15.01	65.56	26.01	200.5	46.61	881.1
k= 0.2	ℓ	5.96	9.07	6.37	10.33	6.38	9.13
	O	37.79	135.1	17.23	50.09	11.65	23.43
	E	32.17	215.3	42.61	245.7	29.28	130.6
k= 0.0	ℓ	6.15	8.61	6.88	9.36	5.75	8.75
	O	27.95	123.1	15.8	39.62	8.04	19.43
	E	46.54	279.7	27.4	241.3	16.83	29.06
k=-0.2	ℓ	6.42	9.47	6.47	9.86	6.34	9.82
	O	22.38	91.79	12.24	26.17	8.57	20.35
	E	35.58	163.1	15.24	42.85	9.70	23.03
k=-0.4	ℓ	7.94	9.70	6.37	8.87	6.75	8.87
	O	28.38	82.10	10.61	19.13	8.86	16.61
	E	48.49	120.9	12.54	23.62	8.98	17.80
k=-0.6	ℓ	6.11	8.88	6.35	9.43	7.20	9.11
	O	14.10	43.64	10.70	23.14	11.12	14.74
	E	17.09	60.85	11.25	23.73	11.31	15.50

Table 6.5 Observed 95 and 99 α -points of statistics for likelihood-based confidence regions, in 1000 samples each of size n , for different underlying values of k .

Regions based on:

ℓ = twice loglikelihood ratio - Wilks' - statistic;
O = observed information matrix;
E = expected information matrix.

the more general speculation: is it possible to find analytically a correction factor for the small-sample distribution of Wilks' statistic, even in cases such as $k=0.4$ when the usual practice of basing an $O(n^{-1})$ correction on the third and higher moments of the score breaks down?

6.2.3 Influence and censoring

In the study of statistical extremes it is important to appreciate the extent to which estimation depends on the few largest or smallest observations. In order to quantify this I now present some results on the asymptotic loss of information due to censoring, and investigate the properties of theoretical influence curves for maximum likelihood estimators.

Consider therefore the distribution

$$F(y) = \begin{cases} \{1-H_c(y)\} [1-(1-ky/\sigma)^{1/k}] + H_c(y) & (k \neq 0) \\ \{1-H_c(y)\} [1-\exp(-y/\sigma)] + H_c(y) & (k=0) \end{cases}$$

($\sigma > 0$, $y > 0$, $1-ky/\sigma > 0$),

which is generalized Pareto for $y < c$ but puts an atom of probability of size $\alpha = (1-ky/\sigma)^{1/k}$ if $k \neq 0$, or $\alpha = \exp(-y/\sigma)$ if $k=0$, at c . The Heaviside function $H_c(y)$ here is zero when $y < c$ and one otherwise. The components of the Fisher information matrix $i_\theta(c)$ for a single observation drawn from this distribution are

$$-E[\partial^2 \ell / \partial \sigma^2] = \{1-(1-kz)^{1/k-2}\} / \sigma^2(1-2k),$$

$$-E[\partial^2 \ell / \partial \sigma \partial k] = -\{1-(1+z-2kz)(1-kz)^{1/k-2}\} / \sigma(1-k)(1-2k),$$

$$-E[\partial^2 \ell / \partial k^2] = \{2-(2+2z+z^2-4kz-3kz^2+2k^2z^2)(1-kz)^{1/k-2}\} / (1-k)(1-2k),$$

where $z=c/\sigma$, provided that $k < 1/2$.

The overall asymptotic information loss for estimation of k and σ is

$$\{\det(i_{\theta}(c))/\det(i_{\theta})\}^{1/2} \times 100\% ,$$

which is given in Table 6.6 for several values of k and probability α that the exact value of Y is unobserved and hence set equal to c . Cox and Hinkley(1974) discuss this as a measure of large-sample relative efficiency of different asymptotically Normal estimators based on the same data, whereas here it is used to compare efficiency of asymptotically optimal estimators derived under different sampling schemes. Large-sample results may not apply to small and medium samples, but they provide useful guidelines.

k	probability $\alpha(\%)$				
	1	5	10	20	50
-0.4	95.4	81.9	69.1	49.9	16.3
-0.2	92.8	76.2	62.2	43.1	13.2
0.0	87.6	67.4	52.9	34.9	9.9
0.2	76.5	53.7	40.1	25.1	6.5
0.4	49.3	30.6	21.5	12.6	3.0

Table 6.6: Asymptotic relative efficiency(%) of upper-truncated maximum likelihood estimation of the generalized Pareto distribution for different truncation probabilities α .

The information loss is severe for positive k even for α as low as 0.05, and for all values of k considered as α increases from 0.1. These results show where information for estimation of extremes comes from: in large samples truncation of the top 1% of the data can lead to the loss of one-half of the total sample information for the

parameters! This is related to the fact that for $k > 1/2$ the uppermost order statistic is a superefficient estimator of the endpoint σ/k . There is also a close connection with the theoretical influence curves for the estimators.

The influence curve $IC_{T,F}(y)$ of an estimator $T(\cdot)$ at a distribution $F(\cdot)$ measures the suitably standardized effect on T of adding a single observation at the point y as the sample size approaches infinity when the model under consideration is correct. Mathematically it is defined as the Frechet derivative of the statistical functional $T(\cdot)$ at $F(\cdot)$ in the direction of the distribution function δy which puts weight on at y :

$$IC_{T,F}(y) = \lim_{\epsilon \rightarrow 0} \frac{T[(1-\epsilon)F + \epsilon\delta y] - T[F]}{\epsilon},$$

provided the limit exists at each y in the domain of F .

Hampel(1968) and Andrews et al.(1972) use the influence curve to compare estimators and to suggest robust alternatives to them. I use it only to assess the sensitivity of maximum likelihood estimators to observations corresponding to various quantiles of the underlying distributions. For such estimators the curves are defined thus:

$$IC_{\theta,F}(y) = i_{\theta}^{-1} \partial \ell(y; \theta) / \partial \theta,$$

the sample version of which is approximately $\hat{\theta} - \theta$ by the usual large-sample theory. These functions are shown in Table 6.7 for different values of k for several percentage points of their respective distributions, together with values of the influence curves of maximum likelihood estimators for the $N(\mu, \sigma^2)$ and exponential(λ) distributions, for comparison. The estimators for the generalized Pareto distribution are more heavily influenced by observations at the upper quantiles of the distribution than for the Normal and exponential distributions.

Probability $Y > y$

k	0.1	0.05	0.02	0.01	0.005	0.002	0.001
0.4	0.79	1.06	1.22	1.16	0.87	0.04	-1.07
0.2	0.95	0.91	0.33	-0.58	-1.95	-4.58	-7.31
0.0	0.95	0.50	-0.83	-2.39	-4.44	-7.88	-11.04
-0.2	0.82	-0.05	-2.02	-4.02	-6.42	-10.09	-13.20
-0.4	0.59	-0.68	-3.13	-5.41	-7.96	-11.64	-14.59
-0.6	0.29	-1.33	-4.14	-6.59	-9.21	-12.85	-15.68

a. Influence curves $IC(y)$ for the maximum likelihood estimator \hat{k} for the generalized Pareto distribution.

k	0.1	0.05	0.02	0.01	0.005	0.002	0.001
0.4	1.57	2.25	2.97	3.32	3.45	3.18	2.48
0.2	2.00	2.51	2.66	2.31	1.49	-0.41	-2.59
0.0	2.26	2.50	2.08	1.21	-0.14	-2.67	-5.14
-0.2	2.38	2.34	1.48	0.30	-1.26	-3.83	-6.11
-0.4	2.41	2.11	0.95	-0.36	-1.94	-4.34	-6.32
-0.6	2.38	1.86	0.52	-0.82	-2.34	-4.50	-6.22

b. Influence curves $IC(y)$ for the maximum likelihood estimator $\hat{\sigma}$ for the generalized Pareto distribution.

	0.1	0.05	0.02	0.01	0.005	0.002	0.001
$\hat{\lambda}/\lambda$	1.30	2.00	2.91	3.61	4.30	5.22	5.91
$\hat{\mu}$	1.28	1.65	2.05	2.33	2.58	2.88	3.09
$\hat{\sigma}/\sigma$	0.32	0.85	1.61	2.21	2.82	3.64	4.28

c. Influence curves $IC(y)$ for maximum likelihood estimators for exponential and Normal distributions.

Table 6.7: Comparison of influence curves

Particular conclusions to be drawn from these results will depend on the use to which they are to be put. A plain implication in the present context is that any inadequacies of MESOS leading to systematic errors in the calculation of the most extreme exposures will exert a very strong influence on statistical models for high exposure episodes based on the generalized Pareto distribution. More generally, in the study of sample extremes through the methods studied here very careful attention must be paid to events leading to the few highest observations, as these have considerable impact on inference - perhaps not a very surprising point. A comment relevant to collection of data is that it is important to ensure that instruments are well-calibrated even out of their usual range, so that events which seem rather unlikely a priori are recorded as accurately as possible. Powerful statistical methods cannot compensate for unreliable data, especially in the statistics of extremes, as is clear from Tables 6.6 and 6.7.

6.3 Some other estimators

6.3.1 Least squares estimators

In the special case when $k_i = k$ for all observations and there is no censoring, naive least squares may be used to find estimates of the parameters θ , as follows. Assume that $\sigma_i = \exp(x_i \beta^T)$, with x_i a $p+1$ vector of covariates including an overall mean effect. Take logarithms of the data Y to see that the n -vector $V = \log(Y)$ satisfies

$$V = X^t \beta + (\beta_0 + \kappa_1) j + \varepsilon ;$$

where ε is a n -vector of independent log-generalized Pareto variates with zero mean; κ_r ($r > 1$) are the cumulants of the log-generalized Pareto distribution with parameters l and k ; and j is an n -vector of

ones. Without loss of generality parametrize so that the columns of X sum to zero: i.e. the mean is orthogonal to regression effects.

The usual least-squares estimate $\tilde{\beta} = (X^tX)^{-1}X^tV$ of β is unbiased and asymptotically Normally distributed with covariance matrix $n^{-1}(X^tX)^{-1}\kappa_2$ (Cox and Hinkley, 1968) and the inverse Fisher information for the maximum likelihood estimates $\hat{\beta}$ of β is $(X^tX)^{-1}(1-2k)$, provided $k < 1/2$. The asymptotic relative efficiency of $\tilde{\beta}$ relative to $\hat{\beta}$, $A_1(k) = (1-2k)/\kappa_2$, is given in Table 6.8 for several values of k. For positive k it drops rapidly as k approaches $1/2$, but is quite high for negative k.

	shape parameter k						
	-1.0	-0.75	-0.5	-0.25	0.0	0.25	0.5
$A_1(k) \times 100\%$	91.2	91.2	83.7	77.8	60.8	35.1	0.0
$A_2(k) \times 100\%$	78.0	69.4	57.1	42.2	26.9	13.5	0.0

Table 6.8 Asymptotic relative efficiency of least squares estimates of parameter θ .

Estimates of β_0 and k may be based on the residual sum of squares of the model; their joint efficiency $A_2(k)$, the square root of the ratio of the determinants of the covariance matrices of their maximum likelihood and least squares estimates, is also displayed in Table 6.8. The asymptotic relative efficiency of θ in a model with p covariates is

$$\{ A_1(k)PA_2(k)^2 \}^{1/(p+2)} \times 100\% ,$$

which approaches A_1 for large p. $A_2(k)$ is substantially less than $A_1(k)$ over the range $|k| < 1/2$ of usual values of k, indicating that the overall loss of information due to use of least squares estimates for θ can be severe. If k is thought to be negative or close to zero,

least squares fitting either to explore data to find suitable models later to be fitted by maximum likelihood or to find consistent starting values for β for numerical maximum likelihood procedures may be useful; but it cannot be recommended if k is thought to be positive.

6.3.2 Moment Estimators

When the appropriate moments of random variables Y_i exist, consistent parameter estimates based on sample moments may be found by equating their sample and theoretical values and solving the resulting expressions simultaneously. In the case of a simple random sample of size n from the generalized Pareto distribution, the equations are

$$\sigma/(1+k) = n^{-1} \sum_{i=1}^n y_i = \bar{Y} ,$$

and

$$\sigma^2/(1+k)(1+2k) = (n-1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 = S^2 .$$

These give estimators k^* and σ^* ;

$$k^* = \{ \bar{Y}^2/S^2 - 1 \} ,$$

$$\sigma^* = \bar{Y} \{ 1 + \bar{Y}^2/S^2 \} ,$$

which only have finite variance-covariance matrix for $k > -1/4$. This matrix is easily found to $O(n^{-1})$ by making Taylor expansions and recalling that in terms of the cumulants κ_r of the generalized Pareto distribution, $\text{Var}(\bar{Y}) = \kappa_2/n$, $\text{Cov}(\bar{Y}, S^2) = \kappa_3/n$, and $\text{Var}(S^2) = \kappa_4/n + 2\kappa_2/(n-1)$.

The asymptotic relative efficiency of k^* and σ^* is given in Table 6.9 for $-1/4 < k < 1/2$, the only range in which they can be compared with the asymptotically optimum maximum likelihood estimates. It is fairly low except for $k \approx 0$, where it is very high

because when k is zero the likelihood equations and those determining the moment estimates are the same. The general conclusion is that except for $k=0$, moment estimates are not very good. However if $k>1$, when the likelihood equations fail to yield even consistent estimates, those based on moments may be useful, although their variances are large so that big datasets will be needed to give the estimators reasonable standard errors - see Table 6.10.

k	ARE(k^*, σ^*)	ARE(k^*)	ARE(σ^*)
-0.25	0.0	0.0	0.0
-0.20	51.2	28.9	32.1
-0.15	75.5	58.4	66.3
-0.10	90.2	81.7	87.8
-0.5	97.8	95.7	97.5
0.00	100.0	100.0	100.0
0.05	98.2	96.4	98.4
0.10	93.4	87.5	94.5
0.15	86.5	75.9	89.3
0.20	78.1	63.5	83.6
0.25	68.6	51.7	77.5
0.30	58.4	41.2	71.4
0.35	47.6	32.2	65.2
0.40	36.2	24.7	59.3
0.45	23.6	18.7	53.4
0.50	0.0	0.0	0.0

Table 6.9: Asymptotic relative efficiency(%) of moment estimates of generalized Pareto distribution

k	Var(k^*)	Var(σ^*)	Cov(k^*, σ^*)
0.0	1.00	2.00	1.00
0.5	1.80	2.10	1.80
1.0	4.80	2.53	3.40
1.5	10.39	3.00	5.52
2.0	19.29	3.49	8.14
2.5	32.23	3.97	11.27
3.0	49.97	4.47	14.89
3.5	73.25	4.96	19.02
4.0	102.80	5.46	23.64

Table 6.10: Standardized variances and covariances of moment estimates k^* and σ^* of generalized Pareto distribution

6.4 Two tests for tail weight

Although the generalized Pareto law is a single parametric family of distributions, different values of k correspond to qualitatively different types of tail behaviour: power-law for $k < 0$; exponential for $k = 0$; and curtailed above for $k > 0$. In some circumstances it may be necessary to test the hypothesis $k = 0$ against either or both alternatives k positive or negative. Hosking(1984) compares a number of tests for this in the closely related case of the generalized extreme-value distribution: here I give a small discussion of the problem for simple random samples of size n from the generalized Pareto law. Two relatively informal graphical procedures which relate to this situation were given in Chapter 5. Here I deal with formal tests.

The scale parameter σ is a nuisance parameter: the null hypothesis $k = 0$ is composite. There are broadly two approaches to obtaining exact tests for composite null hypotheses, based on appeal

to the invariance principle or on the construction of similar critical regions.

Exact similar test statistics may be found by observing that if $k=0$, then the statistic $Z = \sum Y_i$ is sufficient for σ , and hence the conditional distribution of the data $\{Y_i\}_{i=1}^n$ given the observed value z of Z is independent of σ . Z is certainly boundedly complete and so critical regions ω_α such that

$$\Pr[t(Y) \in \omega_\alpha \mid Z=z; k=0] = \alpha$$

for a suitable test statistics $t(Y)$ have Neyman structure and are similar of size α . By the Neyman-Pearson lemma, the uniformly most powerful tests for the alternatives $k>0$ and $k<0$ against $k=0$ are based on the ratio of the conditional likelihoods of the data under the hypotheses, or equivalently the conditional score statistic.

Here a problem arises: the joint density of the data $\{Y_i\}_{i=1}^n$ given the value of Z is unknown under the alternative hypothesis, so it is ^{not} obvious how to find the relevant score. However the unconditional score statistic is easily derived and turns out to be equivalent to

$$G = \sum_{i=1}^n (Y_i/Z)^2$$

conditionally on the observed value z of Z . Then under the null hypothesis the quantities $D_i = Y_i/z$ are distributed exactly as the spacings of a sample of size $n-1$ from the unit uniform distribution. Clearly the distribution of G is independent of σ either unconditionally or conditionally on Z .

The statistic G , or Greenwood's statistic, has a long history. Neyman(1941) derived it as a test for overdispersion in analysis of variance; Greenwood(1946) proposed it as a test for departures from Poissonness of a series of events; and Moran(1947, 1951, 1953) and

others have studied it in the context of point processes. See also Pyke(1965). The distribution of G is unknown even under the null hypothesis, but Burrows(1979), Currie(1981) and Stephens(1981) give its significance points for selected sample sizes up to $n=500$. G is asymptotically Normal, but its approach to Normality is unusually slow. It has support in the interval $[1/n, 1]$, mean $\mu_G = 2/(n+1)$, and variance $\sigma_G^2 = 4(n-1)/\{(n+1)^2(n+2)(n+3)\}$. Large values of $T = (\mu_G - G)/\sigma_G$ indicate $k > 0$, and conversely. Significance points of T are easily found from those of G .

Because of the slow approach to unit Normality by T when $k=0$, it is compared with a statistic

$$S = \left\{ 2 \sum_{r=n}^{n-1} (n-r)Y_{(r)} / Z - (n-1)/2 \right\} \left[12/(n-1) \right]^{1/2}$$

whose null distribution is known to converge to it rapidly.

The statistic S was suggested by the fact that when $k=0$, the quantities $V_1 = nY_{(1)}$, $V_2 = (n-1)(Y_{(2)} - Y_{(1)})$, ..., $V_n = (Y_{(n)} - Y_{(n-1)})$ are independent exponential variates with parameter σ . The statistic $V = \sum_r V_r$ is sufficient for σ , and the joint distribution of the $U_{(r)} = \sum_{i=1}^r V_i$, ($r=1, \dots, n-1$), given that $\sum_r V_r = v$, is the joint distribution of order statistics from a random sample of size $n-1$ from the uniform distribution on $[0, v]$. To see the effect of $k \neq 0$ on the $U_{(r)}$, recall from Chapter 5 that

$$E[Y_{(r+1)} - Y_{(r)}] = \sigma(n-r+k)^{-1} \prod_{i=1}^r (1+ka_{i,n})^{-1}$$

where $a_{i,n} = (n-i+1)^{-1}$, so that

$$E[U_{(r)}] = \sigma \sum_{i=1}^r \prod_{i=1}^i (1+ka_{i,n})^{-1} .$$

If $k > 0$ then the $U_{(r)}/V$ will tend to increase, and if $k < 0$ they will tend to decrease, relative to their expected positions when $k=0$. The

statistic $H = \frac{\sum_{r=1}^{n-1} U_{(r)}}{V} = 2 \frac{\sum_{i=1}^{n-1} (n-i)Y_{(i)}}{\sum_{i=1}^n Y_i}$ has the Irwin-Hall distribution, tending rapidly to Normality with mean $(n-1)/2$ and variance $(n-1)/12$ as $n \rightarrow \infty$, under the null hypothesis; large values of H indicating $k > 0$ and vice versa.

A simulation experiment to compare the power of the statistics S and T in small samples was performed for values of k in the range $-\frac{1}{2}$ to $\frac{1}{2}$, taking 100 samples each of sizes 10, 25 and 50. Table 6.11 shows the observed power of S and T for tests of nominal size 0.05 against one-sided alternatives. Figure 6.1 displays some of the same information. After allowing for sampling variation, it seems that the tests have very similar power for $n=10$, but that for $n=50$ T is a little more powerful for tests of $k < 0$, and S is more powerful for tests of $k > 0$. The rather poor apparent size of the test based on T for $n=25$ and 50 may be due to inaccuracies in Stephens' logNormal approximation for the percentage points of G .

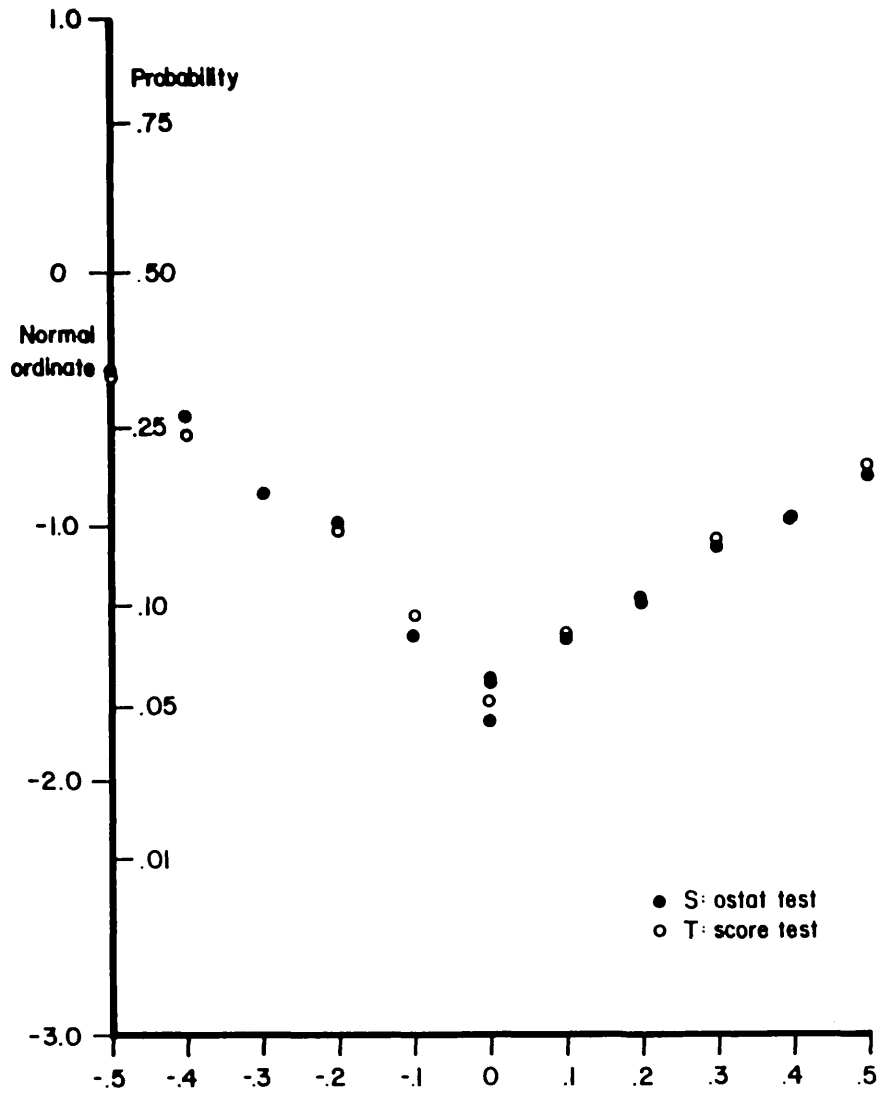
The message from this study is that in smallish samples the statistics S and T have about the same power against the alternatives $k \neq 0$: S will usually be preferred on the grounds that it does not need to be referred to special tables.

Other test statistics could be based on the hazard function $(\sigma - ky)^{-1}$ of the of generalized Pareto distribution. A test for k negative would then be for decreasing failure rate, and one for k positive for increasing failure rate. Proschan and Pyke(1967) give some tests for increasing failure rate. I shall not consider these further here but there are clearly many possibilities.

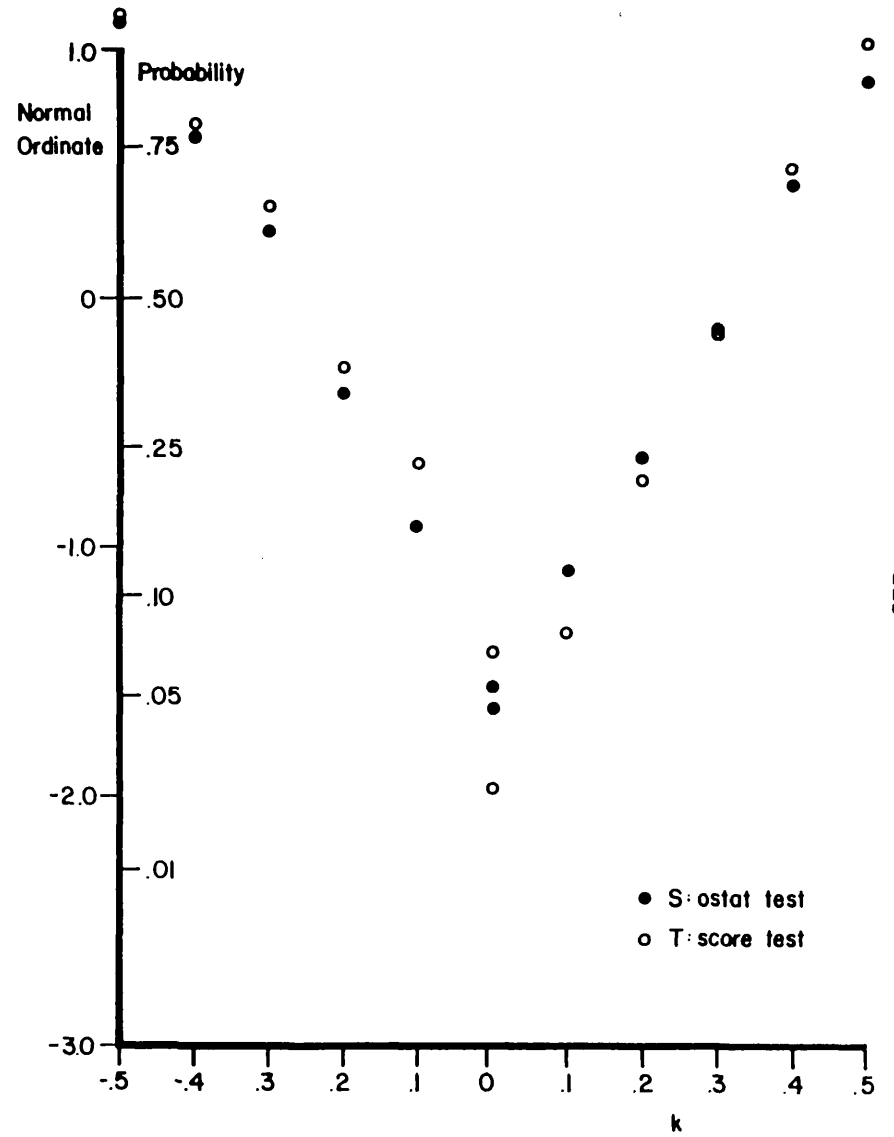
The problem of finding good test statistics for $k \neq 0$ in more complex cases of covariate-dependent data has not been studied here, but it seems likely that it is very hard to find exact tests and that asymptotic theory must be deployed.

	n=10				n=25				n=50			
	S		T		S		T		S		T	
	-	+	-	+	-	+	-	+	-	+	-	+
-.5	.350	.013	.340	.013	.628	.001	.649	.001	.866	.001	.873	.001
-.4	.284	.016	.262	.017	.506	.004	.552	.002	.742	0.0	.757	0.0
-.3	.192	.018	.192	.019	.360	.007	.392	.003	.608	.002	.643	.002
-.2	.161	.017	.152	.017	.256	.012	.317	.007	.352	.006	.391	.004
-.1	.075	.038	.086	.036	.122	.036	.177	.022	.179	.017	.224	.009
0.0	.038	.053	.037	.055	.043	.048	.065	.026	.058	.050	.077	.024
.1	.027	.074	.028	.076	.016	.086	.017	.048	0.010	.138	.004	.090
.2	.016	.099	.010	.097	.003	.170	.003	.107	0.0	.262	0.0	.234
.3	.004	.140	.003	.144	.002	.263	.001	.183	0.0	.446	0.0	.444
.4	.007	.165	.003	.167	0.0	.367	0.0	.278	0.0	.673	0.0	.698
.5	.003	.212	.002	.223	0.0	.509	0.0	.434	0.0	.809	0.0	.846

Table 6.11 Power of S and T for different values of k:
proportion of 1000 simulations significant at the one-sided 5% level.



(a) sample size $n = 10$.



(b) sample size $n = 50$.

Figure 6.1 : Power of S and T in small samples.

7. DIAGNOSTICS

Techniques commonly used for checking goodness of fit in statistical modelling include:

- (i) the inspection of residuals;
- (ii) embedding the fitted model in a more comprehensive alternative and testing fit against the larger model;
- (iii) the study of influence.

I consider some possible modifications of these ideas suitable in the present context.

7.1 Residuals

Plots and other procedures based on residuals are commonly used in statistical modelling. Atkinson(1982), Cook and Weisberg(1982), Green(1984) and McCullagh and Nelder(1983) are some recent references. Several possible definitions of residuals are available: those due to Cox and Snell(1968), and the deviance residuals defined by Pregibon(1981) are discussed here.

Cox and Snell gave a general definition of residuals for a broad class of models, and suggested a method of approximating their means, variances, and covariances. For a model where the observations Y_i have some expression

$$Y_i = g_i(\theta, \epsilon_i) \quad (i=1, \dots, n)$$

in terms of a vector of unknown parameters θ and the independent identically distributed variables ϵ_i , and the equations

$$Y_i = g_i(\hat{\theta}, R_i)$$

have a unique solution for R_i in terms of Y_i and the maximum likelihood estimate $\hat{\theta}$ of θ , they define the crude or unadjusted

residuals R_i as that solution and go on to find means, variances and covariances of the R_i to order n^{-1} , which can then be used to modify the R_i to have the same means and variances to that order. Their results suggest that even in fairly small samples these adjustments make little difference to the R_i for plotting purposes.

With this in mind, define

$$R_i = -\hat{k}_i^{-1} \log(1 - \hat{k}_i y_i / \hat{\sigma}_i) \quad (i=1, \dots, n),$$

which are crude residuals in the sense above. These should be distributed approximately as independent unit exponential variables when the model is correct, and so may be checked for outliers, dependence, and distributional form in the usual way using graphical techniques for the exponential distribution. The values of test statistics should be interpreted with great care, and can be misleading in many cases, as Durbin's contribution to the discussion following Cox and Snell(1968), and their 1971 paper make clear. Levels of significance of test statistics may be seriously underestimated if they are formed from unadjusted residuals, so their systematic use is not recommended.

In some circumstances it may be useful to use the $U_i = 1 - \exp(-R_i)$, which should be approximately uniform on $(0,1)$, rather than the R_i . Two cases when this may be informative are:

(i) when the data is divided into subsamples, in which case plotting the U_i against subsample number or some meaningful physical quantity may reveal distortions in the fitting of the GPD to specific subsamples, or discrepancies between them;

(ii) when plotting the data to investigate the possibility of serial dependence or clustering into groups (possibly occurring together in time) of the observations.

If in these situations the plots are made on the original scale,

the skewness of the R_i may lead to difficulties of interpretation alleviated by plotting the U_i .

Another possible general definition of residuals is motivated as follows. Suppose there are n independent observations y_i with associated unconstrained parameters θ_i , so that the total loglikelihood of the data is

$$\ell = \sum \ell_i(y_i; \theta_i),$$

for the full or maximal model. As it stands this model is not especially useful, since unless the θ_i are linked in some way there will be $\dim\{\theta_i\}$ parameters for each observation, and so at least n parameters in all. Fewest parameters are needed if $\theta_i = \theta$ for all i , which amounts to proceeding as if all the variates were identically distributed: the minimal model. A reasonable summary usually lies between these two extremes, and constrains the parameter vector $\theta = (\theta_1, \dots, \theta_n)$ to lie on a sub-manifold $\theta = \theta(\beta)$ of the full $\dim\{\theta\}$ -dimensional parameter space. The maximum loglikelihood achievable is attained on the full model, and the minimum on the minimal model; other values lie in-between.

The idea behind deviance residuals, exploited by Pregibon(1981) and others, is to measure the discrepancy between the full model and the intermediate model $\theta = \theta(\beta)$ due to individual observations by considering the quantities

$$r_D(y_i; \theta) = \text{sgn}(\tilde{\theta}_{(i)} - \hat{\theta}) / \{ 2(\ell(\tilde{\theta}_{(i)}) - \ell(\hat{\theta})) \},$$

the deviance residuals, where $\hat{\theta} = \theta(\hat{\beta})$ is the maximum likelihood estimate of θ under the intermediate model and $\tilde{\theta}_{(i)}$ is the value of θ whose j^{th} components ($j \neq i$) equal those of $\hat{\theta}$ but whose i^{th} component is chosen to maximise $\ell(\tilde{\theta}_{(i)})$. The quantities r_D should be approximately independently unit Normally distributed if the model

$\theta = \theta(\beta)$ is adequate, and may be inspected to check this.

Jorgensen(1983) gives some more definitions of residuals, all quite closely related to the r_D .

An important point is that the r_D defined exactly as above cannot be used if a generalized Pareto distribution with separate shape k_i as well as scale σ_i parameters is regarded as the maximal model. For then the optimum difference of loglikelihoods is achieved when an atom of probability is put at each observation: all the r_D equal $\pm\infty$! If however the maximal model has a common shape parameter k this problem is avoided and the r_D can be sensibly defined. This may or may not be a relevant point in other contexts - Pregibon did not have to face it since his concern was with logistic regression models - but it is a potential difficulty here. For the case of independent generalized Pareto variates Y_i with difference scales σ_i but the same shape k , the deviance residuals may be written thus:

$$r_{Di} = \text{sgn}(y_i/\hat{\sigma}_i - 1) \sqrt{2} \left[(1/\hat{k} - 1) \log\{(1-\hat{k}) / (1 - \hat{k}y_i/\hat{\sigma}_i)\} - \log\{y_i/\hat{\sigma}_i\} \right].$$

Two types of residual are now defined, and it is of some interest to determine their relationship, especially as McCullagh and Nelder(1983) suggest that deviance residuals are very close to Normality. One possible method of assessing the relationship for variates Y_i with continuous distribution function $F(y;\theta)$ is as follows.

For such variables Cox-Snell residuals can be formed simply by taking the probability integral transform of Y . Then $F(Y;\theta)$ is uniformly distributed in the unit interval and $X = \Phi^{-1}(F(Y;\theta))$ has the standard Normal distribution. Provided the function F^{-1} is well-defined this suggests that the random variables $G(X) = r_D(F^{-1}(\Phi(X);\theta);\theta)$ and X be compared. For if the random variable $r_D(Y;\theta)$ is exactly unit Normal then obviously $G(x)=x$: a

straight line of unit slope through the origin. Departures from Normality manifest themselves as departures from this null behaviour.

The graph may be interpreted as a plot of deviance residuals for a very large sample whose underlying distribution function is F , against Normal order statistics. The correlations among the residuals are asymptotically negligible provided a fixed number of parameters is being estimated, and in the limit $G(x)$ is observed. The effect on the residuals of assuming the deviance corresponding to a distribution function F when the true distribution function of the data is H may be seen by considering $r_D(H^{-1}(\Phi(x)))$ as a function of x . In this case however, the effect of parameter estimation should be taken into account. I shall discuss this no further.

Taking the case $F(y;\theta)=\Phi((y-\theta_1)/\theta_2)$, obviously $r_D(y;\theta)=(y-\theta_1)/\theta_2$ and hence $G(x)=x$, exactly as expected.

For the case of the generalized Pareto distribution, the function $G(x)$ is plotted in Figure 7.1 for several values of k , together with the line $G(x)=x$ for comparison. The residuals are negatively biased by an amount which depends on k , but the plots are almost straight. The function G and its first three derivatives at $x=0$ are tabulated in Table 7.1 for the same range of values of k . The table emphasises the impression gained from the figure: the function G is very nearly linear for the usual range of values of k . The table can be used to modify the r_D to produce residuals with asymptotic slope and intercept unity and zero respectively and $x=0$, which might be valuable in some applications.

Cox-Snell seem preferable to deviance residuals because they are asymptotically unbiased and there are no difficulties defining them when the k_i are not all equal.

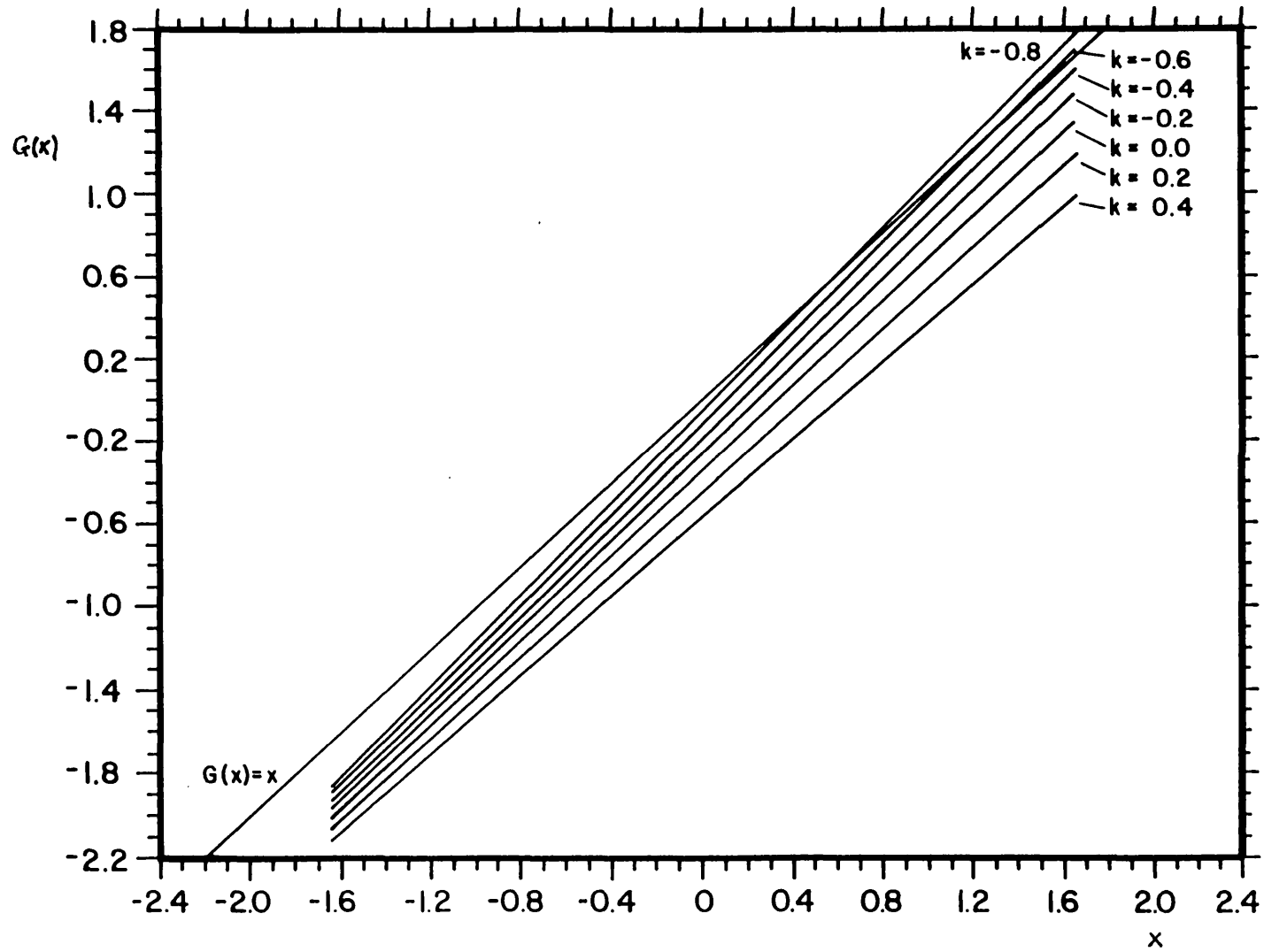


Figure 7.1 : Comparison of deviance and Cox-Snell residuals for generalized Pareto distribution.

k	$G(0)$	$G'(0)$	$G''(0)$	$G^{(3)}(0)$
0.4	-0.551	0.945	-0.017	-0.011
0.2	-0.440	0.987	-0.002	-0.008
0.0	-0.345	1.023	0.007	-0.009
-0.2	-0.262	1.052	0.012	-0.013
-0.4	-0.187	1.076	0.013	-0.020
-0.6	-0.119	1.097	0.011	-0.029
-0.8	-0.057	1.114	0.007	-0.039

Table 7.1 Residual comparison function $G(\cdot)$ and its first few derivatives for several values of k

7.2 A score test

Possible more formal methods of checking fit of a model to data are to test the fit of the model against alternatives:

(i) which make the same distributional assumptions but include more factors or combinations of them in the systematic part of the model;

(ii) representing plausible but separate families of hypotheses (Cox, 1961,1962);

(iii) which embed the random component of the model in some more general distribution;

(iv) in which the systematic part of the model applies to some transformation of the data (Box and Cox, 1964).

Here the possibilities are logically distinct; I pursue (iii) but tests for alternatives of type (i) can be made simultaneously by generalizing the procedure proposed.

The question of transformations is an important one from the point of view of improving the generalized Pareto approximation to a set of sample excesses over some threshold, which as an asymptotic approximation need not be good in small samples, although it may be expected to improve as the threshold increases. As this happens, however, less of the sample will be available and the uncertainty of eventual conclusions will increase, so it may be wise to transform the data and then to apply techniques based on thresholds, thus retaining a reasonable proportion of the data for analysis whilst improving the approximation. Ideally such a transformation should be based on knowledge of the physical processes underlying the data, but in situations where suitable assumptions cannot be made, the possibility of more or less empirical transformations should be considered.

Here a score test for the possibility that not Y_i/σ_i , but $\phi^{-1}\{\exp(\phi(Y_i/\sigma_i)^\lambda)-1\}$, has the generalized Pareto distribution with parameters k and 1 , for some $\lambda > 0$ and $-\infty < \phi < \infty$, is considered. For $\phi=0$ and $\lambda=1$ this reduces to the null hypothesis. Others have constructed score tests in not dissimilar contexts: Atkinson(1973, 1982) gives tests in case (iv) for the Box-Cox family of power transformations and for the folded-power family for proportions (Mosteller and Tukey, 1977); Cook and Weisberg (1983) give a score test for homogeneity of variance in the normal-theory linear model; and Spiegelhalter(1983) has derived several score tests for checking distributional assumptions against specific alternatives. The test is locally uniformly most powerful for the alternatives considered, but is specific to them, so should be used together with residual plots to reveal other possible discrepancies of the data. Significance of the test means that a transformation of the form $\phi^{-1}\{\exp(\phi(y_i/\sigma_i)^\lambda)-1\}$ for some σ_i , ϕ and λ should improve the fit of the generalized Pareto distribution to the data.

Here I construct the test for independent variates Y_i with $\sigma_i = \exp(\beta_0 + \beta^T x_i)$ and common shape parameter k , parametrized so that $\sum_{i=1}^n x_{iu} = 0$ for $u=1, \dots, p$; i.e. the parameters β_1, \dots, β_p are orthogonal to β_0 . These conditions are not essential to the derivation of the test but they clarify the details. For more complex models, perhaps with $k_i \neq k$ for all i , the test statistic is probably best computed numerically.

After some tedious calculations, it emerges that provided $k \neq 0$ and $|k| < 1/2$, the inverse information matrix for the parameters $(\beta, k, \beta_0, \lambda, \phi)$ when $\lambda=1$ and $\phi=0$ is

$$\begin{vmatrix} (X^T X)^{-1}(1-2k) & 0 \\ 0 & P \end{vmatrix}$$

where the 2x2 corner submatrix of P corresponding to (λ, ϕ) is

$$2 \Delta^{-1} \begin{vmatrix} 2k^{-2}(1+2k)(1+k)^2(b-k^2-1) & (k+1)(1-4k^2)(1-2k+2c)k^{-1} \\ (k+1)(1-4k^2)(1-2k+2c)k^{-1} & 2(1-2k)(4k^2+2k+1) \end{vmatrix}$$

with

$$\Delta = (16bk^2 + 8bk + 4b - 32ck^3 - 24k^3 + 16k^2c^2 - 20k^2 + 8kc + 16ck^2 - 4k - 4c^2 - 4c - 5).$$

Here $b = \pi^2/6 + \psi'(-1/k)$ and $c = \log(-k) + \psi(-1/k) - \psi(2)$ when $k < 0$, and $b = \pi^2/6 + \psi'(1+1/k)$ and $c = \log(k) + \psi(1+1/k) - \psi(2)$ when $k > 0$; with $\psi(\cdot)$ and $\psi'(\cdot)$ respectively the first and second derivatives of the log-gamma function.

The corresponding elements of the score vector U are

$$U_\lambda = \sum_i (y_i / \sigma_i) (1 - \frac{1}{2}(1-k)(y_i / \sigma_i)(1 - ky_i / \sigma_i)^{-1})$$

and

$$U_\phi = \sum_i [1 + \ln(y_i / \sigma_i) - (1-k)(y_i / \sigma_i) \ln(y_i / \sigma_i)(1 - ky_i / \sigma_i)^{-1}].$$

When in addition it is required to test $k=0$, the parameters k and ϕ give the same type of first-order departure from the model, so that their components of the score vector are equal and i_θ is singular. One of them must be dropped, in which case we have the $p+3$ -sided matrix

$$i_\theta^{-1} = \begin{vmatrix} (X^T X)^{-1} & 0 \\ 0 & P^* \end{vmatrix}$$

where P^* is the matrix

$$\Delta^{-1} \begin{vmatrix} \pi^2/3 - (\gamma-2)^2 & -\gamma & \pi^2/6 - (\gamma-1)(\gamma-2) \\ -\gamma & 1 & -1 \\ \pi^2/6 - (\gamma-1)(\gamma-2) & -1 & \Delta+1 \end{vmatrix}$$

with $\Delta = \pi^2/6 - (\gamma-1)^2 - 1$ and γ Euler's constant. The parameter vector here is $\theta = (\beta, \beta_0, \lambda, \phi)$. The components of the score statistic are

$U_\phi = \sum_1 [\frac{1}{2}(y_1/\sigma_1)^2 - y_1/\sigma_1]$ and $U_\lambda = \sum_1 [1 + (1 - y_1/\sigma_1) \log(y_1/\sigma_1)]$, so that the test for the hypothesis $\phi=0, \lambda=1$ is based on the statistic

$$W_u = n^{-1} \Delta^{-1} [U_\phi^2 (\Delta+1) - 2U_\phi U_\lambda + U_\lambda^2]$$

which has asymptotic χ^2_2 distribution when it is true.

There are analytical difficulties with this test which arise when λ is included since then maximum likelihood theory is not regular if $k < -\frac{1}{2}$. Since in most environmental applications $|k| < \frac{1}{2}$ (Jenkinson, 1969), these difficulties are unlikely to restrict the use of the test much. For $k < -\frac{1}{2}$, a goodness-of-fit test may be based on the part of the score statistic corresponding to ϕ alone, which has asymptotic chi-squared distribution on one degree of freedom.

The small-sample properties of the test under the null hypothesis were investigated by simulation. One thousand samples of sizes 25, 50 and 100 were generated from the generalized Pareto distribution for values of $k = -0.6$ (0.2) 0.4, the parameters estimated by maximum likelihood, and their values of the statistic W_u grouped according to the value of \hat{k} . The results, shown in Table 7.2, demonstrate that W_u is under-dispersed relative to the χ^2_2 distribution, especially for $k > 0.3$. The nominal significance points of the statistic at levels .90, .95, .975, .99 are 4.61, 5.99, 7.38 and 9.21 respectively. The behaviour of W_u is unknown in more complex situations but Table 7.2 provides at least some guidance for small samples.

Finally, note that contributions to the test made by individual observations can be plotted in order to identify particular points or groups thereof influencing the statistic unduly; see Atkinson(1982) or Cook and Weisburg(1982) for discussion of this idea. This may help to distinguish aberrant values if a single datum or group of data contributes overmuch to the significance of the test, or

	$-.5 \leq \hat{k} < -.3$	$-.3 \leq \hat{k} < -.1$	$-.1 \leq \hat{k} < .1$	$.1 \leq \hat{k} < .3$	$.3 \leq \hat{k} < .5$	
n=25	.90	2.78	2.80	2.65	1.41	0.68
	.95	4.39	3.96	3.78	2.08	1.02
	.975	5.89	5.57	5.02	2.80	1.39
	.99	7.52	8.21	6.84	4.24	1.94
	M	652	818	996	961	862
n=50	.90	2.68	3.30	2.95	1.66	0.71
	.95	3.97	4.72	4.05	2.37	0.96
	.975	5.52	6.24	5.36	3.42	1.26
	.99	7.06	8.70	9.85	4.33	1.54
	M	793	959	994	1088	927
n=100	.90	3.09	3.69	3.44	2.04	0.78
	.95	4.43	5.07	4.67	2.75	1.06
	.975	5.57	6.90	6.23	3.62	1.32
	.99	7.39	9.57	8.18	4.57	1.72
	M	904	991	1039	954	969

TABLE 7.2: Observed significance points of score test of fit W_u
(M = numbers of samples on which points are based)

conversely may show that evidence of failure to fit is spread throughout the entire data.

7.3 Influence

Theoretical results about the influence curves of maximum likelihood estimators for the generalized Pareto distribution were given in Section 6.2.3. Here the more practical matter of assessing the influence, or leverage, that an individual element of a given set of data exerts upon the parameter estimates based on that sample, is discussed. The treatment follows that of Cook and Weisberg (1982, Chapter 5). The $\{Y_i\}_{i=1}^n$ are regarded throughout as independent generalized Pareto variables whose unknown parameters may depend on the values of known covariates.

Suppose that the parameter estimate θ is found by maximum likelihood based on the entire sample. Then one obvious way of investigating the dependence of $\hat{\theta}$ on individuals y_i is to drop them from the sample and to re-estimate $\hat{\theta}$, thus obtaining n estimates $\hat{\theta}_{(i)}$ each based on $n-1$ observations. However this has the drawback that $n+1$ maximizations are required, potentially an expensive procedure. Cook and Weisberg suggest replacing the $\hat{\theta}_{(i)}$ by one-step estimates $\hat{\theta}_{(i)}^1$ arising from a quadratic approximation to the loglikelihood $\ell_{(i)} = \sum_{j \neq i} \ell_j$ of the data without its i^{th} datum. Thus if

$$\ell_{(i)}(\theta) \sim \ell_{(i)}(\hat{\theta}) + (\theta - \hat{\theta})^T \ell'_{(i)}(\hat{\theta}) + \frac{1}{2} (\theta - \hat{\theta})^T \ell''_{(i)}(\hat{\theta}) (\theta - \hat{\theta}),$$

then $\ell_{(i)}(\theta)$ is approximately maximized at

$$\hat{\theta}_{(i)}^1 = \hat{\theta} - (\ell''_{(i)}(\hat{\theta}))^{-1} \ell'_{(i)}(\hat{\theta}),$$

where $\ell'_{(i)}(\hat{\theta})$ and $\ell''_{(i)}(\hat{\theta})$ are respectively the score statistic and observed information for the reduced dataset, at the overall maximum likelihood estimate $\hat{\theta}$. Note that $\ell'_{(i)}(\hat{\theta}) = -\ell'_i(\hat{\theta})$ to see that

$$\hat{\theta}_{(i)}^1 = \hat{\theta} + (\ell''_{(i)}(\hat{\theta}))^{-1} \ell'_{(i)}(\hat{\theta}).$$

The one-step estimates $\hat{\theta}_{(i)}^1$ may be expected to be close to the fully iterated estimates $\hat{\theta}_{(i)}$ provided the loglikelihood is locally quite quadratic. The results of Section 6.3.2 suggest that this is not the case, but some experimentation indicates that the non-quadraticity makes no serious difference to the one-step estimates, which anyway serve only as an approximation designed to give a qualitative idea of the effect of jackknifing the estimates. Sometimes the one-step estimates for the generalized Pareto distribution lie outside the range of allowable parameter values, and then the full iterated estimates must be found. Cook and Weisberg suggest that an overall idea of the effect on the estimates of leaving the i^{th} observation out may be found by considering the likelihood distances

$$\ell d_{(i)} = 2\{ \ell(\hat{\theta}) - \ell(\hat{\theta}_{(i)}^1) \},$$

which should be large if $\hat{\theta}_{(i)}^1$ is far from $\hat{\theta}$, measured in terms of the loglikelihood ℓ of the whole data.

Table 7.3 shows an example of these ideas at work on a set of MESOS exposures to air contamination 800 km north of Mol, due to notional releases of one Curie of radiation every three hours through 1976. The entire sample is plotted in Figure 1.3 : 187 values in all. The threshold was taken to be $0.0011 \mu\text{Cism}^{-3}$ and the resulting set of 28 excesses multiplied by 10^{10} . No declustering was employed: the remaining exposures comprise more than one-half of the total experienced at the receptor in 1976. Both the clustering of the extremes and the huge variation in exposure levels are evident from the table and figure.

The maximum likelihood estimates of the parameters are $\hat{k} = -0.296$ and $\hat{\sigma} = 9.23$, with estimated standard errors 0.245 and 2.809 and

time(hrs) +3	excess Cism ⁻³ x 10 ¹⁰	$\hat{\sigma}_{(i)}$	$\hat{k}_{(i)}$	$\hat{d}_{(i)}$
105	19.93	8.55	-.33	.07
294	1.10	9.85	-.27	.05
295	1.37	9.82	-.27	.04
504	4.40	9.49	-.29	.01
505	8.65	9.70	-.32	.01
510	10.82	8.95	-.32	.02
512	87.81	9.71	-.05	2.71
604	.12	9.97	-.26	.07
606	1.66	9.79	-.27	.04
611	10.76	8.96	-.32	.02
612	7.30	9.21	-.31	.01
613	11.74	8.90	-.33	.02
874	19.48	8.56	-.33	.06
1269	1.97	9.75	-.28	.03
1270	2.20	9.73	-.28	.03
2192	2.29	9.71	-.28	.03
2228	1.39	9.82	-.27	.04
2430	16.49	8.66	-.33	.05
2504	2.88	9.65	-.28	.02
2531	8.05	9.15	-.31	.01
2532	8.00	9.16	-.31	.01
2533	7.28	9.22	-.31	.01
2534	10.49	8.97	-.32	.01
2535	49.70	8.48	-.25	.33
2536	30.65	8.38	-.32	.12
2537	8.01	9.16	-.31	.01
2539	23.89	8.46	-.33	.09
2540	6.35	9.30	-.31	.01

TABLE 7.3: MESOS exposure dataset and influence diagnostics: time-integrated air contamination through 1976 800 km north of Mol due to unit releases of Kr₈₅ during every three-hour period; exceedances of 1.1×10^{-10} Cism⁻³.

covariance -0.427 based on expected information. The likelihood ratio test for $k=0$ is 2.72 as χ^2_1 , just over the 10% level; the score test for lack of fit at 0.14 is not especially notable. The table shows the jackknifed values of $\hat{\sigma}_{(i)}$ and $\hat{k}_{(i)}$ and the $ld_{(i)}$. The only large $ld_{(i)}$ is for $i=7$, corresponding to the largest sample value, for which $\hat{k}_{(i)} \sim -0.05$. Incidentally, this was the only observation for which the fully iterated parameter estimates had to be found. Figure 7.2 confirms the implication of the table, that the parameter estimates are sensitive only to the loss of the largest datum, which contains virtually all the evidence for $k < 0$. The value of $\hat{\theta}$ is close to the centre of the other $\hat{\theta}^1_{(i)}$. If there were prior suspicion that y_7 arose from a recording error or an instrument failure a decision to exclude it from further analysis might be made, but as it is the statistician just has to live with the situation. This example confirms at a data-analytical level the implications of the theory in Section 6.2.3: parameter estimation in extreme-value contexts may depend critically on the few highest values. Given the nature of the problem, any other conclusion would be a surprise.

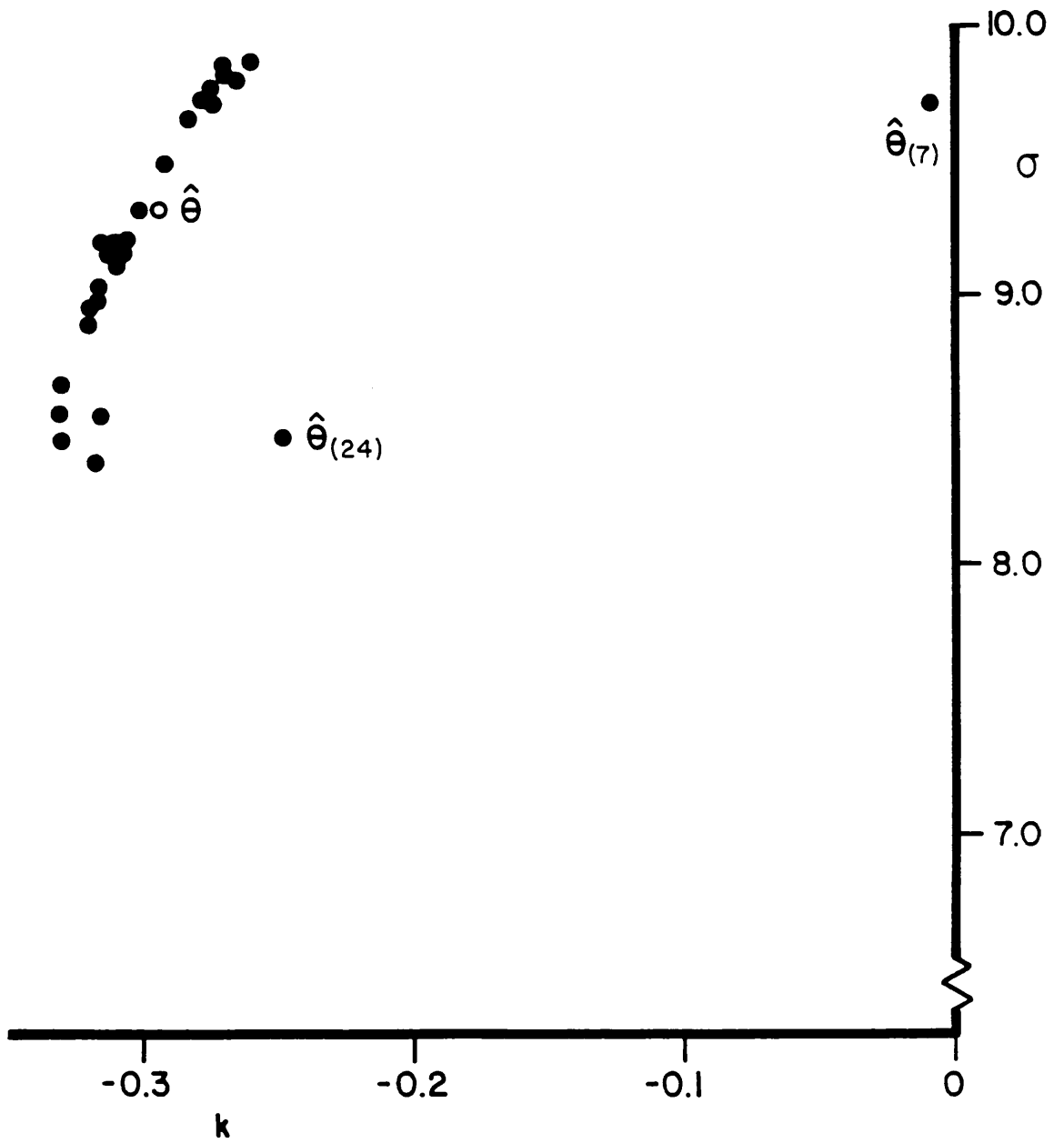


Figure 7.2 : Jack-knifed maximum likelihood estimates for data in Table 7.3.

8. HIGH EXPOSURE EPISODES

8.1 Introduction

Accurate prediction of high levels of exposure to radioactivity as a result of potential releases from nuclear installations is important because high contamination levels pose the gravest direct risks to the individual and population. Another reason for laying particular stress on modelling the upper tails of the MESOS exposure data is that the distributions are so skewed that the upper few observations of each of them contain a substantial proportion of the total annual exposure - see Table 3.9. Figure 1.3 shows that these high values tend to occur in clusters, an important aspect of modelling extremes. In this chapter the statistical methods developed in the previous three chapters are used to analyse the MESOS data.

This analysis relies upon the generalized Pareto distribution as an approximation to the distribution of exceedances of the MESOS data over high thresholds. The choice of threshold is clearly an important practical and theoretical issue: the tension between the need for a high threshold so that the generalized Pareto approximation is adequate and the need to retain sufficient data for analysis has already been mentioned. Here theoretical considerations are bypassed: the thresholds are chosen empirically by regression of the upper percentiles of the MESOS exposure distributions on covariates. That the resulting thresholds are adequately high and retain enough data for analysis is verified below in Section 8.3.

If inference is to be made for given data, a threshold level could be determined by minimizing some function of goodness of fit of the generalized Pareto distribution to the data tail and the number of data which exceed the threshold - the mean squared prediction error for some high return value, say. However an integral part of

the present problem is the prediction of extremes for distributions with as yet unknown covariates. This suggests that thresholds be made to depend on the covariates. One crude way to do this is as follows.

Arguments in Section 3.3 suggest that the scale of an exposure distribution should depend on nuclide parameters λ , v_d , and λ_w , source-receptor distance d , and release duration t hours in such a way that it is a linear function of $d\lambda$, dv_d , $d\lambda_w$, $\log\{d\}$, and $\log\{t/3\}$. This in turn suggests fitting to the data the model

$$\log\{y_{[np]}\} = \xi_0 + \xi_1 d\lambda + \xi_2 dv_d + \xi_3 d\lambda_w + \xi_4 \log\{d\} + \xi_5 \log\{t/3\} + \varepsilon,$$

where $y_{[np]}$ is the sample order statistic corresponding to the $p \times 100\%$ point of the exposure distribution, the ε 's are uncorrelated identically distributed errors, and the ξ 's are unknown parameters to be estimated. Fits of this equation to the exposure distributions - using GLIM - for different values of p in the range 0.9-0.99 reveal that the values of the estimates $\hat{\xi}_1 - \hat{\xi}_5$ are quite stable, and that $\hat{\xi}_0$ increases with p . Treating as fixed the estimated parameters for $p=0.9$, the equation

$$\exp\left\{ \hat{\xi}_0 + 9.463 - 0.1678d\lambda - 0.3497 \times 10^{-3} dv_d - 0.5473 \times 10^{-2} d\lambda_w - 1.182 \log\{d\} - 0.6913 \log\{t/3\} \right\} \quad \dots 8.1.1$$

is henceforth adopted to give thresholds in $\mu\text{Ci m}^{-3}$ for time-integrated air concentration distributions for exposures at a receptor d metres from a source at which a unit release over a period of duration t hours has occurred. Corresponding thresholds in $\mu\text{Ci m}^{-2}$ for dry deposition distributions are obtained simply by multiplying air concentration thresholds by deposition velocity v_d ; the equation for exposures to wet deposition in $\mu\text{Ci m}^{-2}$ is

$$\lambda_w \exp\{ \xi_0 + 11.49 - 0.0901d\lambda - 0.9692 \times 10^{-2} d\lambda_w - 0.9793 \log\{d\} - 0.6645 \log\{t/3\} \} \quad \dots \quad 8.1.2.$$

In both 8.1.1 and 8.1.2, ξ_0 - loosely called the threshold - may take values from 0 to about 2 as p rises from 0.9 to 0.99.

Equations 8.1.1 and 8.1.2 are now regarded as fixed - up to the choice of ξ_0 - and determine thresholds below which exposures are regarded as irrelevant to this analysis of extremes. Clearly it is important that any model proposed for extremes is not critically dependent on the exact value of ξ_0 chosen; hence the data are analysed separately for different values of ξ_0 .

Throughout Chapter 3 emphasis was laid upon prediction of marginal distributions of exposures: variation of exposure levels with time was ignored. Although it is hard to sustain the belief that there is no seasonal variation in exposure levels, it is small compared with the much greater variation due to diurnal changes in the boundary layer. Clustering of exposures due to persistence of weather conditions has also been largely ignored, but it cannot be brushed aside if doing so may lead to serious overestimation of the probability of high values during an arbitrary period because their propensity to occur together has not been taken into account. Thus there is a need for a model for clustering of extremes.

Clustering aspects of the extremes of the MESOS data are studied in Section 8.2, and exceedances themselves analyzed in Section 8.3; the proposed model is tested in Section 8.4.

8.2 A clustering model for extremes

It is important to distinguish between two questions which may be asked of a model for high exposures at a given receptor due to releases of a given nuclide from a known source. The first is this: what is the distribution of extreme exposures due to a single unit release of some known duration t hours which starts at an arbitrary time? This must be answered to satisfy the Euratom Treaty Article 37 requirements. The second is this: what is the distribution of the maximum exposure due to a number of unit releases over successive periods each of duration t hours? This is harder to answer because of the need to take into account clustering of high exposures. The aim of this section is to provide a clustering model for extremes which allows the second question to be addressed: it is then easy to answer the first.

Here the aim of attacking clustering is to enable reasonably good prediction of return values, not to study the small-scale structure in time of extreme episodes. Thus for the present purpose the essential ingredients of a suitable clustering model are: firstly, a mechanism which generates cluster centres; secondly, a mechanism which determines the number of points in a cluster; and thirdly, a mechanism to determine the size of the exposure attached to each point. A cluster is a group - defined in a way below - of exceedances over a threshold. The size, or number of excesses in a cluster is a random positive integer. The excesses in a cluster are regarded as distinct but as all occurring simultaneously at the cluster centre, an approximation which serves the current need.

Another approximation is introduced by thinking of exposure taking place in continuous time, whereas in fact they occur in discrete time. This results in considerable simplification, and the bias thus generated is small unless exposures take place rather

frequently - which is not the case here. The bias is biggest for releases of duration one week, but is not then serious.

The unambiguous definition of a cluster of exposures at a single receptor due to releases of any duration t hours must precede the formulation of a clustering model. A cluster of excesses over a given threshold is said here to begin with a single exceedance and end with a period of one day or more during which none occur - so by definition exposures due to releases of duration more than one day do not cluster. This is crude but not unreasonable: clusters so defined may not overlap, which if they are interpreted as corresponding to meteorological episodes is a sensible requirement which also leads to a simple statistical analysis. The rationale for this definition or one like it is that a day is a natural meteorological unit of time - due to diurnal variation of the mixing layer - and so it is plausible that if there is a gap of one day or more in high levels of exposure, they have arisen as a result of different dispersion episodes. This does not preclude the possibility that a single episode may last several days, although in the MESOS data they rarely last more than two days. The choice of a one day gap is not crucial - defining a cluster as ending with no excesses in any period between twelve hours and a few days makes little difference to the sizes of the clusters.

Leadbetter, Lindgren, and Rootzèn (1983) prove that under weak conditions on a stationary stochastic process, the process of its exceedances over a high threshold should converge to a homogeneous Poisson process as the threshold increases. This suggests that for releases of duration t hours the process of cluster centres - which correspond to meteorological situations leading to high exposures at a given receptor - be modelled as a homogeneous Poisson process of rate ρ_t , say. That is, the numbers of clusters in non-overlapping time periods are independent of each other, and the number of

clusters in any period of length T has the Poisson distribution with mean Tp_t . The i^{th} cluster centre has attached to it a cluster of exposures $\{Y_{ij}\}_{j=1}^{M_i}$, where M_i is the (random) cluster size.

The next ingredient is a recipe to determine cluster sizes - which are necessarily positive. Analysis of the MESOS data suggests that the geometric distribution is appropriate; that is

$$\text{Prob}(M_i = k) = (1-q)q^{k-1}$$

for some q in the range $(0,1)$. This distribution has probability generating function

$$G(u) = u \frac{(1-q)}{1-qu} \quad \dots \quad 8.2.1;$$

its mean is $1/(1-q)$, and its variance is $q/(1-q)^2$. The probability q changes with release duration and threshold ξ_0 but is assumed independent of receptor location, nuclide type, and source.

The final ingredient is the specification - for a given release duration and threshold - of the joint distribution of the exposures in the i^{th} cluster given its size. An appropriate model is this: given $M_i = m_i$, the values of the exceedances $\{Y_{ij}\}$ over the threshold are independent and identically distributed in some distribution $F(\cdot)$. Moreover $F(\cdot)$ does not depend on seasonal or other time-dependent factors. Under this model the distribution of the maximum exposure Y_T experienced at the receptor due to T successive t -hour releases at the source is

$$\text{Prob}(Y_T < y) = \exp\{ Tp_t [G(F(y)) - 1] \} \quad \dots \quad 8.2.2.$$

These assumptions are now tested. Results are reported for the analysis of time-integrated air concentration data at all 16 receptors for releases of Kr_{85} from Mol during 1976. Results for other sources and nuclides are similar.

Cox and Lewis(1966, Chapter 6) discuss methods of testing if a point process observed for a fixed length of time T is a homogeneous Poisson process. The Anderson-Darling statistic with $U_{(i)} = (t_i - t_0)/T$, where t_0 is the time at which the process began to be observed and the t_i are the times of successive events, provides a powerful test for departures from Poissonness. This test applied to the data with $\xi_0=0$ yields two out of sixteen values significant at the 5% but not the 1% level, weak evidence of departures from Poissonness; however for $\xi_0=0.5$ or more there is no such evidence. The Poissonness of clusters is not in doubt.

Table 8.1 shows for $\xi_0=0$ and releases of duration three hours the observed distribution of cluster sizes compared with their expected distribution based on assuming them geometrically distributed. The maximum likelihood estimate of q is then 0.49, and the total number of clusters is 194. The observed distribution is a little overdispersed relative to the expected one; but for higher values of ξ_0 and longer release durations the match is better, and there is no reason to question the geometricity of cluster sizes. The likelihood ratio statistic for the hypothesis of different values of q at different receptors against the null hypothesis of the same value of q at all receptors is 24.89, nominally χ^2_{15} . Just not significant at the 5% level, this casts doubt on the null hypothesis. However the effect is not big enough to be physically important, nor is it marked in other datasets: thus it is adequate to use a single value of q for given ξ_0 and release duration at all receptors. Values of q appropriate for exposures in all conditions are given in Table 8.2(a).

cluster size :	1	2	3	4	5	6	>6
observed :	115	37	16	11	7	3	5
expected :	98.3	48.5	23.9	11.8	5.8	2.9	2.8

Table 8.1 : Comparison of observed and expected cluster sizes for exposures to air contamination due to three-hour releases of Kr₈₅ from Mol in 1976; threshold $\xi_0=0$

The assumptions about the m_i exceedances Y_{ij} ($j=1, \dots, m_i$) in the i th cluster may be checked without assumptions about the form of $F(\cdot)$, as follows. The first assumption is that the mean values of the Y_{ij} do not depend on the cluster index i . Denote the overall mean of the exceedances Y_{ij} by $Y_{..}$ and the cluster means by $Y_{i..}$. Then if the true cluster means are different, the statistic

$$A = \frac{\sum_i (Y_{i..} - Y_{..})^2}{\sum \sum_{ij} (Y_{ij} - Y_{i..})^2}$$

should be larger than would be expected otherwise. An approximation to the full randomization distribution of A under the null hypothesis of no variation of the mean values of the Y_{ij} between clusters is obtained by randomly permuting them - say 99 times - and calculating the value A_{perm} of A corresponding to each permutation, then finding the significance level of A_{obs} , the observed value of A . That is,

$$p_{\text{obs}} = \{ \#A_{\text{perm}} < A_{\text{obs}} \} / 100$$

is an approximation to the actual significance level of A_{obs} on the full permutation distribution of the Y_{ij} under the null hypothesis. Combination of the values of the p_{obs} for all 16 different receptors and $\xi_0=0$ by noting that $-2\log\{p_{\text{obs}}\}$ has the χ^2_2 distribution on the null hypothesis leads to the test statistic 30.45, distributed as χ^2_{32} : there is no evidence against the hypothesis that the cluster means are the same. This is confirmed by analysis of other datasets.

ξ_0	release duration			
	3 hours	6 hours	12 hours	1 day
0	0.5	0.35	0.25	0.10
0.5	0.45	0.30	0.20	0.05
1.0	0.40	0.25	0.15	0.0

(a) Exposures to air contamination

ξ_0	release duration			
	3 hours	6 hours	12 hours	1 day
0	0.35	0.20	0.10	0.05
0.5	0.25	0.10	0.05	0.02
1.0	0.15	0.0	0.0	0.0

(b) Exposures to wet deposition

TABLE 8.2: Values of clustering probability q for different thresholds and release durations.

The assumption of no correlation among the exceedances Y_{ij} in a single cluster may be checked in a similar way. The subscript j now indicates the order in time of the excesses within a cluster. If $m_i > 1$, then estimates $\hat{\gamma}_k$ of the lag k ($k < m_i$) correlation coefficients γ_k of the Y_{ij} may be obtained from the Y_{ij} . A statistic which should be relatively larger if there are indeed non-zero correlations among the excesses in a cluster is

$$B = \sum_i m_i \sum_k \hat{\gamma}_k^2.$$

Once again an approximation to the full randomization distribution of B under the null hypothesis of no autocorrelation among the (identically distributed) Y_{ij} can be found by permuting the Y_{ij} . The combined test statistic for all 16 receptors and $\xi_0=0$ is 33.16 as χ^2_{32} , giving no evidence against the null hypothesis. Excesses in the same cluster may reasonably be held to be uncorrelated: for the purpose of this analysis I make the stronger assumption that they are independent. This is appropriate to the simple model being built here, but would merit closer examination in a formulation aimed primarily at modelling the dynamics of pollution episodes.

A similar analysis was performed for $I_{131}(p)$ wet deposition data for releases from Heysham through 1973. For $\xi_0=0$ there is again weak evidence for departure from Poissonness of the process of cluster centres, but the evidence vanishes as ξ_0 increases. Table 8.3 shows a comparison of observed and expected cluster sizes for $\xi_0=0$, based on the assumption that they are geometrically distributed. The fit is very good. The likelihood ratio statistic for different values of q at the different receptors - nominally χ^2_2 - is 8.29: there is no evidence of overdispersion. Values of q appropriate for different release durations and thresholds are given for exposures to wet deposition in Table 8.2(b).

Table 8.3 shows that most clusters have size one, so there is little information to be gained in using test statistics A and B to check assumptions about excesses between and within clusters. The Poisson clustering model posited above seems satisfactory for both wet and dry extremes.

cluster size :	1	2	3	4	5	6	>6
observed :	64	20	7	3	0	0	0
expected :	64.5	20.2	6.4	2.0	0.6	0.2	0.1

Table 8.3 : Comparison of observed and expected cluster sizes for exposures to wet deposition due to three-hour releases of $I_{131}(p)$ from Heysham during 1973; $\xi_0=0$

The Poisson process parameter ρ_t at a single receptor depends on release duration t hours and possibly on other factors. Important factors may be found by assuming that it is a function of covariates, - nuclide parameters, source-receptor distance, release duration, etc - and using GLIM to regress the observed numbers of clusters on the covariates using a Poisson likelihood. It emerges that nuclide parameters and source-receptor distance have little effect on the observed numbers of clusters and that appropriate prescriptions are

$$\rho_t = \exp\{ \alpha + 0.6\log\{p(\theta)\} + \log\{t/3\} \} \dots 8.2.3,$$

for exposures in all conditions, and

$$\rho_t = \exp\{ \alpha + 0.4\log\{p_w(\theta)\} + \log\{t/3\} \} \dots 8.2.4$$

for exposures in wet conditions, where $p(\theta)$ and $p_w(\theta)$ are the source windrose probabilities used in Chapter 2 and t is the release duration in hours. Since ρ_t is a linear function of t , the expected number of clusters of exposures during a given time-period at a receptor is fixed, regardless of the duration of the releases

considered for that time-period. Values of α depend on the source windrose type: for exposures to air contamination, sources in the Mediterranean area have higher values of α than those with more uniform windroses, but the opposite is true for exposures to wet deposition. Values of $\hat{\alpha}$ are given for several values of ξ_0 in Table 8.4. Values appropriate for exposures in all conditions are lower for sources with more uniform windroses: values of ρ_t for Heysham, Karlsruhe, and Mol are lower than those for Cadarache and Ispra by a factor of 0.6-0.3 depending on the value of ξ_0 . The high dependence of clustering parameter ρ_t on source windrose for exposures to time-integrated air concentrations shows that for all sources the incidence of high exposures is more closely related to the source windrose than the bulk of exposures. It is interesting that the dependence of ρ_t on source windrose is weaker for exposures to wet deposition - presumably because of local rainfall effects such as orographic rain. Model fit improves as ξ_0 increases. For lower values of ξ_0 the numbers of clusters seem overdispersed as judged by the deviances for their regressions, but since counts for different nuclides at the same receptor are not independent the deviances are artificially high.

Under this model the number of clusters of exposures which have at least one exceedance of a threshold ξ_0 experienced at a receptor due to T releases each of duration t hours has approximately the Poisson distribution - with mean $T\rho_t$ and variance also $T\rho_t$ - conditionally on the estimated value of α . It is easy to see from the standard errors of the parameter estimates $\hat{\alpha}$ in Table 8.4 that any extra variation introduced by use of the estimates α rather than their true values is tiny and can safely be ignored.

ξ_0	α_{Uniform}	$\alpha_{\text{Mediterranean}}$
0.0	-4.18 (0.027)	-3.74 (0.035)
0.5	-4.89 (0.038)	-4.23 (0.047)
1.0	-5.88 (0.063)	-5.11 (0.070)

(a) Exposures in all conditions

ξ_0	α_{Uniform}	$\alpha_{\text{Mediterranean}}$
0.0	-4.84 (0.054)	-5.44 (0.091)
0.5	-5.40 (0.071)	-6.07 (0.125)
1.0	-6.35 (0.115)	-6.93 (0.192)

(b) Exposures in wet conditions

Table 8.4: Dependence of Poisson rate ρ_c on threshold parameter ξ_0 estimate (standard error)

8.3 High exposure levels

The arguments of Chapters 4 and 5 suggest strongly that the generalized Pareto distribution

$$F(y) = \begin{cases} 1 - (1 - ky/\sigma)^{1/k} & (k \neq 0) \\ 1 - \exp(-y/\sigma) & (k = 0) \end{cases} \quad \dots 8.3.1.$$

is apt to describe the sizes of exposures which exceed high thresholds. Further, the previous section suggests that it will be adequate to treat high exposures at a single receptor due to releases of a given nuclide over a known release duration at a fixed source as independent and identically distributed. However the unknown scale parameter σ - and possibly also k - will probably depend upon source-receptor distance, nuclide characteristics, and release duration. This section elucidates this dependence for exposures in all and in wet conditions.

To recap, if the shape parameter k of the generalized Pareto distribution is zero, the distribution is exponential; if k is

positive, the distribution has an upper terminal beyond which no values can occur; and increasingly negative values of k correspond to power-law tails with no upper terminal and increasing weight.

The data used to study extremes in all conditions consist of 3524 observations exceeding the thresholds defined at equation 8.1.1 with $\xi_0=0$, and in wet conditions of 2390 observations exceeding the thresholds defined at equation 8.1.2 with $\xi_0=0$. These datasets were split into equal halves and analysis only performed for one half, in order to cross-validate the models.

The argument which began Section 3.3 suggests that a suitable form for the dependence of σ on covariates is

$$\sigma = \exp\{ \beta_0 + \beta_1 d \lambda + \beta_2 d v_d + \beta_3 d \lambda_w + \beta_4 \log(d) + \beta_5 \log(t/3) \} \dots 8.3.2.$$

Here d is the source-receptor distance in metres; λ (s^{-1}), v_d (ms^{-1}) and λ_w (s^{-1}) are nuclide decay constant, deposition velocity and washout coefficients respectively; and t is release duration in hours. The parameters β are to be estimated from the data and may be interpreted in terms of physical quantities related to pollutant transport in conditions leading to high exposures.

Table 8.5(a) shows for exposures in all conditions the successive reductions in model deviance - twice the minus loglikelihood of the data - due to fitting to the data by maximum likelihood the covariates in the order given in the table. The reduction achieved by fitting each of them decreases as ξ_0 increases. The effects of release duration, lateral broadening, half-life, and washout coefficient are large at all levels ξ_0 .

The corresponding parameter estimates and their standard errors based on the observed information matrix are displayed in Table 8.6. The parameter estimates are similar for all values of ξ_0 ; however the effect of release duration on exposure is more marked, and the tail

Reduction in Deviance

Covariate	$\xi_0 = 0$	$\xi_0 = 0.5$	$\xi_0 = 1.0$
$\log\{t/3\}$	142.0	44.4	22.4
$\log\{d\}$	1438.0	769.2	320.0
$d\lambda$	432.0	274.6	158.0
$d\lambda_{\omega}$	66.0	21.6	6.4
dv_d	16.0	5.8	0.8
# observations	1762	751	280

(a) Exposures to air contamination

Covariate	$\xi_0 = 0$	$\xi_0 = 0.5$	$\xi_0 = 1.0$
$\log\{t/3\}$	109.0	110.4	49.2
$\log\{d\}$	528.0	221.8	92.8
$d\lambda$	18.0	8.4	15.2
$d\lambda_{\omega}$	6.0	11.0	10.0
dv_d	2.0	0.2	0.4
# observations	1195	622	279

(b) Exposures to wet deposition

Table 8.5 Reductions in deviance due to introduction of successive covariates for different threshold levels ξ_0 .

Threshold Level

Parameter	$\xi_0 = 0$		$\xi_0 = 0.5$		$\xi_0 = 1.0$	
	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.
β_0	23.441	0.502	24.987	0.748	25.050	1.177
β_1	-0.135	4.933×10^{-3}	-0.137	6.671×10^{-3}	-0.143	9.464×10^{-3}
β_2	-2.304×10^{-4}	5.510×10^{-5}	-1.799×10^{-4}	7.676×10^{-5}	-8.863×10^{-5}	1.196×10^{-4}
β_3	-9.893×10^{-3}	3.144×10^{-3}	-8.691×10^{-3}	4.583×10^{-3}	-1.112×10^{-2}	7.323×10^{-3}
β_4	-1.227	3.999×10^{-2}	-1.316	5.986×10^{-2}	-1.294	9.423×10^{-2}
β_5	-0.720	3.704×10^{-2}	-0.840	4.901×10^{-2}	-1.065	7.850×10^{-2}
k	-0.233	3.101×10^{-2}	-0.121	4.089×10^{-2}	-5.227×10^{-2}	6.289×10^{-2}

Table 8.6 Parameter estimates for different thresholds for exposures to time-integrated air concentrations (μCism^{-3}).

of the data apparently becomes more nearly exponential ($k \rightarrow 0$), as ξ_0 increases. It is noteworthy that the value of $\hat{\beta}_2$ decreases at high levels: as ξ_0 increases dry deposition is increasingly less important. The estimates suggest that mean puff speed $u \approx 7 \text{ ms}^{-1}$; that the mean height of the puff over its travel time is $h \approx 600 \text{ m}$; and that the proportion of travel time over which rain is experienced is about 0.07. The values of $\hat{\beta}_4$ suggest that the relative effect of plume broadening with distance is greater for high than for all levels of exposure.

The partially maximized loglikelihood $\hat{\ell}_{\max}(k)$ is shown for $\xi_0 = 0$ in Figure 8.1. It is close to being quadratic about its maximum and yields $(-0.27, -0.19)$ as an approximate 95% confidence interval for k , with those based on the inverse observed and expected information matrices $(-0.29, -0.17)$ and $(-0.29, -0.18)$ respectively. For $\xi_0 = 0$ the hypothesis $k = 0$ is quite untenable, although the excesses are closer to exponentiality for higher thresholds.

It might be thought that different dispersion conditions might apply to releases over periods of twelve hours or less and those over periods of one day or more - which should be affected to a greater extent by the diurnal mixing cycle. The likelihood ratio test for different values of k in these two sets of release durations - but a common value of β - gives evidence for this for $\xi_0 = 0$ at between the 5% and 1% levels. The value of \hat{k} for exposures due to releases of duration twelve hours or less is -0.25 , but that for exposures due to releases of longer duration is -0.08 . This confirms the idea that exposures due to releases of duration a day or more are lighter-tailed than those due to shorter releases.

When the hypothesis that different values of k apply to different distances from the source, a similar effect is found for $\xi_0 = 0$. The likelihood ratio test for the hypothesis that the value of

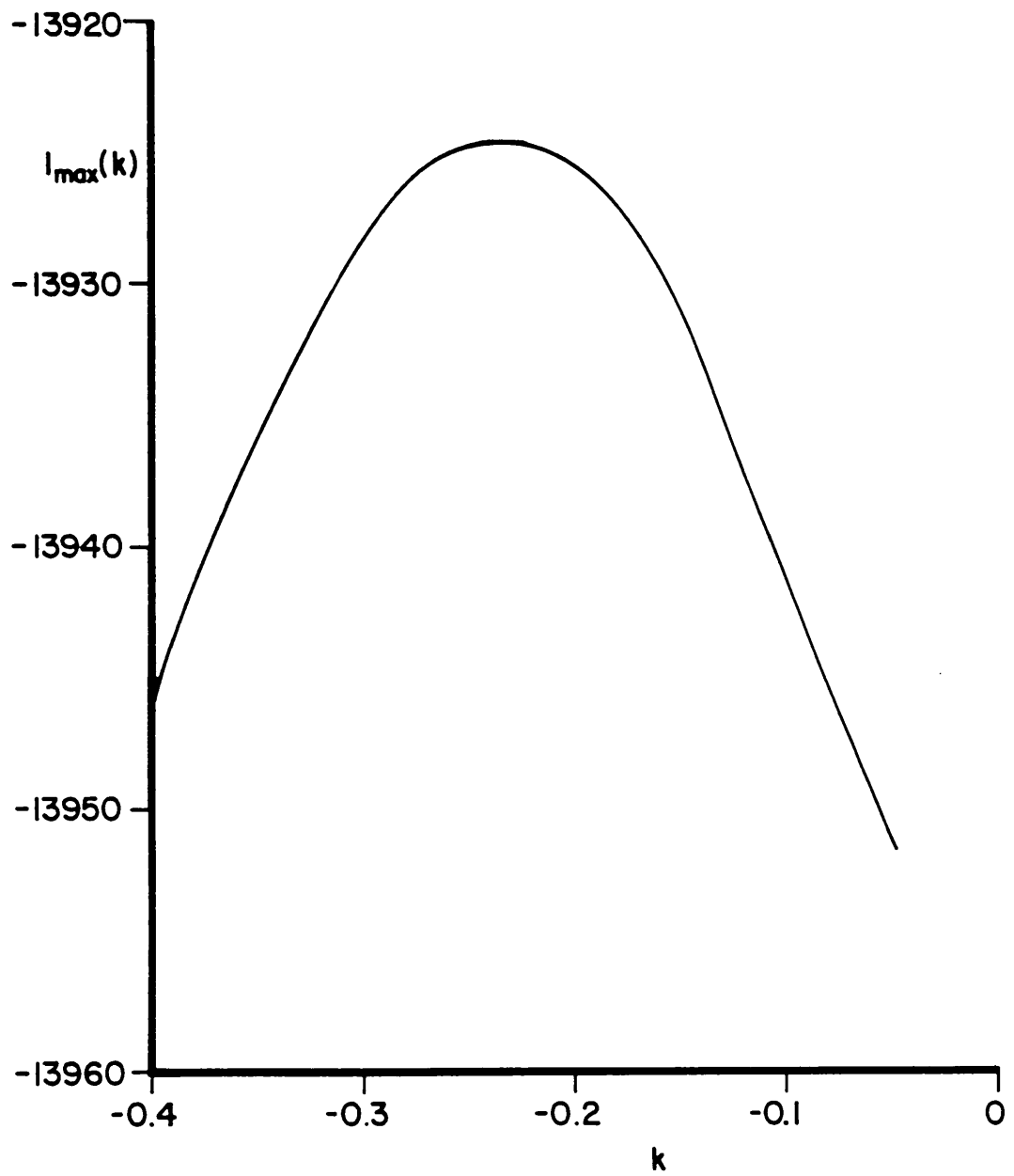
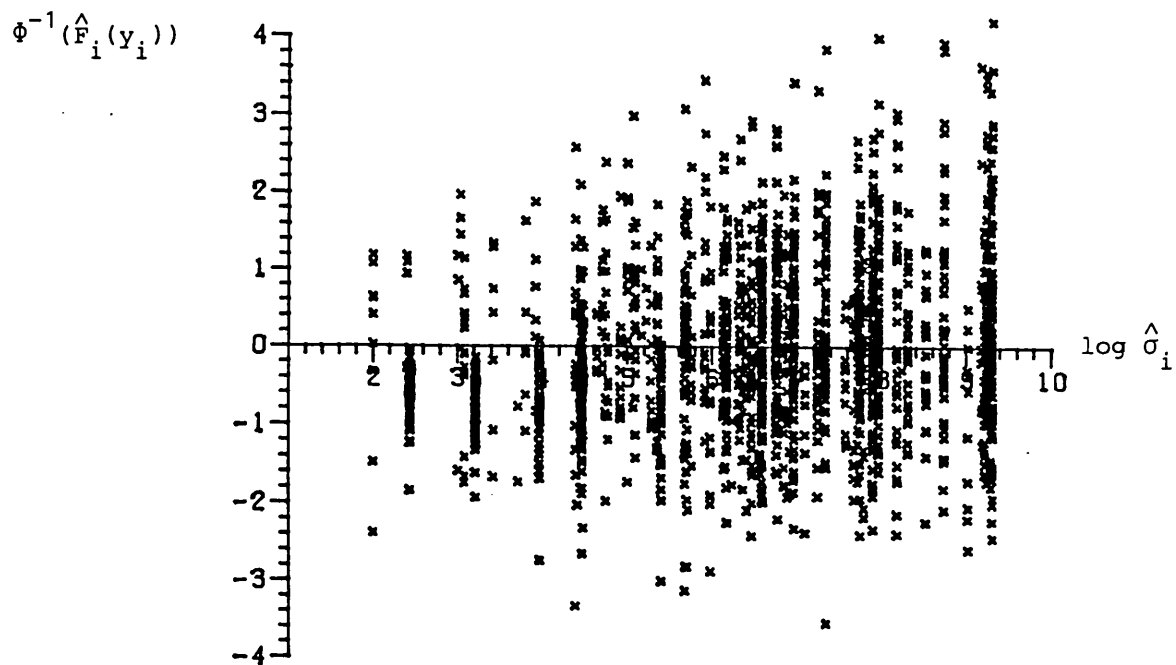


Figure 8.1 : Partially maximized loglikelihood $l_{\max}(k)$ for high exposures to air contamination; threshold $\xi_0=0$.

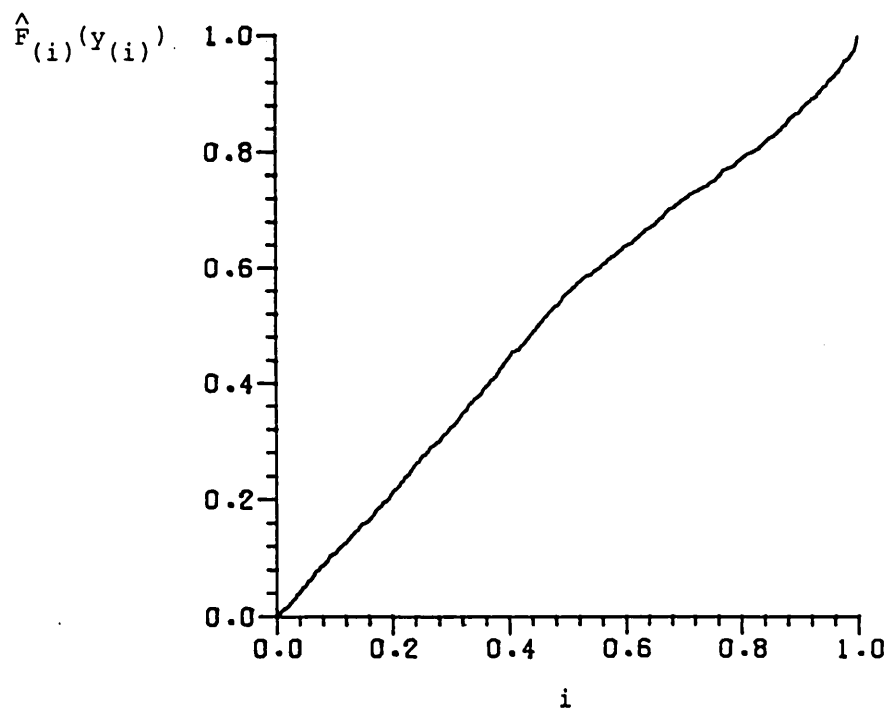
k at receptors 100 kilometres from their sources is different to that at greater distances is significant at between the 5% and 1% levels. This lends support to the notion that dispersion mechanisms which affect high exposure levels at distances of more than 100 km have less effect on those at 100 km or less, with the consequence that the tail behaviour of extreme exposures is different over the two ranges of source-receptor distances. The value of \hat{k} at the longer distances is -0.19, whereas that at receptors 100 km from their sources is -0.36: exposures have heavier tails closer to their sources.

The value of Wilks' statistic for a test of the hypothesis that the values of k differ for depositing and non-depositing isotopes is 0.04. This gives no evidence in support of the hypothesis.

The score statistics for goodness of fit are 4.90, 2.12, and 1.73 for $\xi_0=0, 0.5,$ and 1.0 respectively. None are significant at their nominal 5% level, although the first is significant at the χ^2_2 10% level, possibly as a result of the heterogeneity of the values of k mentioned above. The plot of Normal Cox-Snell residuals when $\xi_0=0$ - that is, $\Phi^{-1}[\hat{F}_1(Y_1)]$, with $\Phi(\cdot)$ the standard Normal distribution function and $\hat{F}_1(\cdot)$ the estimated generalized Pareto distribution of the excess Y_1 - against the linear predictor $\log\{\hat{\sigma}_1\}$ in Figure 8.2(a) shows no outlying values, although there seems to be more variation for big values of the linear predictor than for smaller ones - as a result of the variation of k . The plot in Figure 8.2(b) of the empirical distribution function of the $\hat{F}_1(Y_1)$ - which should be roughly uniform - shows the same effect: rather too many high values of the $\hat{F}_1(Y_1)$, due to their overdispersion. However an Anderson-Darling test using the parameter estimates in Table 8.6 applied to the half of the data previously retained for cross-validation yields for $\xi_0=0$ the very low test statistic value 0.084; not nearly significant even at the 15% level. Calculation of the



(a) Normal Cox-Snell residuals $\Phi^{-1}(\hat{F}_i(y_i))$ vs. $\log \hat{\sigma}_i$



(b) empirical distribution function of uniform residuals $\hat{F}_i(y_i)$

Figure 8.2 : Residual plots for model for high exposures to air contamination; threshold $\xi_0=0$.

likelihood differences based on one-step jackknifed parameter estimates reveals no single particularly influential observations. This seems reasonable: the sample size is 1762 and the covariate matrix is quite well balanced. Similar calculations show better model fit at higher thresholds; in particular the evidence for heterogeneity of the values of k disappears.

The conclusion to be drawn is that there is good evidence for some heterogeneity of values of k for exposures due to releases of different durations and at different distances from their sources, but model fit for air contamination - and by implication dry deposition - data is adequate for the present purpose, especially if the likely sizes of inaccuracies in MESOS exposure levels are recalled to mind. The high values of dry deposition distributions - in $\text{pCi}\cdot\text{m}^{-2}$ - may be predicted by multiplying the predicted scale parameter σ of the nuclide time-integrated air concentration distribution by the deposition velocity v_d appropriate to the radionuclide in question.

A form for σ suitable for exposures to wet deposition is

$$\sigma = \lambda_w \exp\{ \beta_0 + \beta_1 d\lambda + \beta_2 d v_d + \beta_3 d \lambda_w + \beta_4 \log\{d\} + \beta_5 \log\{t/3\} \} \dots 8.3.3.$$

Successive reductions in model deviance due to fitting the covariates in order are given for exposures to wet deposition in Table 8.5(b). The reductions due to fitting the effects of release duration, source-receptor distance, and half-life are biggest at all threshold levels. The reduction due to fitting the effect of deposition velocity is not statistically significant.

Parameter estimates and their standard errors based on the observed information matrix are displayed in Table 8.7. The values of all the parameter estimates are less stable for wet exposures than for air contamination. Although pseudo-nuclides were generated to

Threshold Level

Parameter	$\xi_0 = 0$		$\xi_0 = 0.5$		$\xi_0 = 1.0$	
	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.
β_0	23.238	1.088	21.845	1.580	18.971	2.534
β_1	-7.214×10^{-2}	2.532×10^{-2}	-5.514×10^{-2}	3.615×10^{-2}	-0.133	5.045×10^{-2}
β_2	-8.250×10^{-5}	6.642×10^{-5}	-4.084×10^{-5}	9.445×10^{-5}	-9.544×10^{-5}	1.488×10^{-4}
β_3	-1.184×10^{-2}	3.906×10^{-3}	-1.919×10^{-2}	5.884×10^{-3}	-2.743×10^{-2}	8.684×10^{-3}
β_4	-0.822	9.279×10^{-2}	-0.688	0.135	-0.413	0.216
β_5	-0.647	3.553×10^{-2}	-0.703	4.780×10^{-2}	-0.784	7.817×10^{-2}
k	-0.397	3.906×10^{-2}	-0.460	5.854×10^{-2}	-0.505	9.008×10^{-2}

Table 8.7 Parameter estimates for different threshold levels for exposures to wet deposition (pCim^{-2}).

assess the effect of decay constant λ on high exposures, $\hat{\beta}_1$ is not very accurately determined, but it has the right sign and roughly the right size. The same is true for $\hat{\beta}_2$. The values of $\hat{\beta}_3$ are reasonably consistent for different threshold levels: they are bigger in absolute value than those for air contamination exposures. Values of $\hat{\beta}_4$ - representing the effect of plume broadening with distance - increase as ξ_0 increases. This suggests that the effect of plume broadening is increasingly weaker for higher levels of ξ_0 , though why this should be so is not clear. On the other hand the value of $\hat{\beta}_5$ decreases as ξ_0 increases, indicating a slightly bigger effect of release duration for exposures over higher thresholds. The effect of release duration on exposures to wet deposition is smaller than on exposures to air contamination. Finally, the apparent weight of the distribution tail - measured by the value of \hat{k} - increases with ξ_0 , unlike its behaviour for exposures to air contamination. This very different type of tail behaviour may be due to the fact that rainfall rate varies for different rainfall episodes. Such variation tends to overdispersion exposures to wet deposition relative to those to air contamination, but it is not clear why such overdispersion should increase with ξ_0 .

In view of the indeterminacy of $\hat{\beta}_1$ and $\hat{\beta}_2$, it is possibly misleading to interpret the values of the $\hat{\beta}$ for exposures to wet deposition in terms of mean puff height during its travel-time, its mean speed, and so on.

Figure 8.3 shows the partially maximized loglikelihood $\hat{\ell}_{\max}(k)$ for $\xi_0=0$ for exposures to wet deposition. Its shape is rather different to Figure 8.1 insofar as Figure 8.3 is quite non-quadratic even near its maximum; moreover despite the large sample size there is clearly less information about the value of k . The 95% confidence interval for k based on Figure 8.3 is roughly $(-0.51, -0.31)$, but

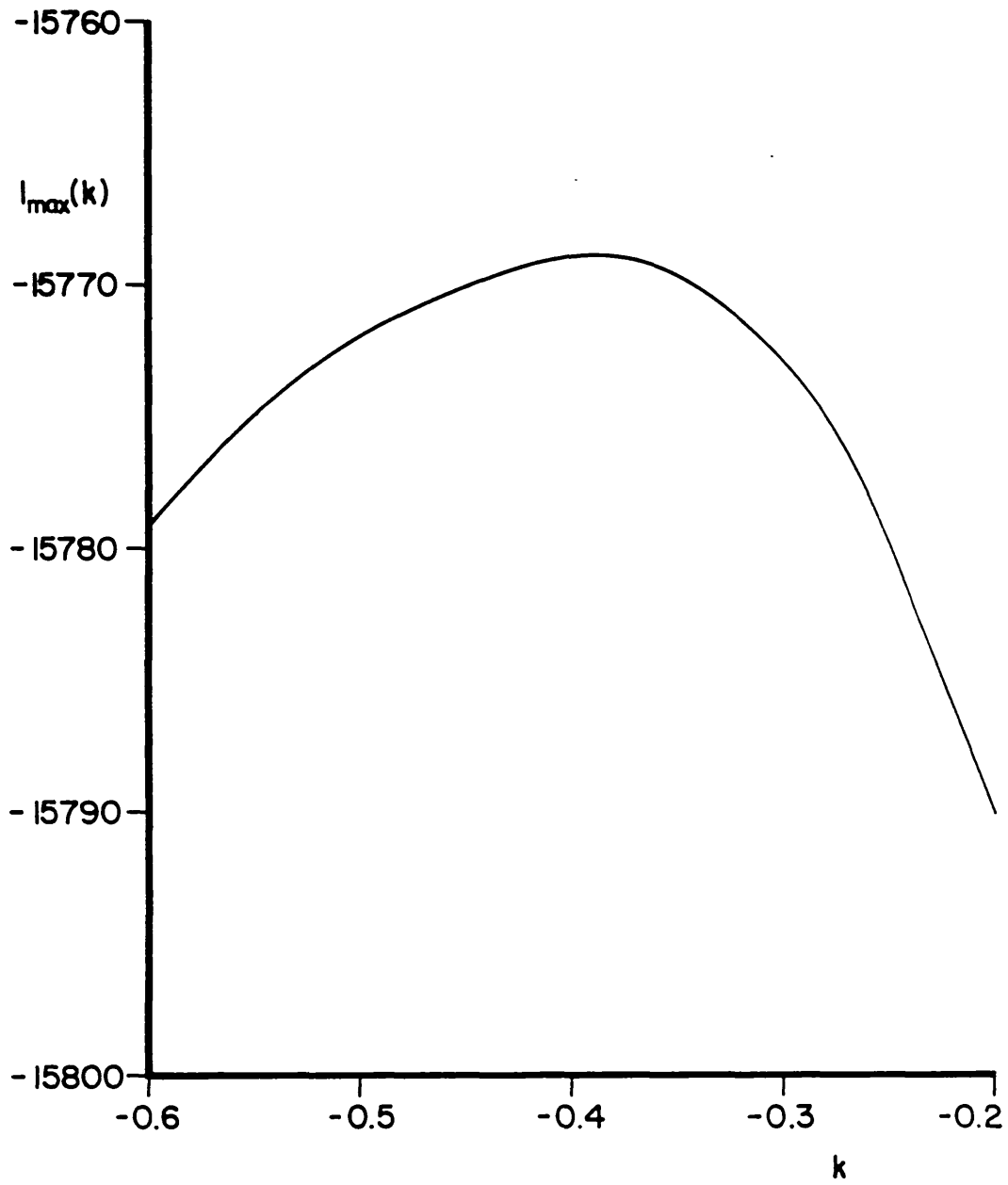


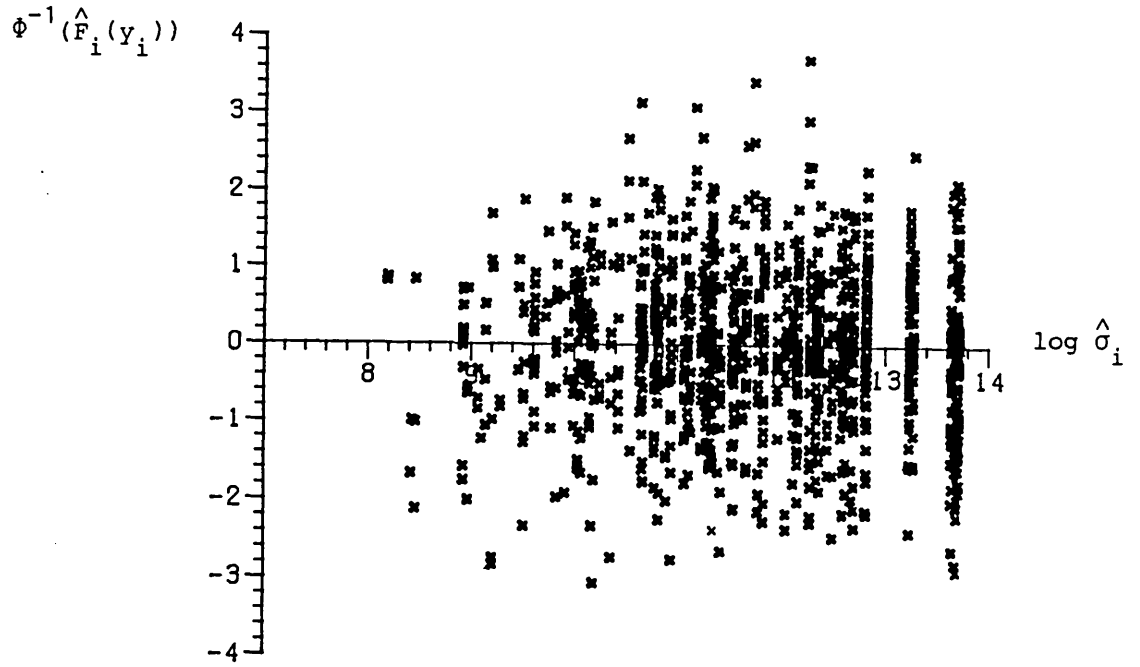
Figure 8.3 : Partially maximized loglikelihood $l_{\max}(k)$ for high exposures to wet deposition; threshold $\xi_0=0$.

those based on inverse expected and observed information are $(-0.48, -0.32)$ and $(-0.47, -0.32)$ respectively. The maximum likelihood estimate of k is about -0.4 . The data plainly do not have an exponential tail. Testing the hypothesis of one value of k at receptors 100 km from their source but a different one at all other distances gives a Wilks' statistic significant at the 5% but not at the 1% level. The test for different values of k for exposures due to releases of duration one day or more, or twelve hours or less, is not quite significant at the 0.1% level. Thus there is again good evidence of heterogeneity of values of k .

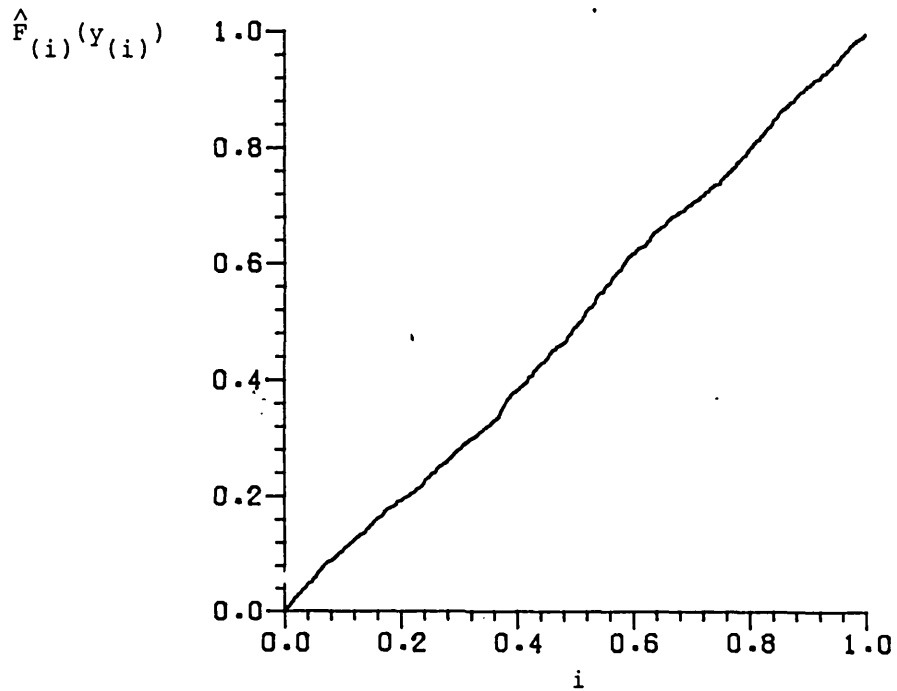
The score test statistics for goodness of fit - nominally χ^2_2 if k has absolute value less than one-half - are 1.77 and 0.35 for $\xi_0=0$ and 0.5 respectively; whereas that for $\xi_0=1$, nominally standard Normal, is 0.52. None of these indicate gross departures from the model. Nor do the residual plot in Figure 8.4(a) and the plot of the empirical distribution function of the $\hat{F}_1(Y_1)$ in Figure 8.4(b), for $\xi_0=0$. The one-step jackknifed parameter estimates show that some of the observations which correspond to pseudo-nuclides have large influence on one or more of the estimated parameters. This is because there is little information for determining β_1 in exposures due to the other, longer-lived, isotopes.

The Anderson-Darling statistic for the fit of the model at equation 8.3.3 with $\xi_0=0$ and the corresponding parameter estimates in Table 8.7 to the data retained for cross-validation is 2.12 - significant at between its 10% and 5% levels. The model fits the data adequately, despite the heterogeneity of the values of k and the fluctuations in the parameter estimates at different thresholds.

It is recommended that the threshold level $\xi_0=0$ be chosen for prediction of high exposures to time-integrated air concentration, dry deposition, and wet deposition, and the relevant parameter values



(a) Normal Cox-Snell residuals $\Phi^{-1}(\hat{F}_i(y_i))$ vs. $\log \hat{\sigma}_i$.



(b) empirical distribution function of uniform residuals $\hat{F}_i(y_i)$

Figure 8.4 : Residual plots for model for high exposures to wet deposition; threshold $\xi_0=0$.

in Tables 8.2, 8.4, 8.6 or 8.7 used. For releases of duration three hours an exposure will exceed the threshold $\xi_0=0$ with probability about 0.005-0.01; the probability of such an exceedance increases for increasing release durations.

Under the clustering model proposed in the previous section, the probability that a single release of duration t hours will lead to an exposure which exceeds the appropriate threshold is $\rho_t/(1-q_t)$, where q_t is taken from Table 8.2 - $q_t=0$ if t is greater than 24 hours. The distribution of an excess over the threshold for an exposure to time-integrated air concentration is the generalized Pareto form 8.3.1 with scale parameter σ defined in equation 8.3.2 and with parameter estimates taken from Table 8.6; for a wet deposition distribution it is the generalized Pareto form with scale parameter σ defined in equation 8.3.3 and parameter estimates taken from Table 8.7.

Consider exposures to a nuclide with parameters λ , v_d , and λ_w , at a distance d metres from a source, for releases of duration t hours. Since $k \neq 0$ for $\xi_0=0$, the estimated return value \hat{x}_p for such exposures for a given probability p , conditionally on the threshold having been exceeded, is

$$\hat{x}_p = \hat{\sigma} \{ 1 - (1-p)^k \} / k \quad \dots 8.3.4.$$

If D is the matrix whose diagonal elements are $s.e.(\hat{\beta}_0)$, $s.e.(\hat{\beta}_1)$, ..., $s.e.(\hat{\beta}_5)$, $s.e.(\hat{k})$ and zeroes elsewhere, and V is the correlation matrix for the estimates - in Table 8.8(a) or (b) - then the variance of the predicted return value \hat{x}_p is

$$\text{Var}(\hat{x}_p) = \hat{x}_p^2 \tilde{x} D V D \tilde{x}^T \quad \dots 8.3.5,$$

where \tilde{x} is the row vector

$$(1, d\lambda, dv_d, d\lambda_w, \log\{d\}, \log\{t/3\}, -\log\{1-p\}(\sigma/kx_p - 1) - 1/k).$$

	parameter						
	β_0	β_1	β_2	β_3	β_4	β_5	k
β_0	1.0	.482	.289	.093	-.995	-.194	.110
β_1		1.0	.060	.216	-.517	.093	.006
β_2			1.0	-.702	-.271	-.551	.072
β_3				1.0	-.129	.316	-.032
β_4					1.0	.148	-.055
β_5						1.0	-.105
k							1.0

(a) Exposures to air concentration

	β_0	β_1	β_2	β_3	β_4	β_5	k
β_0	1.0	-.101	.626	.662	-.997	-.022	-.023
β_1		1.0	-.198	-.353	.095	.290	-.007
β_2			1.0	.207	-.646	.044	-.012
β_3				1.0	-.676	-.039	-.049
β_4					1.0	-.021	.050
β_5						1.0	-.042
k							1.0

(b) Exposures to wet deposition

TABLE 8.8: Correlation matrices for estimated parameters of generalized Pareto distribution for $\xi_0 = 0$

Thus an approximate $(1-2\alpha)\times 100\%$ confidence interval for the value of x_p which would have been obtained had MESOS been used for the calculation is $(\hat{x}_p + z_\alpha \sqrt{\text{Var}(\hat{x}_p)}, \hat{x}_p - z_\alpha \sqrt{\text{Var}(\hat{x}_p)})$, where $\Phi(z_\alpha) = \alpha$, α is one-half or less, and $\Phi(\cdot)$ is the standard Normal integral.

The distribution of the maximum exposure Y_{\max} at a receptor due to T successive t -hour releases each of duration t hours is found using equations 8.2.1 and 8.2.2; it is

$$\text{Prob}(Y_{\max} < y) = \exp\left[T\rho_t \left\{ \frac{(1-q_t)F(y)}{(1-q_t)F(y)} - 1 \right\} \right].$$

The percentile y_p for Y_{\max} corresponding to a given probability $p > \exp\{-T\rho_t\}$ may be found by solving the equation

$\text{Prob}(Y_{\max} < y_p) = p$, and is

$$\hat{y}_p = \hat{\sigma} \left\{ 1 - (1-p^*)^k \right\} / k \quad \dots 8.3.6$$

since $k \neq 0$, where

$$p^* = \frac{T\rho_t + \log(p)}{T\rho_t + q_t \log(p)} \quad \dots 8.3.7.$$

If $p < \exp\{-T\rho_t\}$, then $\hat{y}_p = 0$ since there is an atom of probability of size $\exp\{-T\rho_t\}$ at $y=0$, corresponding to the event of no exceedances of the threshold in an interval of length T .

The estimation error of ρ_t is very small compared with that of $\hat{\beta}$ and \hat{k} , so it is suggested that it be ignored in finding confidence intervals for the MESOS values of y_p . Thus the variance of \hat{y}_p is

$$\text{Var}(\hat{y}_p) = y_p^2 \hat{x}^* \text{DVD} \hat{x}^{*T} \quad \dots 8.3.8,$$

where \hat{x}^* is the vector of covariates defined above but with p^* substituted for p in its last term.

These formulae for variances of the estimated return values \hat{x}_p and \hat{y}_p depend on several approximations: first, Normal approximations

to the distribution of the parameter estimates and return values; second, the approximations introduced by use of the generalized Pareto distribution; and third, the approximations introduced by using MESOS rather than experimental or observational data. The third of these dominates the others. To anticipate the conclusions of Section 8.4: although the model for tail behaviour is apparently very accurate, it is only as trustworthy as the MESOS calculations on which it is based.

8.4 A verification study

It is meet and right that the model for extremes be verified by comparison with data not used to derive it - only thus can it be properly assessed. The data are for MESOS calculation for releases from Hannover and Stuttgart through 1973. The notional radionuclides concerned - labelled Cases 1-4 for convenience - mostly have deposition parameters larger than those used to develop the model. Their nuclide characteristics are summarized in Table 1.4. Case 1 has deposition parameters in the same range as those used to derive the model, and an infinite half-life; Case 2 has a larger deposition velocity v_d , and an infinite half-life; Case 3 has a half-life of just less than three days and large deposition velocity and washout coefficient λ_w ; and Case 4 has an infinite half-life and the same deposition parameters as Case 3.

Consider first the verification of the model for episodes of high exposure to time-integrated air contamination, for threshold levels defined at equation 8.1.1 with $\xi_0=0$ and releases from Hannover. The three elements of the model to be considered are: the occurrence of clusters of exposures; the sizes of the clusters; and the sizes of the individual excesses in the clusters.

Under the model the numbers of clusters observed at receptor j

due to T releases each of duration t hours is Poisson with mean $E_j = T\rho_t$ and ρ_t is defined at equation 8.2.3. If the observed number of clusters at the receptor over the period is O_j , then the statistic

$$X^2 = \sum_j (O_j - E_j)^2 / E_j,$$

where the sum is over all 16 receptors, should be distributed approximately as χ^2_{16} if the model is a good fit to the data, and should be rather large otherwise.

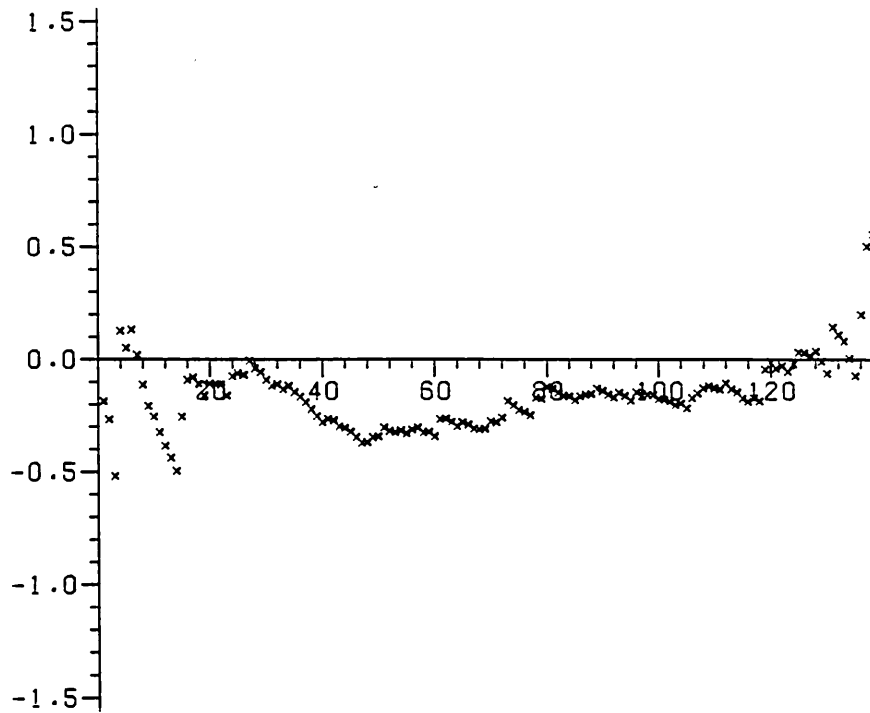
For Case 1, and releases of duration 3 hours, 12 hours, and 3 days, the values of X^2 are respectively 18.54, 22.42, and 11.11, none of which casts serious doubt on the model. The largest contributions to the values of X^2 come from the receptors 100 and 200 km south of Hannover, where the numbers of clusters are underestimated for shorter release durations. The numbers of clusters of exposures exceeding the relevant thresholds for releases of Case 2, Case 3, and Case 4 - all of which have deposition parameters bigger than those used to derive the model - are considerably underestimated, especially far from Hannover. This implies that for time-integrated air concentrations and dry deposition threshold, dependence on nuclide deposition velocity is too strong, leading to too low thresholds at long distances. As a result, the numbers of clusters of exposures experienced at receptors more than 300 km from their sources are generally underpredicted by a factor of between three and five.

In addition to the numbers of clusters, the observed and predicted sizes of the clusters of exposures at all 16 receptors can be compared using the statistic X^2 . The O_j are now the observed sizes of clusters and the E_j their expected sizes, and the statistic should approximately the χ^2_7 distribution. For Case 1 and releases of duration 3 hours and 12 hours, X^2 has values 13.22 and 11.82

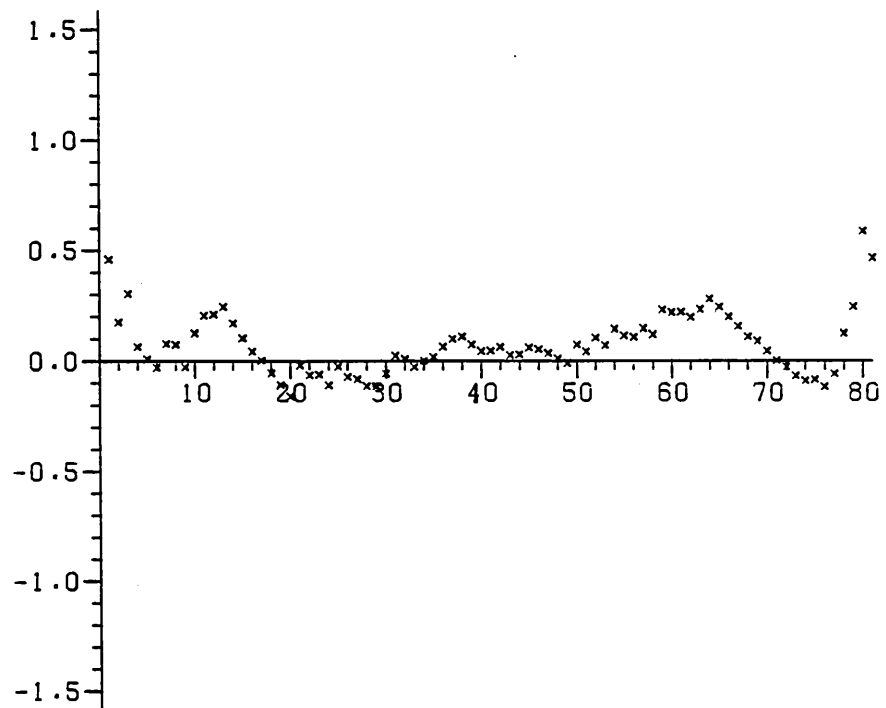
respectively, both of which indicate slight overdispersion of the data relative to the model. For the other three notional isotopes the model fits very poorly because the threshold is too low so that far from the source clusters are bigger than the model predicts - leading to massive overdispersion. For Case 2, for example, the values of X^2 for releases of duration 6 hours and one day are 30.17 and 132.3 respectively: the low threshold renders the clustering model inadequate.

The situation improves considerably when the sizes of the excesses themselves are considered. For Case 1 and $\xi_0=0$, conditionally upon the appropriate parameter values in Table 8.6, the Anderson-Darling statistics for fit of the generalized Pareto distribution to excesses at all 16 receptors are 1.088, 0.725, and 0.159 for releases of duration 3 hours, 12 hours, and 3 days respectively. All of these indicate a good fit of the distribution to the data. The picture is equally rosy for other radionuclides: for example, the Anderson-Darling statistic for exposures to Case 3 due to releases of duration one week is 1.252; and that for exposures to Case 4 due to releases of duration 3 hours is 2.094. The nominal 5% level of the statistic is 2.492 for samples of size 5 or more, so there is no evidence of any statistical discrepancies between the observed and predicted distributions.

To assess the practical significance of such discrepancies as do occur, Figure 8.5 shows for various nuclides and release durations plots of $\log\{o_{(i)}/e_{(i)}\}$ against i , where $o_{(i)}$ is the (appropriately scaled) observed i^{th} order statistic of the excesses, and $e_{(i)}$ is the expected i^{th} order statistic for a generalized Pareto distribution with shape parameter $k=-0.233$. If the expected and observed distributions matched perfectly, the plot would be a straight line of gradient zero through the origin.

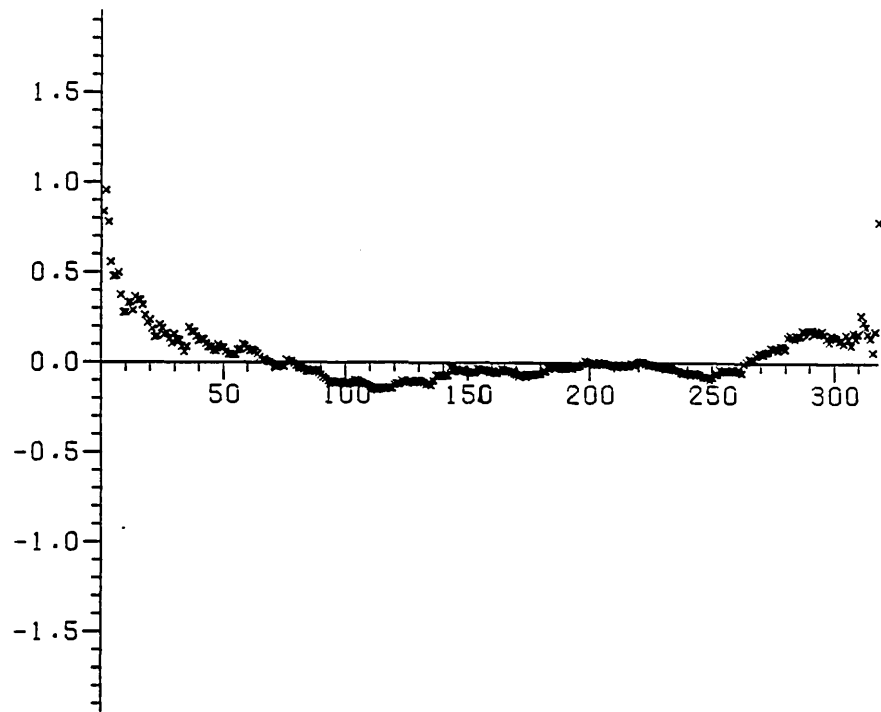


(a) Case 1, release duration 3 hours.

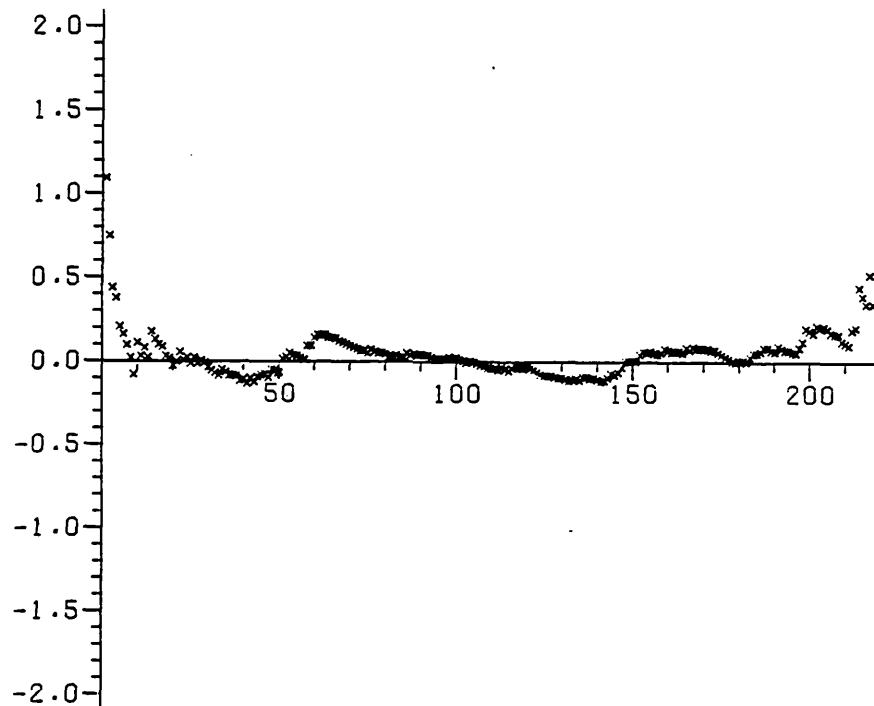


(b) Case 1, release duration 1 day.

Figure 8.5 : Plots of differences of log-observed and log-expected order statistics for high levels of air contamination due to releases from Hannover.

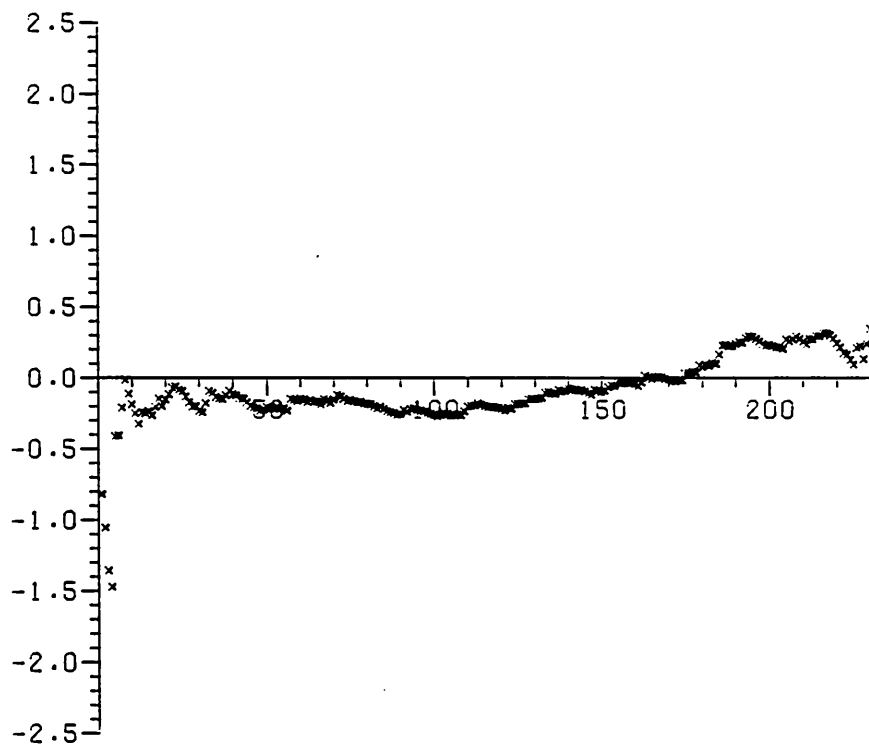


(c) Case 2, release duration 6 hours.

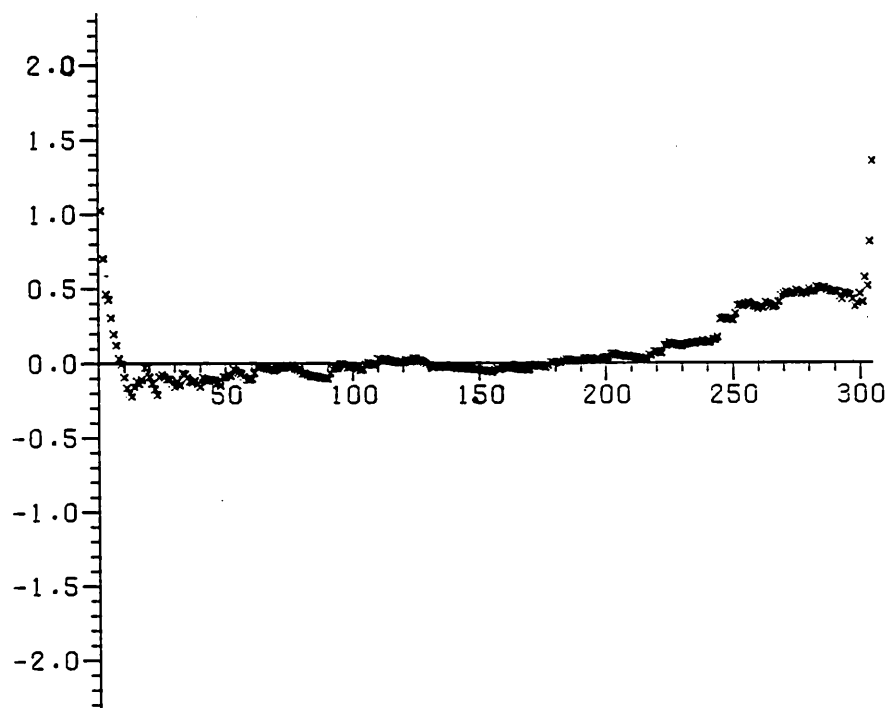


(d) Case 2, release duration 1 day.

Figure 8.5 : Plots of differences of log-observed and log-expected order statistics for high levels of air contamination due to releases from Hannover.



(e) Case 3, release duration 3 hours.



(f) Case 4, release duration 3 hours.

Figure 8.5 : Plots of differences of log-observed and log-expected order statistics for high levels of air contamination due to releases from Hannover.

Figure 8.5(a, b) shows the plot for Case 1 exposures due to releases of duration 3 hours and 3 days. In Figure 8.5(a) the expected distribution levels are too high by a factor 1.3 or so for almost all values of i ; part (b) of the figure shows a very good fit throughout the range of the data. The rest of the figure shows the same pattern: observed and predicted values generally lie within a factor 1.3 or so of each other, although the bottom few exposures are usually a factor 1.6 or so too low and the uppermost few the same factor too high.

To draw conclusions from these data, the model for high levels of exposure to time-integrated air concentrations and dry deposition is very good for radionuclides whose parameters lie in the range of those used to build the model. There is evidence that probabilities of high levels of exposure are slightly overdispersed relative to those predicted by the model, but the prediction of high levels of contamination is very accurate indeed. The model for probabilities of high exposures to isotopes with big deposition and washout parameters is less adequate because thresholds are too low at moderate and long distances from the source: uncritical use of the model could lead to probabilities of high exposures being underestimated by a factor of up to five. However the model for the sizes of these high exposures is about as accurate for such radionuclides as for those used to derive the model.

Consider now the verification of the model for episodes of high exposures to wet deposition for threshold levels defined at 8.1.2 with $\xi_0=0$.

For Case 1 the X^2 statistics for comparison of expected and observed numbers of clusters at all 16 receptors for releases from Hannover of duration three hours and one day are 26.31 and 15.17 respectively; for releases from Stuttgart of duration 3 hours, 12

hours and one week they are 122.4, 85.53, and 28.25. These statistics clearly indicate some massive overdispersion of the data relative to the model. Comparison of the observed and expected numbers of clusters shows that major discrepancies arise only at a few receptors: those 100 and 200 km east of Hannover; and those 100 km north-east, 100 and 200 km south-east, and 600 km south-west of Stuttgart. Overdispersion due to discrepancies at the same receptors is evident - though to a lesser extent - for exposures to wet deposition of Case 2. This implies two things: that the frequency of high exposures to wet deposition is underpredicted by a factor of up to four at receptors generally downwind of the source in wet conditions; and that effects such as orographic rain may have a substantial effect on the frequency of such incidents. For Cases 3 and 4, with their higher washout coefficients, thresholds are rather high and the numbers of clusters of high exposures tend to be slightly overpredicted, but the observed and expected numbers are not significantly different.

The χ^2 statistic for comparison of observed and predicted cluster sizes for exposures to wet deposition of Case 1 are 6.08 and 2.22 for releases from Hannover of duration 3 hours and one day; and 25.72 and 15.17 for releases from Stuttgart of duration 3 hours and 12 hours respectively. This is good evidence that observed cluster sizes are overdispersed relative to the model; in the Stuttgart data this is mostly a result of the high number of exposures at the receptors 100 and 200 km south-east of the source. The same is true of exposures to Case 2 wet deposition, but there is no evidence of lack of fit for exposures due to wet deposition of Cases 3 or 4.

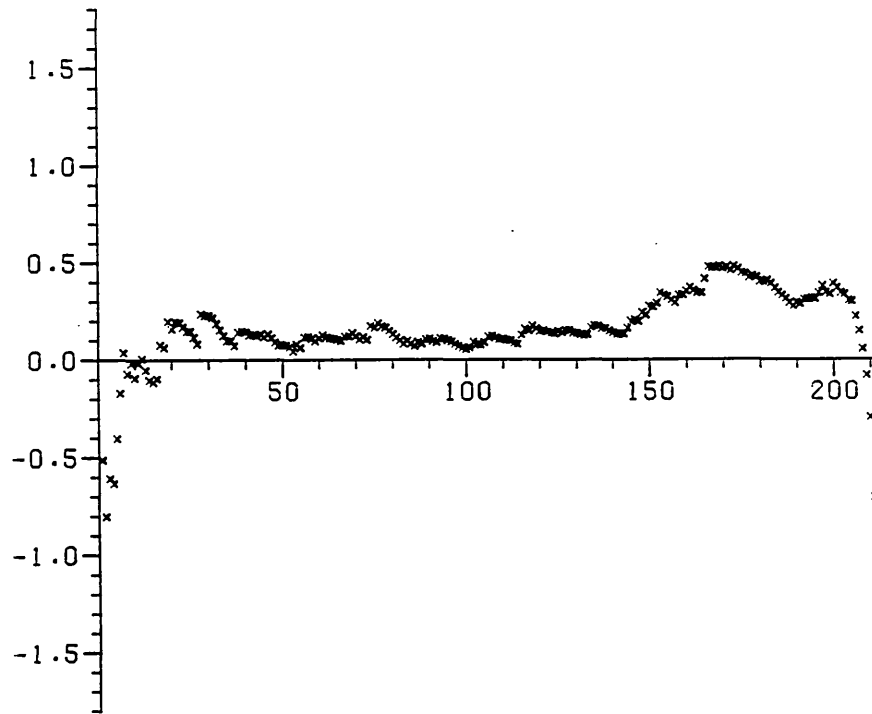
The Anderson-Darling statistics for excesses of Cases 1 exposures to wet deposition are 2.228, 2.776, and 0.473 for releases from Stuttgart of duration 3 hours, 12 hours, and one week

respectively; and are 0.204 and 1.764 for releases from Hannover of duration 3 hours and one day. The Stuttgart results give some evidence of lack of fit. The same is true of exposures due to releases of Case 2 from Stuttgart - for releases of duration 6 hours and one day the Anderson-Darling statistics are 4.693 and 4.318 respectively, both significant at at least the 1% level. However there is no evidence of discrepancies between the data and the model for exposures to wet deposition of Cases 3 or 4.

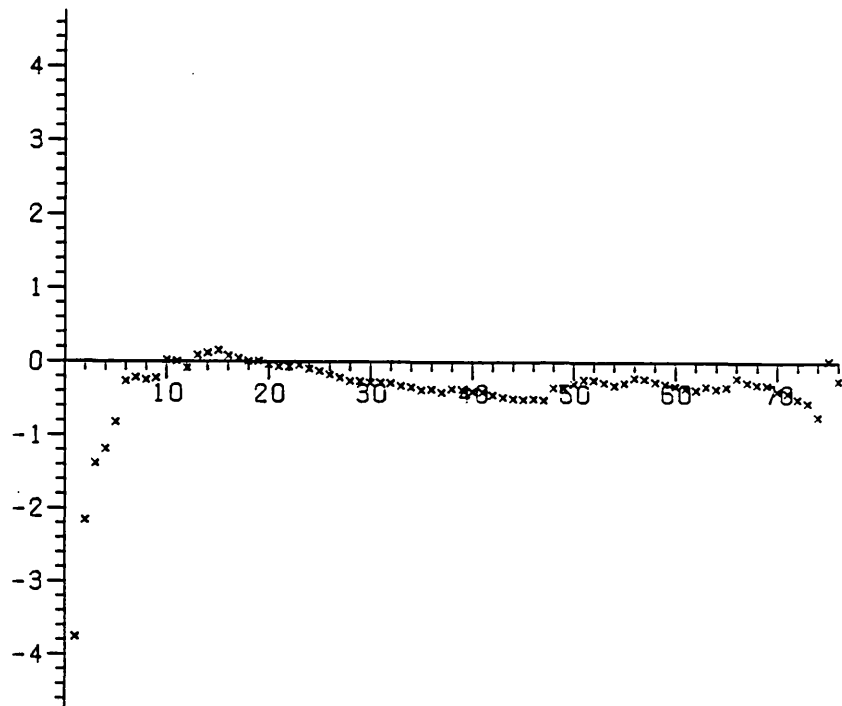
In Figure 8.6 the practical significance of these differences is assessed using the order statistic plots described above. Figure 8.4 (a, b) shows their size for releases of duration three hours from Stuttgart and one day from Hannover. Apart from overprediction of the top and bottom few order statistics, the distributions lie within a factor 1.6 of each other. This is true for other nuclides, as evidenced by the rest of the figure.

Thus for episodes of exposures to wet deposition of nuclides with parameters in the range of those used to derive the model it may be concluded that: first, probabilities of such episodes are generally accurately predicted but may be underestimated by a factor of up to four at receptors either downwind of the source in rainy conditions, or at those where there is orographic enhancement of rain; and second, that the distributions of high values derived using the statistical model generally lie within a factor 1.6 of the MESOS results, and often closer. For isotopes whose washout coefficients are bigger than those used to derive the model, probabilities of exposure to wet deposition are slightly but not seriously underestimated using the model; and the values of high exposures are generally accurate to within a factor of 1.6.

For short release durations the absolute values of the probabilities of the extreme events being considered here are often

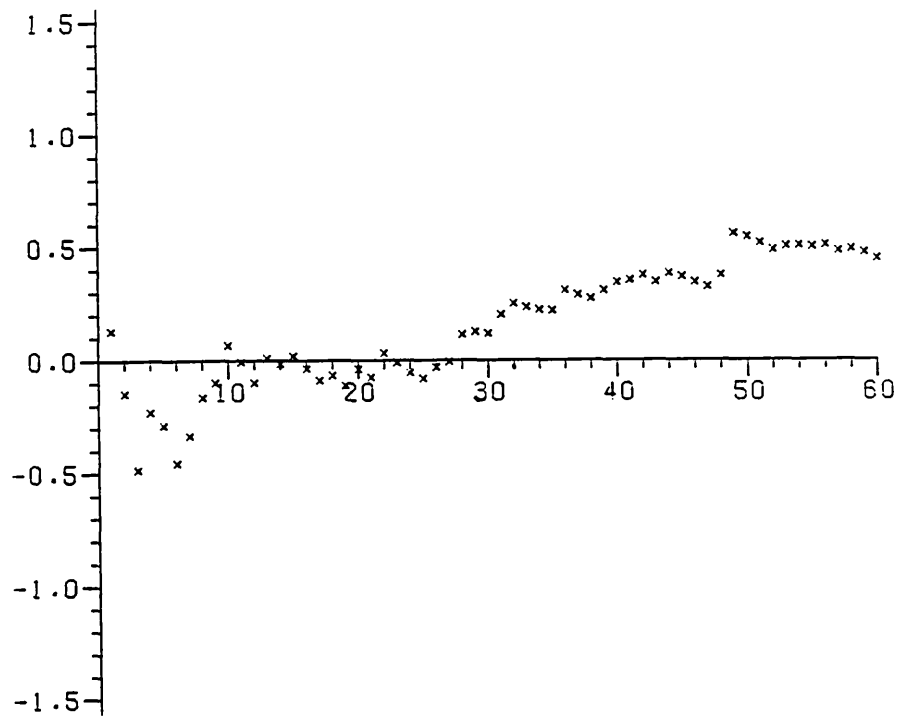


(a) Case 1, releases of duration 3 hours from Stuttgart.

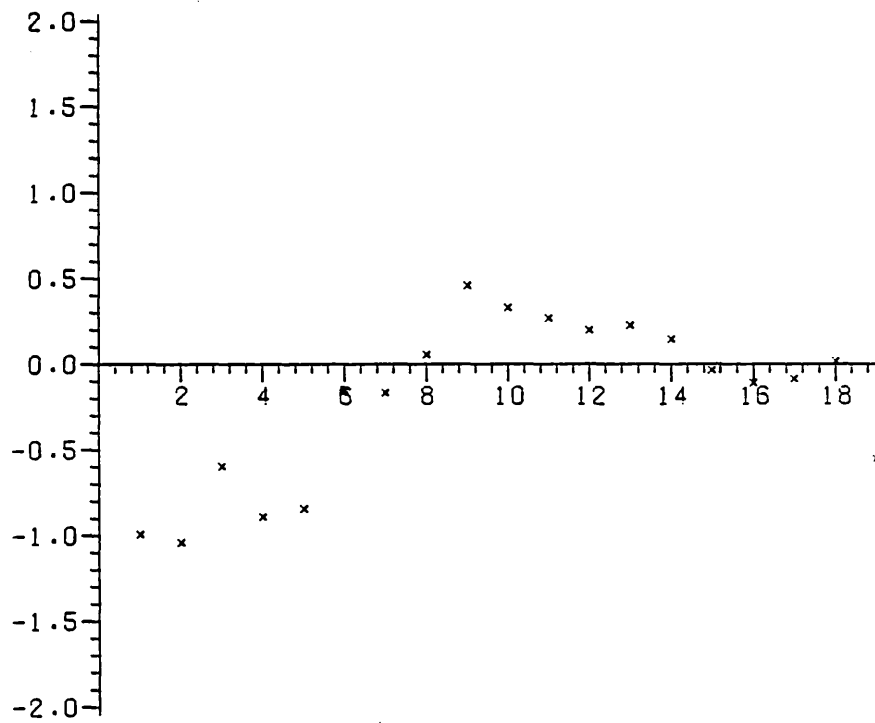


(b) Case 1, releases of duration 1 day from Hannover.

Figure 8.6 : Plots of differences of log-observed and log-expected order statistics for high levels of wet deposition.

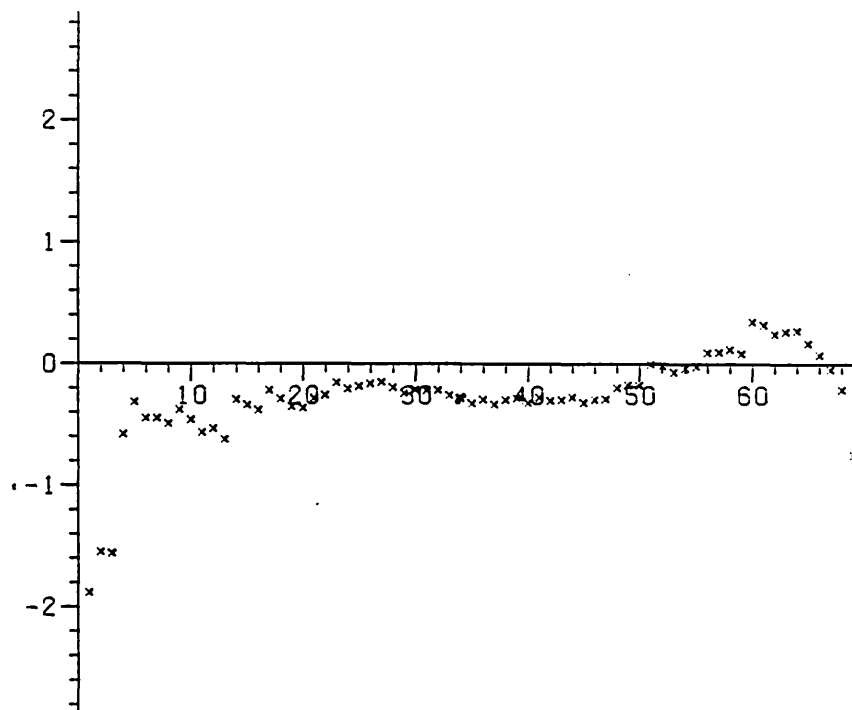


(c) Case 2, releases of duration 1 day from Stuttgart.

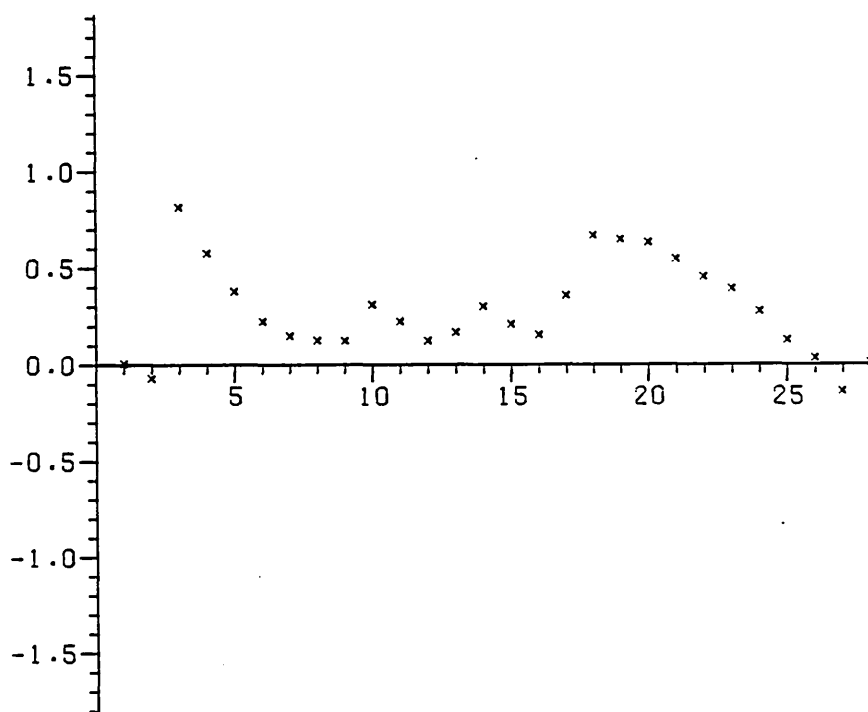


(d) Case 3, releases of duration 3 days from Stuttgart.

Figure 8.6 : Plots of differences of log-observed and log-expected order statistics for high levels of wet deposition.



(e) Case 4, releases of duration 3 hours from Stuttgart.



(f) Case 4, releases of duration 1 day from Stuttgart.

Figure 8.6 : Plots of differences of log-observed and log-expected order statistics for high levels of wet deposition.

of the order of 0.01 or less, so even when multiplied by the factors of four or five by which they may be in error, the events are rare. Prediction of probabilities of exposure to air contamination, dry deposition, and wet deposition, is generally better for releases of duration longer than one day. Although not perfect, this model for high exposure episodes generally works well.

9. SUMMARY, CONCLUSIONS, AND DISCUSSION

This thesis has two main themes. The first is the provision of a method of estimating the distributions of exposure due to airborne transport of radioisotopes from a source situated at any point in Western Europe to a receptor between 100 and 1000 kilometres away, based on a statistical analysis of the results of the MESOS model for long-range atmospheric dispersion. The second is the development and study of the properties of statistical techniques for modelling exceedances of continuous random variables over high - or under low - thresholds, based on the use of the generalized Pareto distribution.

Chapter 1 provides background and motivation for the task undertaken in Chapters 2, 3 and 8: the statistical analysis of the MESOS exposure data. MESOS - a complex puff trajectory model for long-range atmospheric pollutant transport - and its associated input databases of meteorological measurements and output databases of puff and receptor histories are described. Attempts to verify MESOS by internal calibration and comparisons with other models and to validate it by reference to observational data are discussed, and the current lack of data suitable for a comprehensive validation of such models is lamented. Some areas of uncertainty in MESOS - notably its modelling of rainfall, and simple parametrizations of complex physical processes - are discussed; so too are the strong statistical assumptions made for the purpose of later data-analysis.

A statistical analysis of the MESOS probabilities of exposure to air contamination and wet deposition is performed in Chapter 2. Two equations which account for much of the observed variation in exposure probabilities are derived by a mixture of heuristic physical, probabilistic, and empirical arguments; and parameters in them are estimated from the MESOS data. Parameter values depend on the type of exposure which they describe - air contamination or

wet deposition - and the climatology of the source. Investigation of the second leads to the study of geostrophic windroses in Section 2.5, where naive and more sophisticated classifications of windroses are made. In Section 2.6 verification of the equations leads to the conclusion that prediction of probabilities is accurate except where trajectories are systematically diverted by features such as mountain barriers. Prediction of probabilities of exposure to wet deposition is less accurate than to air contamination because of local effects - such as orographic rainfall - which it is hard to include in such a model. Notwithstanding this, prediction is generally good, and it will often be possible to assess qualitatively the effects of trajectory turning and localized rainfall.

Attention turns in Chapter 3 to the estimation of the distribution of contamination at an exposed receptor. This is more complicated for a variety of reasons, chief among them that exposure distributions depend on many factors whose combined effect may be assessed qualitatively but is difficult to parametrize in a simple form capable of direct physical interpretation. In the absence of physical arguments for a particular probability distribution of exposures, an empirical approach is taken which leads to the choice of the two-parameter Weibull form to represent the MESOS exposure data. Weibull distributions fit the MESOS exposures well enough for the present purpose. Heuristic physical arguments suggest equations for the dependence of exposure distributions on factors such as nuclide characteristics, source-receptor distance, and release duration; this leads to the estimation of associated parameters - some of which have a direct physical interpretation - from the data. In Section 3.4 verification of the equations indicates that they are generally sufficiently accurate for their intended purpose. Statistically predicted distributions usually lie within a factor of

two to three of their MESOS counterparts - which are thought to be generally accurate to a similar factor - over the important part of their ranges. However, if extrapolated beyond their ranges of validity they may lead to serious prediction errors. For these reasons this part of the model must be regarded as less satisfactory than the prediction of exposure probabilities, but despite this is simple to use and correctly used is accurate to the same order as MESOS.

Chapter 4 gives a short review of modelling excesses of continuous random variables over high thresholds - which has roots in the hydrological 'Peaks Over Threshold' models, and introduces a major tool in such analysis: the generalized Pareto distribution. The distribution is studied in some detail in Chapter 5, where its basic properties are described, with particular emphasis on those useful for statistical inference; as an aside it is characterized by a generalization of the Hamel equation.

Statistical inference for the distribution is considered at length in Chapter 6, where a general formulation is given for maximum likelihood estimation in complex data. Estimation - mostly by maximum likelihood - in simple random samples is studied in some detail, and the information in the tails of the distribution is weighed by consideration of the effect of censoring and influence curves for maximum likelihood estimates. It is concluded by use of simulation that uncritical use of large-sample theory for confidence regions can be dangerous since convergence to asymptotic distributions of estimates is slow. Much information about the values of parameters is contained in the tails of the data, with the particular implication that any systematic errors in the calculated MESOS high exposures will have a profound effect on statistical models for them. Estimation by moments and least squares are

discussed in special cases and their efficiency assessed: they are not generally to be recommended. Two tests for tail weight are considered in Section 6.4, one of which is to be preferred on the basis of its small-sample properties.

In Chapter 7 diagnostic techniques are developed for the models described in Chapter 6, based on residuals, a score test for goodness of fit, and sample influence calculations. A simple example shows that in small samples critical information about parameter values may be contained in the biggest sample order statistic.

Chapter 8 describes the analysis of high exposure episodes in the MESOS exposure database, based on the techniques developed in the previous few chapters. A simple model for clustering of high values is proposed and the MESOS exposures examined to see if it fits them. The generalized Pareto distribution is fitted to excesses over empirically determined thresholds, and resulting models for extreme exposures are discussed and verified. The statistical model for high exposure levels gives results very close to the MESOS data, even for isotopes not used to derive it. Statistically predicted and MESOS high exposures generally lie within a factor 1.6 or so of each other - and often less. Prediction of the probabilities of episodes of high exposures to air contamination and dry deposition is good for radionuclides with parameters in the range of those used to derive the model, but is less accurate for others. There is underprediction of probabilities of high exposures to wet deposition at receptors usually downwind of the source in wet conditions or due to orographic enhancement of rain, which is not taken account of in the model.

The statistical model developed in Chapters 2, 3, and 8 is a useable and useful tool which predicts exposure probabilities and distributions to about the same accuracy as MESOS, for a wide variety of radionuclides, release durations, and sources in Western Europe.

Its use is illustrated in a forthcoming CEC report. Its main virtues are: its simplicity and conciseness in describing a complex physical situation; its accuracy over the range for which it was derived; and the fact that its statistical basis enables explicit statements to be made about the probable degree of accuracy of quantities derived from it. This last quality is highly desirable, especially in view of use within the nuclear industry of probabilistic risk assessment.

However statements of statistical uncertainty based on this work are always made relative to the results of the dispersion model MESOS, not to some physical experiment designed to validate such results. When - if - suitable results from such experiments become available, it may be possible to give more positive guidance about the relative importance of statistical errors and those introduced in the course of physical modelling.

The main vice of the empirical statistical approach taken in this thesis is undoubtedly that although its elements are generally physically motivated, not all of them are capable of direct interpretation in terms of physical processes. A drawback of this is that it does not necessarily give fresh insight into the processes. On the other hand to generate such insight is not a primary aim of this work, and whilst it would be possible to build a stochastic model for dispersion based on analysis of the MESOS puff histories there is no guarantee that it would be as simple or concise as that developed here. Moreover it is likely that any attempts to incorporate a stochastic element into a trajectory dispersion model will have empirical components like those here.

The main contribution of Chapters 4-7 lies not in the originality of the basic ideas - covariate-dependent data, estimation by maximum likelihood, diagnostic techniques, and so on - but in their wholesale application to the modelling of sample extremes, and

in particular to excesses over high thresholds using the generalized Pareto distribution. To my knowledge this is new. It makes possible the construction of flexible and potentially rather complex models for extremes, together with apparatus for fitting them, assessing their fit, and testing relevant hypotheses. Big issues avoided include: the choice of threshold over which to apply the generalized Pareto approximation; the related question of so-called penultimate approximations, raised by the behaviour of the tails of the data in Section 8.3; problems - such as non-independence of nearby excesses - associated with the clustering of high values; problems posed by transformations chosen to accelerate convergence of data to asymptotic extremal distributions; and the analytic study in small samples of maximum likelihood estimates. Nor has attention been paid to computational problems which may arise in maximum likelihood estimation. Bayesian inference and decision-theoretic issues have not been covered, but this is not to deny their potential importance in other extremal situations.

As a final point, it cannot have escaped the readers' notice that by focusing on exposures at single receptors - albeit sited anywhere in a large annulus centred on the source - the spatial structure of the data has been emasculated. As a result, questions about simultaneous exposures at two or more receptors cannot be answered. Stochastic models capable of tackling these problems will almost certainly have a strong underlying physical basis. Statistics of extremes for such data are almost unknown.

REFERENCES

- Aitken, M. and Clayton, D.(1980). The fitting of exponential, Weibull, and extreme-value distributions to complex censored survival data using GLIM. *Appl. Statist.* 29, 156-163.
- Alecio, N.G.(1983). A critical review of radiological aspects of siting of nuclear installations. PhD thesis, University of London.
- Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. and Tukey, J.W.(1972). *Robust Estimates of Location: Survey and Advances*. Princeton University Press, Princeton, N.J.
- ApSimon, H.M. and Goddard, A.J.H.(1983). Atmospheric transport of radioisotopes and the assessment of population doses on a European scale: application of the MESOS code to the meteorological dispersion of radioactive discharges from notional nuclear sites in the European Community with particular reference to the mesoscale. Commission of the European Communities, Radiation Protection Report EUR 9128 EN.
- Atkinson, A.C.(1973). Testing transformations to Normality. *J. Roy. Statist. Soc. B*, 35, 473-479.
- Atkinson, A.C.(1982). Regression diagnostics, transformations, and constructed variables (with discussion). *J. Roy. Statist. Soc. B*, 44, 1-36.
- Atkinson, A.C.(1985). *Plots, Diagnostics, and Regression*. Oxford University Press.
- Baker, R.J. and Nelder, J.A.(1978). *General Linear Interactive Modelling (GLIM)*, Release 3. Numerical Algorithms Group, Oxford.
- Bartlett, M.S.(1952). Approximate confidence intervals, II. *Biometrika*, 40, 306-317.
- Berger, A., Melice, J.L. and Demuth, Cl.(1982). Statistical distributions of daily and high atmospheric SO₂-concentrations. *Atmos. Envir.*, 16, 2863-2877.
- Box, G.E.P. and Cox, D.R.(1964). An analysis of transformations (with discussion). *J. Roy. Statist. Soc. B*, 26, 211-246.
- Burrows, P.M.(1979). Selected percentage points of Greenwood's statistic. *J. Roy. Statist. Soc. A*, 142, 256-258.
- Cheng, R.C.H. and Amin, N.A.K.(1983). Estimating parameters in continuous univariate distributions with a shifted origin. *J. Roy. Statist. Soc. B*, 45, 394-403.
- Clarke, R.H.(1979). A model for short and medium range dispersion of radionuclides released to the atmosphere. National Radiological Protection Board, report NRPB-R91.
- Cook, R.D. and Weisberg, S.(1982). *Residuals and Influence in Regression*. Chapman and Hall, London.

- Cook, R.D. and Weisberg, S.(1983). Diagnostics for heteroscedasticity in regression. *Biometrika*, 70, 1-10.
- Cox, D.R.(1961). Tests of separate families of hypotheses. *Proc. Fourth Berkeley Symp.*, 1, 105-123.
- Cox, D.R.(1962). Further results on tests of separate families of hypotheses. *J. Roy. Statist. Soc. B*, 24, 406-424.
- Cox, D.R.(1982). Statistical significance tests. *Br. J. clin. Pharmac.*, 14, 325-331.
- Cox, D.R. and Hinkley, D.V.(1968). A note on the efficiency of least squares estimates. *J. Roy. Statist. Soc. B.*, 30, 284-289.
- Cox, D.R. and Hinkley, D.V.(1974). *Theoretical Statistics*. Chapman and Hall, London.
- Cox, D.R. and Isham, V.(1980). *Point Processes*. Chapman and Hall, London.
- Cox, D.R. and Lewis, P.A.W.(1966). *The Statistical Analysis of Series of Events*. Chapman and Hall, London.
- Cox, D.R. and Oakes, D.(1984). *Analysis of Survival Data*. Chapman and Hall, London.
- Cox, D.R. and Snell, E.J.(1968). A general definition of residuals (with discussion). *J. Roy. Statist. Soc. B.*, 30, 248-275.
- Cox, D.R. and Snell, E.J.(1971). On test statistics calculated from residuals. *Biometrika*, 58, 589-594.
- Crompton, S.(1982). *Collective doses from nuclear installation discharges*. PhD thesis, University of London.
- Currie, I.D.(1981). Further percentage points of Greenwood's statistic. *J. Roy. Statist. Soc. A*, 144, 360-363.
- Davis, H.T. and Feldstein, M.L.(1979). The generalized Pareto law as a model for progressively censored survival data. *Biometrika*, 66, 299-306.
- Davison, A.C.(1983). Modelling excesses over high thresholds, with an application. In J. Tiago de Oliveira (ed.), *Statistical Extremes and Applications*. D. Reidel, Dordrecht.
- DeMouchel, W.H.(1983). Estimating the stable index α in order to measure tail thickness: a critique. *Ann. Statist.*, 11, 1019-1031.
- Doury, A.(1976). Une methode de calcul pratique et generale pour la prevision numerique des pollutions vehiculees par l'atmosphere. *Rapport CEA-R-4280*, (rev. 1), SACLAY.
- Draper, N. R. and Smith, H.(1981). *Applied Regression Analysis*. Second edition. Wiley, New York.
- Feller, W.J.(1968). *An Introduction to Probability Theory and its Applications*, Volume I. Third edition. Wiley, New York.

- Ferber, G.J. and Heffter, J.L.(1983). CAPTEXT '83: Cross-Appalachian tracer experiments revised plan, March 1983.
National Oceanic and Atmospheric Administration Air Resources Laboratory, Rockville, M.D.
- Flood Studies Report(1975). Natural Environmental Research Council.
- Georgopoulos, P.G. and Seinfeld, J.H.(1982). Statistical distributions of air pollutant concentrations.
Environ. Sci. Technol., 16, 401A-416A.
- Green, P.J.(1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussion).
J. Roy. Statist. Soc. B, to appear.
- Greenwood, M.(1946). The statistical study of infectious diseases.
J. Roy. Statist. Soc., 109, 85-110.
- Gumbel, E.J.(1958). Statistics of Extremes. Columbia University Press, New York.
- Haldane, J.B.S.(1953). The estimation of two parameters from a sample. Sankhya, 12, 313-320.
- Hampel, F.R.(1968). Contributions to the theory of robust estimation. PhD thesis, University of California, Berkeley.
- Hemming, C.R., Kelly, G.N. and Charles, D.(1983). The effect of non-linear dose-response relationships on the predicted consequences of accidental releases of radioactive material. National Radiological Protection Board, report NRPB-M91.
- Holland, D.M. and Fitz-Simons, T.(1982). Fitting statistical distributions to air quality data by the maximum likelihood method. Atmos. Envir., 16, 1071-1076.
- Hosking, J.R.M.(1984). Testing whether the shape parameter is zero in the generalized extreme-value distribution.
Biometrika, 71, 367-374.
- Jenkinson, A.F.(1969). Statistics of extremes. WMO technical note 98, Chapter 5, 183-227.
- Johnson, R.A. and Haskell, J.H.(1983). Sampling properties of estimators of a Weibull distribution of use in the lumber industry. Can. J. Statist., 11, 155-169.
- Jorgensen, B.(1983). Maximum likelihood computation via the delta-algorithm. Submitted to Biometrika.
- Leadbetter, M.R., Lingren, G. and Rootzen, H.(1983). Extremes and Related Properties of Random Sequences and Series.
Springer-Verlag, New York.
- Maguire, B.A., Pearson, E.S. and Wynn, A.H.A.(1952). The time intervals between industrial accidents.
Biometrika, 39, 168-180.
- Manning, P.M.(1984). Private communication.

- Mardia, K.V., Kent, J.T. and Bibby, J.M.(1979). Multivariate Analysis. Academic Press, London.
- Maul, P.R.(1980). Atmospheric transport of sulphur compound pollutants. PhD thesis, University of London.
- McCullagh, P. and Nelder, J.A.(1983). Generalized Linear Models. Chapman and Hall, London.
- Moran, P.A.P.(1947). The random division of an interval. J. Roy. Statist. Soc. Suppl., 9, 92-98.
- Moran, P.A.P.(1951). The random division of an interval - part II. J. Roy. Statist. Soc. B, 13, 147-150.
- Moran, P.A.P.(1953). The random division of an interval - part III. J. Roy. Statist. Soc. B, 15, 77-80.
- Mosteller, F. and Tukey, J.W.(1977). Data Analysis and Linear Regression. Addison-Wesley, Reading, Mass.
- Neyman, J.(1941). On a statistical problem arising in routine analyses and in sampling inspections of mass production. Ann. Math. Statist., 12, 46-76.
- North, M.(1980). Time-dependent stochastic model of floods. J. Hydraulics Div., ASCE, 106, 649-665.
- Pickands, J. III(1975). Statistical inference using extreme order statistics. Ann. Statist., 3, 119-131.
- Pollack, R.I.(1975). Studies of pollutant concentration frequency distributions. Report no. EPA-650/4-75-004. US Environmental Protection Agency, Triangle Park, N. Carolina.
- Pregibon, D.(1981). Logistic regression diagnostics. Ann. Statist., 9, 705-724.
- Proschan, F. and Pyke, R.(1967). Tests for monotone failure rate. Proc. Fifth Berkeley Symp. Math. Statist. Prob., 3, 293-312.
- Pyke, R.(1965). Spacings (with discussion). J. Roy. Statist. Soc B, 27, 395-449.
- Seber, G.A.F.(1977). Linear Regression Analysis. Wiley, New York.
- Shenton, L.R. and Bowman, K.O.(1977). Maximum Likelihood Estimation in Small Samples. Griffin, London.
- Smith, F.B.(1980). The influence of meteorological factors on radioactive dosages and depositions following an accidental release. Proc. CEC seminar, 22-25 April 1980, Riso, Denmark, 1, 223-245.
- Smith, R.L.(1983). Threshold models for the analysis of sample extremes. In J. Tiago de Oliveira (ed.), Statistical Extremes and Applications. D. Reidel, Dordrecht.
- Smith, R.L.(1985). Maximum likelihood estimation in a class of non-regular cases. Biometrika, to appear.

- Smith, R.L. and Weissman, I.(1985). Maximum likelihood estimation of the lower tail of a probability distribution. J. Roy. Statist. Soc. B, to appear.
- Spiegelhalter, D.J.(1983). Diagnostic tests of distributional shape. Biometrika, 70, 401-410.
- Stephens, M.A.(1977). Goodness of fit for the extreme-value distribution. Biometrika, 64, 583-588.
- Stephens, M.A.(1981). Further percentage points for Greenwood's statistic. J. Roy. Statist. Soc. A, 144, 364-366.
- Todorovic, P.(1979). A probabilistic approach to analysis and prediction of floods. Proc. Int. Statist. Inst., 48, 1, 113-124.
- Todorovic, P. and Rouselle, J.(1971). Some problems of flood analysis. Water Resour. Res., 7, 1144-1150.
- Todorovic, P. and Woolhiser, D.A.(1972). On the time when the extreme flood occurs. Water Resour. Res., 8, 1433-1438.
- Todorovic, P. and Zelenhasic, E.(1970). A stochastic model for flood analysis. Water Resour. Res., 6, 1641-1648.
- Weissman, I.(1983). Statistical estimation in extreme value theory. In J. Tiago de Oliveira (ed.), Statistical Extremes and Applications. D. Reidel, Dordrecht.
- Wrigley, J.(1982). Long-range dispersion of radioisotopes. PhD thesis, University of London.