ELSEVIER

Journal of Hazardous Materials A133 (2006) 46–52

Journal of Hazardous Materials

www.elsevier.com/locate/jhazmat

# Consequences of major accidents: Assessing the number of injured people

Andrea Ronza, Miguel Muñoz, Sergi Carol, Joaquim Casal *

*Centre d'Estudis del Risc Tecnològic (CERTEC)[1], Department of Chemical Engineering, Universitat Politècnica de Catalunya, Diagonal 647, 08028 Barcelona, Catalonia, Spain*

Received 19 May 2005; received in revised form 6 October 2005; accepted 7 October 2005
Available online 28 November 2005

**Abstract**

Quantitative risk assessment studies of accident scenarios usually involve estimating the number of fatalities that can be expected. The number of people injured, however, is seldom evaluated because it implies significant additional effort and often the information required to perform this evaluation is not available. However, the number of injured people can be very important for emergency planning, especially in relatively large accidents. In this paper, a set of 975 accidents were selected for analysis, with the aim of searching for a relationship between the number of people killed and the number of people injured. As the data were scattered, principal component analysis and clustering analysis were applied to identify the data subsets that could undergo a selective, specific statistical treatment. Further treatment of these subsets led to mathematical expressions that are used to estimate the probable number of injured people as a function of the number of fatalities for all accidents, as well as for gas cloud, fire and explosion events, respectively.


*Keywords:* Accident consequences; Fatalities; Injured people

## 1. Introduction

In quantitative risk assessment (QRA) of accident scenarios involving hazardous materials, rough hypotheses are often used to evaluate the magnitude of the consequences. These hypotheses are sometimes essential in order to draw iso-risk curves, as they complement the physical effect and vulnerability calculations. For instance, if more precise data are not available, it is widely accepted [1] that

- the indoor mortality rate in the case of toxic exposure is 10% of the corresponding outdoor rate;
- for blasts (vapour cloud explosions), the fraction of people that die outdoors and indoors is 100% if $\Delta P > 0.3$ bar. If 0.3 bar $> \Delta P > 0.1$ bar, nobody dies outdoors and 2.5% die indoors;
- about 10% of the houses outside the cloud and inside the $\Delta P = 0.1$ bar contour are severely damaged and about one in eight people in a severely damaged house is killed.

These hypotheses, which are mostly rules of thumb based on experience, do not necessarily agree with real data from actual cases, although they are normally of the same order of magnitude. Moreover, they save a lot of effort and reduce the time needed to carry out QRA.

Risk analysts, especially when major accident scenarios are concerned, seek to estimate/predict the overall number of people affected, i.e. fatalities, injuries and possibly evacuees. Standard QRA focuses mainly on calculating the number of fatal victims, as well as calculating the distances that define the areas to be evacuated. The number of injured people is seldom evaluated, as it would involve significant additional effort and in most cases little or no information is available.

These estimations are all based on calculating the effects of an accident, and they are independent from each other. Each estimation refers to a specific vulnerability criterion, and may include some of the shortcuts mentioned above. So far, no statistical relationship has been proposed to link the number of fatalities ($N_K$) in an accident involving hazardous materials to the number of injured people ($N_I$). Nevertheless, such a relationship could be especially useful in certain cases. For example, if $N_K$ is calculated for the area surrounding a plant where a major

* Corresponding author. Tel.: +34 93 401 67 04; fax: +34 93 401 19 32.
*E-mail address:* joaquim.casal@upc.edu (J. Casal).
[1] http://certec.upc.es/.

0304-3894/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhazmat.2005.10.011

**Nomenclature**

| | |
|---|---|
| $n$ | number of records |
| $N_I$ | number of injured people |
| $N_K$ | number of killed people |
| $R$ | correlation coefficient |
| $\Delta P$ | explosion overpressure (bar) |

*Greek letters*

| | |
|---|---|
| $\sigma$ | standard deviation |

accident may hypothetically happen, a shortcut for estimating the expected number of injured people would be useful for emergency planning purposes. An injured person, by definition, is someone who needs to be hospitalised quickly. From the point of view of emergency management, estimating the number of injured people quickly in the case of HazMat accidents is therefore of vital importance.

In this paper, we study the possible existence of a relationship between $N_K$ and $N_I$. A general equation is proposed, as well as specific equations for three accident scenarios (fires, gas clouds and explosions). While all these equations can be considered to be predictive, it is important to point out that they are probabilistic. In fact, as shown in the following sections, data variance is so high that it would be quite arbitrary to search for a deterministic relationship based on best fitting.

## 2. The data sample

We used a sample of 975 accident records to carry out the study. Data were taken from the February 2005 release of the Major Hazard Incident Data Service (MHIDAS). This database, which is maintained by the UK Health and Safety Executive, is one of the most frequently used databases for historical analysis of incidents involving dangerous substances [2]. It contains data on accidents that have occurred in 95 countries since the beginning of the 20 century. The data fields are codified in such a way as to make searching for data subsets quite easy. Although the version used includes 12,674 records, some information is unknown and some fields are empty. When an accident involves more than one dangerous substance, it is recorded more than once in the database, which is a problem when selecting data. In order to fix this problem and to use the data to their full potential, a new database was developed in Microsoft Access. This was programmed to import the incident data from the original database and to store them in a relational structure. The codification remained the same, but queries were now possible. These queries allowed the incidents to be grouped and filtered using either the MHIDAS code fields or user-designed code fields. For example, it is possible to classify the database's location code field according to the continent and to filter the incident using this new criterion. Moreover, new records or more reliable information can be introduced into the Microsoft Access database, which improves the incident data.

There are also other data bases on industrial accidents, which were not used in this study. They could be used in the future to enlarge somewhat the sample and confirm or improve the results.

Like almost every accident database, MHIDAS is somewhat selective in recording information. The more information that is available (to the data compilers), the more likely it is that the accident will appear in the database. For this reason, accidents in distant or developing countries are underrepresented, as well as accidents that happened decades ago (in this case accidents are only recorded if they were very severe).

Some measures were taken to reduce data bias. First, accidents that occurred before 1975 were excluded, since they happened in a technological setting that is very different from the present one. Safety measures, risk planning and urban planning before 1975 cannot be compared with the current situation. In addition, the following sets of accidents were excluded: (a) accidents due to handling and/or manufacturing of conventional explosives (TNT, dynamite, gunpowder, ammunition), as these materials are specifically designed to damage structures and harm people and so have a very high potential for causing severe accidents, (b) accidents due to sabotage, for the same reason and (c) the few accidents with more than 2000 fatalities, because of their exceptional and atypical nature (in fact, there is only one such event after 1974, that is, the 1984 Bhopal gas leak accident).

Two more restrictions were imposed on the data: (1) accidents for which $N_K = 0$ or $N_K$ is not defined were excluded and (2) accidents for which $N_I$ is not defined were also excluded. These restrictions were introduced to limit the scope of the work to those data needed to derive some correlation between $N_I$ and $N_K$, and also to avoid the bias caused by considering accidents with an undefined killed/injured record as events that are not fatal or not harmful.

The sample of 975 accidents was therefore defined as described above. It should be noted that the information used was not always accurate and that discrepancies were detected when the consequences of certain accidents were compared to other bibliographic sources. This inaccuracy is sometimes found, for example, in the number of injured people, when figures such as 201 are used to imply "more than 200". Nevertheless, using a large dataset balances the effect of the biased data.

In Fig. 1, the number of injured people is plotted against the number of fatalities. As both $N_I$ and $N_K$ are discrete variables, many points on the graph overlap (e.g. the pair $N_K = 1$, $N_I = 1$ recurs 61 times). To give a better idea of the data distribution, the thickness of the data points in the chart is proportional to the number of times they are repeated in the sample, as indicated in the legend. It can be seen that there is significant data scattering. For example, for $N_K = 1$ the number of injured people ranges from 0 to 600, and for $N_I = 100$, $N_K$ ranges from 1 to 60. This is why a log–log scale was used.

A general, overall trend is quite clear, i.e. $N_I$ increases with $N_K$. However, it is obvious that a basic statistical treatment aimed at a simple regression (e.g. based on least squares) of this cloud of points would be useless and completely unreliable. Therefore, a more complex treatment was applied using two multivariate

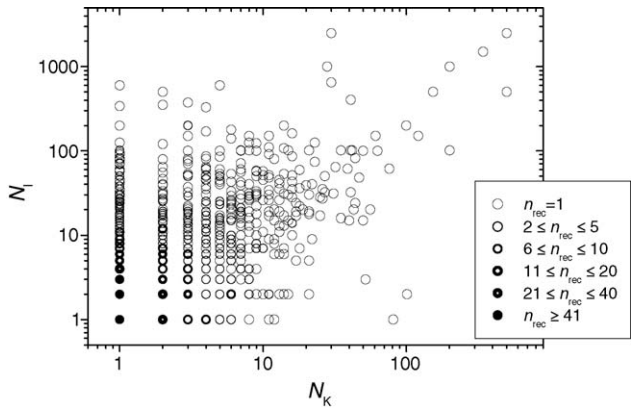


Fig. 1. Data sample. $n$ is the number of records for a given pair ($N_K$, $N_I$).

analysis techniques: principal component analysis and clustering.

## 3. Principal component analysis

Multivariate analysis is essentially aimed at finding subsets in a sample that correlate better than the sample as a whole, i.e. they have a smaller variance. New variables, which are listed in Table 1, were introduced to define these groups. Applying multivariate statistical analysis will make it clear whether these variables are really important in describing the data.

Principal component analysis (PCA) is a multivariate statistical tool that is used to describe a data sample as a function of a group of random variables [3]. The main aim of PCA is to select the variables that have the most influence on the variance of the sample. In fact, this technique allows us to describe the dependence of a variable on the other variables.

In the present case PCA was carried out on a system of 13 variables (see Table 1):

- the $N_I/N_K$ ratio
- $N_K$
- *Location*
- three accident types
- seven possible origins

The results of the PCA for the first two principal components are shown in Figs. 2 and 3. Fig. 2 is a loading plot, i.e. a representation of the variables in the form of vectors (eigenvectors of a correlation matrix). An ideal loading plot should be drawn in a space with $n$ dimensions, where $n$ is the number of variables. Each dimension represents a principal component of the data, that is, a "direction" along which data are distributed, in such a way that the first principal component has the largest covariance, the second principal component has the second largest covariance, etc. In the present case, it is impossible to represent a plot like this, so only the first two principal components are represented. Unfortunately, these two components do not describe a large part of the data variance. Six components are needed to account for 60% of the variance, while the first two components only describe 24%. On a very general level, the quite obvious conclusion can be drawn that the data are not describable in an unambiguous way by any combination of variables. In other words, none of the variables or combinations of variables are decisive in defining the relationship between $N_I$ and $N_K$.

Table 1
Variables used in the PCA analysis.

| Variable | Description | Type |
|---|---|---|
| $N_I$ | Number of injured people | Discrete (minimum = 0; maximum = 2500) |
| $N_K$ | Number of deaths | Discrete (minimum = 1; maximum = 501) |
| $N_I/N_K$ | Injured to killed ratio | Rational (minimum = 0; maximum = 601.0) |
| *Location* | Location (1, EU 15; 2, rest of the first world[a]; 3, rest of the world) | Discrete |
| Accident type | | |
| *Explosion* | Presence of an explosion | Logical (yes = 1; no = 0) |
| *Fire* | Presence of a fire | |
| *GasCloud* | Presence of a gas cloud | |
| Origin of accident | | |
| *DomCom* | Incident originated in domestic or commercial premises | |
| *Process* | Incident originated in a process plant | |
| *Storage* | Incident originated in a storage plant | |
| *Transfer* | Incident originated during loading or unloading | |
| *Transport* | Incident originated during transport of the material external to the plant, including pipelines | |
| *Warehouse* | Incident originated in a warehouse | |
| *Waste* | Waste storage or disposal areas, including settling ponds, material dumps, bulk waste files, but excluding materials being used in plant production | |

[a] Includes accidents occurred in Australia, Canada, Hong Kong, Iceland, Japan, Malta, New Zealand, Norway, Switzerland and the USA.

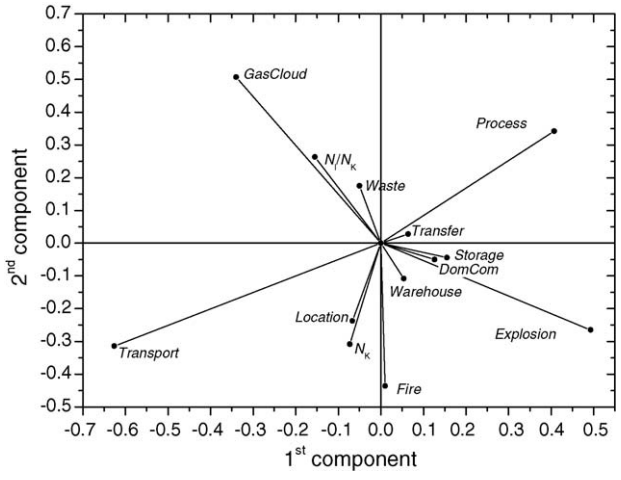


Fig. 2. Principal component analysis loading plot of the 13 variables.

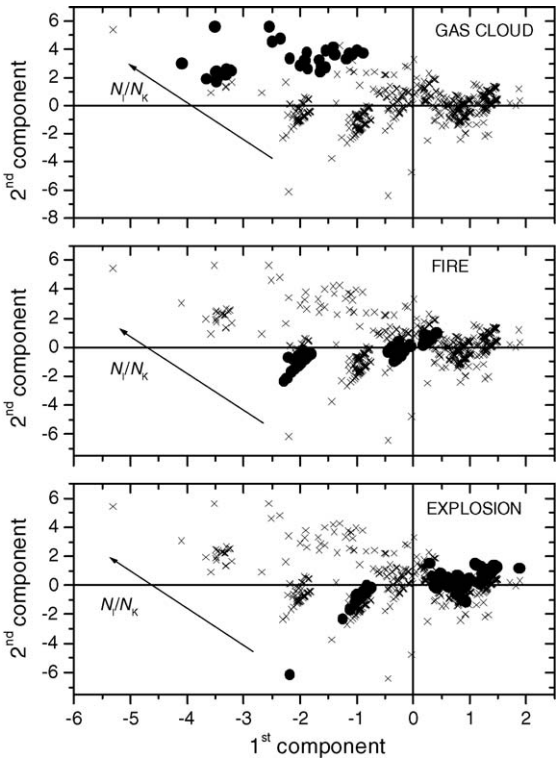


Fig. 3. Score representation for the PCA analysis.

In spite of this, some important conclusions can be inferred from the analysis. Fig. 2 makes it clear that

- The three variables describing the accident type are very important as descriptors of the variance of the system and correlate very well with $N_I/N_K$. It is apparent that gas cloud accidents tend to have a higher ratio of injured people to fatalities, while fire and especially explosion events show a tendency towards smaller than average ratios. The score plots in Fig. 3 confirm this. A score plot represents the data sample in the *n*-dimension space defined by the PCA. Score plots can be useful for supporting inferences derived from a loading plot. In Fig. 3, the gas cloud data subset is displaced in the direction of the $N_I/N_K$ eigenvector, fire is centred around the origin of the axes, while explosion shifts slightly in the opposite direction to $N_I/N_K$. For clarity's sake, the accidents for which both an explosion and a fire were reported are not shown on the graph; otherwise, a more explicit displacement in the opposite direction to $N_I/N_K$ would be seen. This means that if an accident involves a gas cloud, the ratio of injured people to fatalities is higher than in fire accidents, while, on average, explosion accidents show a tendency towards lower ratios.
- *Location* has little influence on $N_I/N_K$. There is a slight tendency for accidents that happen in developing countries to have a higher ratio of injured people to fatalities. *Location* does, however, have a clear effect on the number of fatalities ($N_K$), as can be seen in Fig. 2 (the vectors corresponding to *Location* and $N_K$ have the same direction and similar length). This pattern is confirmed by literature [4].
- Of the variables related to the origin of an accident, only *Transport* and *Process* significantly influence the variance of the system. *Waste*, *Storage*, *DomCom* and *Warehouse* play a secondary role, while *Transfer* has very little or no importance.
- As their eigenvectors are practically perpendicular to that of $N_I/N_K$, *Process* and *Transport* are very weakly related to the ratio of injured people to fatalities, despite their large statistical weight.

## 4. Clustering

PCA was complemented with a clustering analysis of the sample in order to better identify the sets of accidents that could undergo selective, specific statistical treatment. The analysis was performed using the Knowledge SEEKER (KS) Software. Clustering consists in designing a classification tree based on one variable by grouping data according to the values of the other variables. In the present case, data were evaluated based on the $N_I/N_K$ ratio and classified in structured subgroups defined by other variables associated with the incidents.

The technique used to design a classification tree is a refinement of the CHAID algorithm proposed by Kass [5] for determining the best multiway partitions of data based on a significance test. The algorithm recursively splits each subset or node into *k* new nodes, starting with all the observations at the initial node. The process continues until no more significant splits can be found. This method finds groups that maximise similarity within the groups and dissimilarity between the groups [5–7].

At each node, all predictor variables are considered in turn as candidates for splitting the node. The best split of each variable is found and the significance of that split is used to rank variables according to how well they split the node. For categorical associations the F statistic is used.

In order to perform data clustering, the sample was modified slightly to avoid possible ambiguities. Clustering works better when the variables mutually exclude each other. Therefore,

- Ninety accidents were not considered, because they are not classified under any accident type. These accidents are mostly events that MHIDAS describes as some kind of "release" of a hazardous material. This category is quite confusing, as a release or loss of containment in itself cannot harm people unless a toxic cloud is formed, the material ignites or an explosion is produced.
- As mentioned above, MHIDAS can assign an accident to more than one accident type. To avoid confusion, each accident was reassigned to a single accident type according to the following criterion: If *Explosion* = 1, then the accident was considered as an explosion, even if it was *also* a fire and/or a gas cloud. The remaining accidents were considered to be fires if *Fire* = 1, despite their possible additional status as gas clouds. Otherwise, the accidents were considered to be gas clouds.[2] In other

[2] This prioritisation (explosion > fire > gas cloud) is due to the average scale of severity of the accidents: explosions are generally more severe than fires, which in turn are more severe than gas clouds [4].

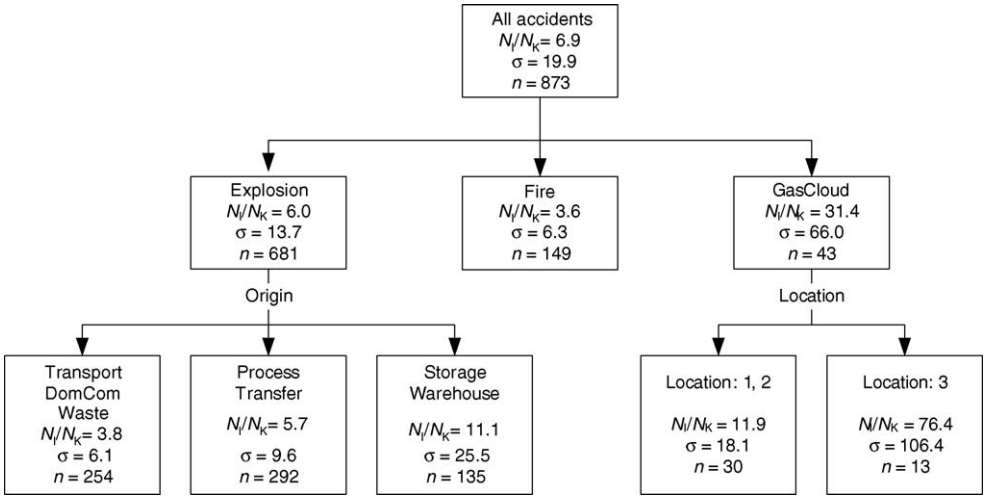


Fig. 4. Results of the clustering analysis.

words, only one variable was used to describe the accident type (instead of three, as in the PCA). The possible *values* of this variable are *Explosion*, *Fire* and *GasCloud*.

- Twelve more records were excluded to which MHIDAS does not assign any definite accident origin. The *Origin* of the accident (which is not affected by the ambiguity in the accident type) is then taken as an independent variable, whose possible values are *DomCom*, *Process*, *Storage*, *Transfer*, *Transport*, *Warehouse* and *Waste*.

Thus, 873 accidents were retained for clustering. The $N_I/N_K$ ratio was used as a dependent variable, while *Origin*, *Accident Type* and *Location* were considered as independent. Fig. 4 shows the results of the analysis. The most significant variable is *Accident Type*. This is evidence that the pattern revealed by the PCA is significant. Moreover, by looking at the average $N_I/N_K$ for each group it can be seen that the ratio is lower for fires and higher for gas clouds.

The other variables are of minor importance and only contribute to a slightly more detailed definition of the three major subgroups. In the case of the *Explosion* subset (681 cases, with an average value of $N_I/N_K = 6.0$), the *Origin* variable proved to be relevant. In addition, three more accident subsets were outlined: (a) *Transport*, *DomCom* and *Waste*, with an average value of $N_I/N_K = 3.8$, (b) *Process* and *Transfer*, with an average value of $N_I/N_K = 5.7$ and (c) *Storage* and *Warehouse*, with an average value of $N_I/N_K = 11$.

In the Fire sample (149 accidents), which had an average value of $N_I/N_K = 3.6$, no further specific sets were identified.

Finally, the gas cloud subset, which had a rather reduced number of accidents (43), gave the highest values for the ratio of injured people to fatalities: $N_I/N_K = 31$ on average. However, this figure is significantly influenced by the data corresponding to the accidents that occurred in developing countries (*Location* = 3), as can be seen in the classification tree. It was possible to obtain two nodes as a function of the level of development of the country where the accident happened. Thus, for the subset corresponding to categories 1 (EU 15) and 2 (rest of the first world), $N_I/N_K = 12$ on average, while for category 3 (a very limited sample) the mean $N_I/N_K$ is 76. This is further proof that accidents occurring in developing countries are, on average, more severe than in industrialised countries.

## 5. A simple statistical treatment of $N_I$ versus $N_K$

The first conclusion that can be drawn from the multivariate analysis is that there is no reason to correlate the data with a generalised linear model or other regressions, since not only does $N_I$ correlate with $N_K$ very weakly (see Fig. 1), but it correlates with the other variables even less. Attempts to define a linear model based on 10–12 variables give discouraging results: $R^2$ is never more than 0.4, this value is achieved with two independent variables ($N_K$ and *GasCloud*) and does not improve when more variables are introduced.

A simple statistical analysis aimed at revealing major patterns in the relationship between $N_I$ and $N_K$ was carried out using the datasets for which the multivariate analysis had indicated that a significant correlation existed, i.e. the three major data subgroups defined according to accident type and the whole sample overall. Given that $N_K$ is very dispersed, data were analysed by dividing the number of people killed into certain ranges. For each range the number of people injured was represented by several statistical parameters, such as mean, standard deviation, median and percentiles. The intervals were selected so that they contained at least 2% of the total number of records. As can be seen in Fig. 1, there are a greater number of records corresponding to accidents with a small number of people killed. In fact, the first nine groups have a width of just one point. For the rest of the ranges, the average number of people killed was evaluated based on a weighted average:

$$\bar{N}_{K,r} = \frac{\sum_{i=n_{r,\min}}^{n_{r,\max}} n_i N_{K,i}}{\sum_{i=n_{r,\min}}^{n_{r,\max}} n_i} \quad (1)$$

where $n_i$ is the number of records with $N_K = N_{K,i}$ and $n_{r,\min}$ and $n_{r,\max}$ are the left and right limits of the $r$ range.

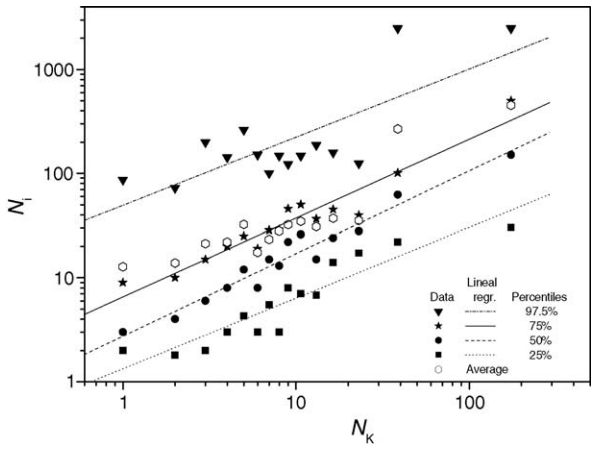


Fig. 5. Percentiles of the distribution and average, as calculated based on $N_K$ ranges.

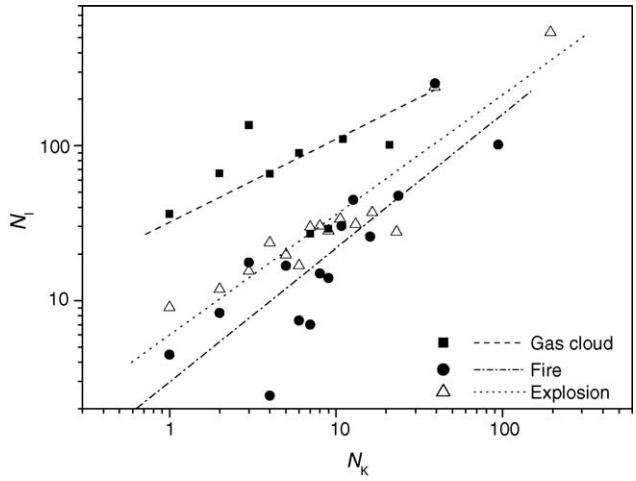


Fig. 6. Fitting regressions for $N_I$ vs. $N_K$ in the case of gas cloud, fire and explosion accidents.

Fig. 5 shows the 25%, 50% (median), 75% and 97.5% percentiles and the average of the $N_I$ distributions for each $N_K$ range and for all the data. The abscissae of each point in the graph are the weighted averages of the corresponding range (according to Eq. (1)). The lines resulting from the linear regression of each percentile are also shown.

Percentiles give an idea of the probability of a certain number of people being injured in relation to the number of fatalities in an accident. In general, percentiles are a good tool for describing data distributions that are unresponsive to regressions. Firstly, percentiles give evidence of the variance of the data. This parameter grows exponentially with the number of people killed. In fact, the lines resulting from the percentile regression are practically parallel to one another in a log–log scale. Moreover, this shows that the shape of the distribution in each range is similar: distributions are asymmetric, their median is nearer to the minimum and they are dispersed widely for high numbers of injured people.

The 50% and 75% percentiles fit into straight lines very well, while obviously the peripheral 25% and 97.5% percentiles are more irregularly scattered. By observing the 97.5% percentile, it can be inferred that, in an industrial accident with less than 25 deaths, it is highly unlikely that the number of injured people will exceed 150.

Due to the aforementioned asymmetry of the distribution, the mean values for each range correspond almost perfectly with those of the 75% percentile. There is only one exception: the $28 < N_K < 50$ range, for which the average number of injuries is unusually high. This is attributable to the Toulouse accident of September 2001, which has a very high $N_I/N_K$ ratio (MHIDAS reports 2500 people injured and 30 killed in this accident).

In Fig. 6 mean values for $N_I$ are plotted for each $N_K$ range and for the three accident types. The corresponding linear regressions are also shown. As mentioned before, the accidents that involve gas clouds result in the greatest number of people injured, followed by explosion and fire accidents, respectively. The fitting lines for the three subsets and the whole dataset, have the following form (validity intervals must be considered as indicative):

$$\text{Gas clouds} \quad N_I = 34 N_K^{0.54}, \quad \text{for } 1 < N_K < 30 \tag{2}$$

$$\text{Fires} \quad N_I = 3 N_K^{0.86}, \quad \text{for } 1 < N_K < 100 \tag{3}$$

$$\text{Explosions} \quad N_I = 7 N_K^{0.76}, \quad \text{for } 1 < N_K < 200 \tag{4}$$

$$\text{All accidents} \quad N_I = 6 N_K^{0.77}, \quad \text{for } 1 < N_K < 200 \tag{5}$$

An interesting feature is that, in spite of having the highest $N_I/N_K$ ratio (Fig. 6), gas cloud accidents show a tendency for this ratio to progressively decrease.

## 6. Conclusions

In QRA, estimating the number of injured people requires a significant amount of additional work. Furthermore, a standard QRA approach with probit estimations often entails rough approximations, as the real figures depend on boundary conditions that are sometimes unpredictable. Therefore, in the same way that approximate QRA values are applied when exact information is lacking, an approximate criterion for evaluating $N_I$ as a function of $N_K$ would be very useful. This criterion was identified based on a historical analysis, i.e. by analyzing the "experimental data" that are available.

The sample used in this survey follows a trend of $N_I$ increasing with $N_K$. However, due to the degree of data dispersion, a conventional statistical correlation approach would be useless and certainly unreliable. This is why more sophisticated methods were applied to identify the data subsets that can be correlated fairly accurately. We applied two multivariate analysis techniques: principal component analysis and clustering.

These procedures have shown that the variables describing the accident type have the most influence on the ratio of injured people to fatalities. As for the rest of the variables, the geographical location is not significant for $N_I/N_K$ (even if it has a clear influence on the number of fatalities, as accidents occurring in developing countries are more severe). Both the principal component analysis and clustering reveal a pattern according to

which the highest values of $N_I/N_K$ pertain to gas cloud accidents, followed by fires and then explosions.

This is confirmed by plotting the mean $N_I$ data against proper ranges of $N_K$ for the three different accident types.

A simple correlation (Eq. (5)) that is independent of accident type and estimates the mean number of injured people was obtained. Due to the particular form of the data distribution, this equation returns $N_I$ values which are not expected to be exceeded (with a probability of 75%). The same was done for the three subsets associated with accident types, which resulted in the corresponding regressions (Eqs. (2)–(4) for gas cloud, explosion and fire accidents). Since in this case there are significantly less data, these three equations were obtained by correlating the means of data ranges, instead of the 75% percentile.

These equations can be helpful in the case of fatal accidents involving hazardous materials. By giving an idea of the number of people that are expected to be hospitalised, hospital facilities and the response to the emergency in general can be better managed. Moreover, using these criteria a priori is a way of saving time in estimating the number of injured people without resorting to effects calculations and probit techniques.

## References

[1] TNO, Guidelines for Quantitative Risk Assessment (CPR18E. Purple Book), TNO, Apeldoorn, 1999.
[2] Health and Safety Executive, MHIDAS, February 2005 [CD-ROM], Health and Safety Executive, London, 2005.
[3] G.H. Dunteman, Principal Components Analysis, Sage, Newbury Park, CA, 1989.
[4] S. Carol, J.A. Vílchez, J. Casal, Study of the severity of industrial accidents with hazardous substances by historical analysis, J. Loss Prev. Process Ind. 15 (6) (2002) 517–524.
[5] G.V. Kass, An explanatory technique for investigating large quantities of categorical data, Appl. Stat. 29 (2) (1980) 119–127.
[6] G.V. Kass, Significance testing in automatic interaction detection (A.I.D.), Appl. Stat. 24 (2) (1975) 178–189.
[7] D. Biggs, B. De Ville, E. Suen, A method of choosing multiway partitions for classification and decision trees, J. Appl. Stat. 18 (1) (1991) 49–62.