# USE OF CONTINGENCY TABLES TO VALUE VARIABLES FOR SPATIAL MODELS

**Olga Špatenková and Jukka Matthias Krisp**

Laboratory of Geoinformation and Positioning Technology
Department of Surveying
Helsinki University of Technology
P.O.Box 1200, FIN-02015, Finland
olga.spatenkova@hut.fi, jukka.krisp@hut.fi
http://www.tkk.fi/Units/Cartography/research/rem/index.html

**KEY WORDS:** data mining, contingency tables, spatial modeling, risk model, fire & rescue services

**ABSTRACT:**

An expressive and comprehensive situation picture is necessary for a reliable decision making in various application fields. The domain knowledge, however, is often too complex to be handled individually, and thus geographical information systems (GIS) with powerful modeling tools are nowadays availed to support the process. Data to be considered are becoming available at an increasing speed and level of detail, thus the challenge of obtaining a useful resulting model lies in utilization of suitable methods. In our research, we deal with a systematic risk assessment model for Helsinki Fire&Rescue services. The model shall serve as a basis for preparedness of fire brigades. In this paper, we aim to use contingency tables, which are known from statistics, to assist valuing new variables for the developing risk model. In the case study, we analyze spatial relationships between the incident data points and distribution of population age. Derived information shall be implemented into the spatial model, which is the basis for further risk modeling process. The methods for the analysis of spatial data suggested in this paper support reliability of the risk model and advance understanding of how GIScientist can contribute to the process of decision making.

## 1. INTRODUCTION

The amount of data collected in spatial information science and social science is increasing rapidly and very often the researchers who acquire data are interested in the relationships of associations between different variables. Thus, a knowledge discovery and data mining as techniques established in the domain of information technology have found a new application field.

The general aim of our research is to assist the Fire&Rescue services in their planning and preparation procedures within the Metropolitan area of Helsinki, Finland. The starting point is an already implemented analysis tool, which generates a risk zone map. The risk level refers to the required preparedness of fire brigades in a sense of response time, and is calculated on the basis of: (1) the population density; (2) the floor area in square meters, and (3) an index for the probability of traffic accidents. Clearly, the model is rather simple; our goal is to investigate other variables that might contribute to enhancement of this model.

To evaluate the significance of potential variables that might be included into the risk model, we want to assess how they match spatially (and temporally) with the incidents. Our intention is therefore to use statistical methods including the spatial component to explore and identify relevant variables. In this paper, we propose the use of contingency tables and measures of association, and try to evaluate the usefulness of this method for a development of spatial analysis models.

One of the relevant questions for the Fire&Rescue services is "How does age of the population affect occurrence of particular incident types?". Thus, in the case study, we exemplarily analyze the incident data points in relation to the population age distribution.

Contingency tables as a background theory of this paper are presented in the next Section 2 followed by our proposal for their utilization to enhance spatial modeling in the Section 3. The application of the approach is described in the case study in the Section 4. The utilized method is discussed in the Section 5, and finally the results are concluded in the Section 6.

## 2. CONTINGENCY TABLES

The term contingency table was first used in (Pearson, 1904) in connection with a generalized theory of association. Nowadays, contingency tables are extensively used in statistics (Everitt, 1992, Agresti, 2002) to analyze relationships between two or more categorical variables that are exclusive (categories do not overlap) and exhaustive (categories include all possibilities). They aggregate the data according to the studied variables, and in this way enable to get a deeper insight into the relationships between them.

Assume that X and Y are discrete variables that can be assigned a finite number of values $1, \ldots, r$ and $1, \ldots, c$ respectively. The contingency table $(n_{ij})$ can be expressed in a form of matrix as shown in the Table 1. The table fields $n_{ij}$ represent empirical frequencies of an existing particular combination of the variable values in the dataset. The corresponding column sums $n_{.j}$ and row sums $n_{i.}$ are called marginal frequencies. (Anděl, 2002)

| $X/Y$ | $1$ | $\ldots$ | $c$ | $\sum$ |
|---|---|---|---|---|
| $1$ | $n_{11}$ | $\ldots$ | $n_{1c}$ | $n_{1.}$ |
| $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ |
| $r$ | $n_{r1}$ | $\ldots$ | $n_{rc}$ | $n_{r.}$ |
| $\sum$ | $n_{.1}$ | $\ldots$ | $n_{.c}$ | $n$ |

Table 1: A contingency table

For easier understanding, let us consider the following example: the color of eyes and hair was recorded for a population of 6800 men. There are four hair color values: blond, brown, black and red, and three eyes color values: blue, gray or green and brown. The distribution of the observed variables is expressed in the contingency table in the Table 2.

| Eyes/Hair | Blond | Brown | Black | Red | $\sum$ |
|---|---|---|---|---|---|
| **Blue** | 1768 | 807 | 189 | 47 | 2811 |
| **Gray or Green** | 946 | 1387 | 746 | 53 | 3132 |
| **Brown** | 115 | 438 | 228 | 16 | 857 |
| $\sum$ | 2829 | 2632 | 1223 | 116 | 6800 |

Table 2: Example of a contingency table

Testing of independence of the two variables is often task of the analysis. The two variables are independent iff the probability distribution $p_{ij} = p_{i.}p_{.j}$ is true for all the couples $i = 1, \ldots, r$ and $j = 1, \ldots, c$, thus iff the conditional probability is equal to the unconditional probability.

If we replace the notion of probability by the relative frequencies: $p'_{ij} = \frac{n_{ij}}{n}$, we get approximately: $\frac{n_{ij}}{n} \sim \frac{n_{i.}}{n}\frac{n_{.j}}{n}$. The question is how large difference between actual and expected cell counts can be considered as approximately equal. The answer is in the Pearson's chi-square test. We proceed further to absolute frequencies: $(\frac{n_{ij}}{n} - \frac{n_{i.}}{n}\frac{n_{.j}}{n})^2 \rightarrow (n_{ij} - \frac{n_{i.}n_{.j}}{n})^2$. According to the probability theory, large expected cell counts vary more than the small ones, thus we have to account the difference in expected cell counts by weighting. The variable

$$\chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}} \qquad (1)$$

has the $\chi^2$ distribution with $(r-1)(c-1)$ degrees of freedom. Thus, when we get $\chi_0^2 \geq \chi^2_{(r-1)(c-1)}(\alpha)$, where $\alpha$ is the level of significance, we refuse the null hypothesis about the independence of the two variables. (Anděl, 2002)

To compare associations between several contingency tables, various measures based on $\chi_0^2$ have been proposed. The most suitable norming was proposed by Cramér:

$$V = \sqrt{\frac{\chi_0^2}{n \quad \{min[(r-1),(c-1)]\}}}. \qquad (2)$$

The Cramér's $V$ is suitable for general (asymmetric) tables. The values lie between 0 and 1, while the maximum is reached for a complete association. (Bishop et al., 1975)

## 3. PROPOSED APPROACH TO VALUE MODEL VARIABLES

A reliable modeling of any phenomenon can be only based on proper understanding of relationships between all involved events, objects and its features. In case the causality is in focus, dependency relationships are of the main interest. The $\chi^2$ statistic reflects the difference between the observed and expected cell counts in the contingency table assuming the independence of the analyzed variables. High values of $\chi^2$ indicate that the variables are not independent: the higher the number, the stronger the relationship. Hence, we use measures based on $\chi^2$ to value the importance of the potential variables for the new risk model.

Further study of contingency tables may reveal more details describing the relationships between the studied variables. Besides the strength of the dependence, we can also specify the table cells invoking that: "What particular combination of variables values is the most critical from the dependency point of view?". The discovered knowledge should be then reflected by the model.

We believe that by performing the analysis by using contingency tables and measures of association, we can capture the appropriate relationships in the studied phenomenon, which must be thoroughly discussed with the domain experts though. Then, involving the variables selected on the basis of statistics into the model will increase reliability of the application.

## 4. CASE STUDY

The proposed approach was applied to value the possible variables for the risk model for Helsinki Fire&Rescue services. The incidents dataset was analyzed in connection to the population age distribution.

### 4.1 Data Description

Finland is one of the countries in which spatial data collection, digital databases, and the information society in general is well developed and organized, but this is no guarantee that information is available and interoperable. Organizational boundaries are clearly visible also in the domain of emergency planning (Jolma et al., 2004). We do have to keep in mind that the unequal quality of spatial data requires individual models and scales for a risk analysis in each community.

Municipalities in Finland are required to collect register data on their population, buildings and land use plans. Because of this, the Helsinki Metropolitan Area Council (YTV) has been working since 1997 on the production of a data package (named SeutuCD) which covers the whole metropolitan area. This data package is gathered from the municipal register and other sources, and it is updated once a year. It includes register data on buildings; metadata information and land use plans as well as the enterprises and agencies located within the metropolitan area. (YTV, 1999) Within the building register layer each of the buildings acts as a spatial object, which has coordinates and a set of attributes. This enables the analysis to be done despite of municipal boundaries in the metropolitan area. The attribute we are interested in is the age of people, which is related to the ones who are registered in a particular building. An interesting pattern of the spatial distribution of the population average age groups entitling us for further analysis of this information is shown in Figure1. The population data originates from the Finnish Population Register Center. This population information system contains information for the whole of Finland on all Finnish citizens and foreigners who are permanently residing in Finland.

Additionally, we have acquired a dataset from the Fire&Rescue services in Helsinki. This dataset contains records of all the fire alarms, rescue missions and automated fire alarm systems missions within Helsinki city area for the years 2001 to 2003. The material provided includes selected information, such as mission type codes, dates, addresses and spatial coordinates. From the incident data we select the most frequent incident types that were recorded more than 50 times during the year 2003 (see the Figure 2).

### 4.2 Data Preprocessing

We are not dealing with the automated fire alarms (coded with A10), as they have a very high significance and disturb the analysis of other data. As we are looking for the causality relationships and focus now on the population age data, we remove also the storm damage incidents from the analysis. The most frequent accidents include for example: T111 - fire in block building, P410 - oil harm on the solid ground, P761 - non-urgent assistance, etc.

The population records refer to the age of the inhabitants registered in a particular building. Based on the amount of the population, the average age for a particular building is calculated. In

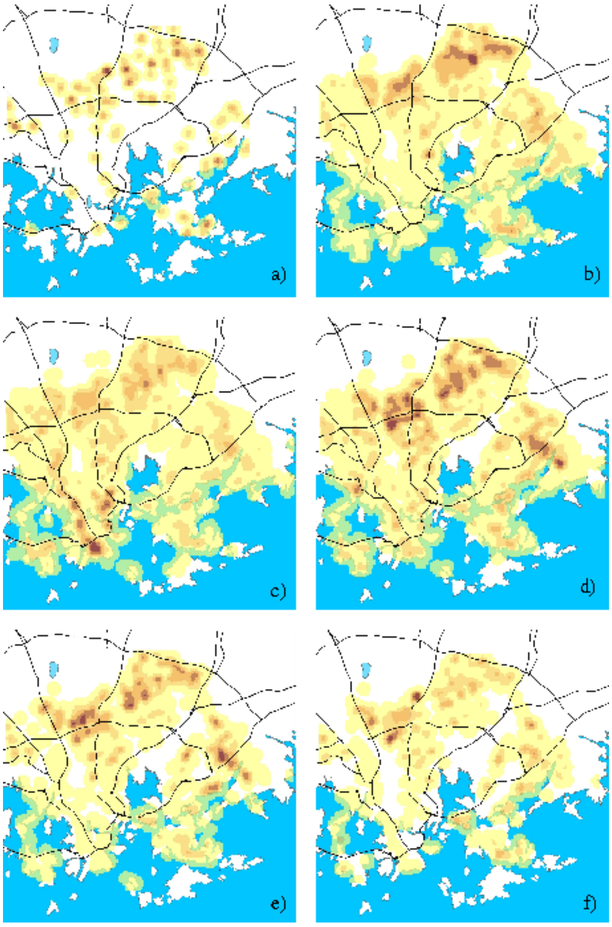

Figure 1: Spatial distribution of the population age groups in Helsinki: a) 8-29, b) 30-37, c) 38-45, d) 46- 52, e)53-63, f) 64-86; dark color indicates high density, light color indicates low density

the close neighborhood, incidents might be related to the building and inhabitants of this building. Thus, information about the average age is related to the incidents based on proximity; distances between incidents and nearest buildings are within 10 meters in most cases. The average inhabitant age in relation to occurrence of the most frequent accidents is classified into six categories based on natural breaks classification.

### 4.3 Results

By organizing the studied variables into the contingency tables, we have discovered some discrepancies in the data. Although the colleagues form the Fire&Rescue services were sure about the correctness of their dataset, simple sorting according to the temporal attribute have unveiled missing records of certain incident types for several months. This experience demonstrates the significance of data verification especially for the analysis based on the data, and need of iterative discussions between the parties involved.

The Table 3 describes the distribution of the incident types in relation to the average age of population in the neighborhood. The observed pattern was compared to expected values based on the $\chi^2$ statistic. The calculated value was $150.0$, which was higher than the $\chi^2_{110}(0.05) = 135.5$, thus we refuse the null hypothesis about independence between the two variables.

We have expressed the strength of the relationship between incidents and population age by the Cramér's $V$: $V = 0.109$. This measure can be then used to value various potential variables and compare their importance. To illustrate the process, we have calculated the $\chi^2$ statistics and Cramér's $V$ for temporal attributes of the incident dataset. Comparing the incident types and days of the week they have happened, we get $\chi^2 = 179.7$, which was higher than $\chi^2_{132}(0.05) = 159.8$, $V = 0.109$. Similarly, for incident types and daytime, we get $\chi^2 = 308.6$, which was higher than $\chi^2_{66}(0.05) = 86$, $V = 0.201$. The derived statistics indicate an association between the incidents and the studied variables. The higher the Cramér's $V$, the stronger evidence against the null hypothesis of independence. Thus, we could sort the studied variables according to Cramér's $V$: the daytime shows the strongest association to the incident types, day of the week and age of population are on the same level of importance.

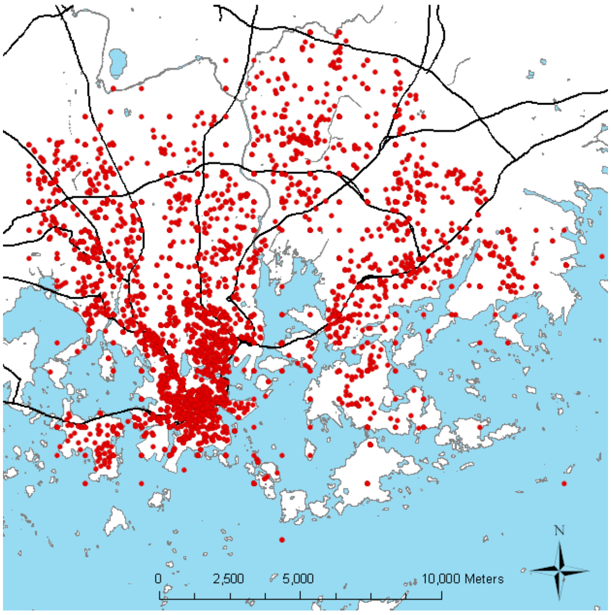

Figure 2: Distribution of the most frequent incident types

Further, by detailed study of the Table 4, we can trace more details. For example the highest number, and therefore the table field with the biggest influence on the relationships between the variables, corresponds to the incident type P732 (which is an ani-

| Inc | 8-29 | 30-37 | 38-45 | 46-52 | 53-63 | 64-86 | $\sum$ |
|---|---|---|---|---|---|---|---|
| P211 | 24 | 26 | 36 | 15 | 10 | 1 | 112 |
| P221 | 15 | 16 | 17 | 14 | 2 | 3 | 67 |
| P310 | 11 | 24 | 32 | 12 | 7 | 2 | 88 |
| P410 | 20 | 47 | 61 | 31 | 20 | 1 | 180 |
| P470 | 9 | 19 | 21 | 20 | 5 | 2 | 76 |
| P520 | 10 | 25 | 21 | 17 | 2 | 0 | 75 |
| P730 | 13 | 21 | 30 | 12 | 5 | 4 | 85 |
| P732 | 18 | 16 | 15 | 5 | 10 | 1 | 65 |
| P760 | 14 | 32 | 32 | 14 | 10 | 6 | 108 |
| P761 | 21 | 56 | 46 | 33 | 19 | 5 | 180 |
| P790 | 5 | 27 | 25 | 16 | 10 | 2 | 85 |
| T110 | 7 | 19 | 31 | 7 | 4 | 0 | 68 |
| T111 | 41 | 68 | 74 | 49 | 18 | 6 | 256 |
| T213 | 12 | 13 | 23 | 14 | 2 | 1 | 65 |
| T221 | 10 | 33 | 26 | 10 | 12 | 3 | 94 |
| T320 | 11 | 24 | 15 | 4 | 6 | 1 | 61 |
| T430 | 12 | 20 | 24 | 9 | 7 | 1 | 73 |
| T490 | 24 | 43 | 60 | 32 | 9 | 3 | 171 |
| T491 | 9 | 31 | 46 | 13 | 11 | 3 | 113 |
| T621 | 17 | 27 | 34 | 12 | 13 | 3 | 106 |
| T711 | 20 | 42 | 40 | 24 | 15 | 3 | 144 |
| T811 | 19 | 50 | 67 | 27 | 10 | 4 | 177 |
| T821 | 9 | 28 | 37 | 5 | 2 | 4 | 85 |
| $\sum$ | 351 | 707 | 813 | 395 | 209 | 59 | 2534 |

Table 3: Contingency table for incidents and average age of population

| Inc | 8-29 | 30-37 | 38-45 | 46-52 | 53-63 | 64-86 | $\sum$ |
|---|---|---|---|---|---|---|---|
| P211 | 4,6 | 0,9 | 0,0 | 0,3 | 0,1 | 1,0 | 6,9 |
| P221 | 3,5 | 0,4 | 0,9 | 1,2 | 2,2 | 1,3 | 9,6 |
| P310 | 0,1 | 0,0 | 0,5 | 0,2 | 0,0 | 0,0 | 0,9 |
| P410 | 1,0 | 0,2 | 0,2 | 0,3 | 1,8 | 2,4 | 5,9 |
| P470 | 0,2 | 0,2 | 0,5 | 5,6 | 0,3 | 0,0 | 6,8 |
| P520 | 0,0 | 0,8 | 0,4 | 2,4 | 2,8 | 1,7 | 8,2 |
| P730 | 0,1 | 0,3 | 0,3 | 0,1 | 0,6 | 2,1 | 3,5 |
| P732 | 9,0 | 0,3 | 1,6 | 2,6 | 4,0 | 0,2 | 17,7 |
| P760 | 0,1 | 0,1 | 0,2 | 0,5 | 0,1 | 4,8 | 5,8 |
| P761 | 0,6 | 0,7 | 2,4 | 0,9 | 1,2 | 0,2 | 5,9 |
| P790 | 3,9 | 0,5 | 0,2 | 0,6 | 1,3 | 0,0 | 6,4 |
| T110 | 0,6 | 0,0 | 3,9 | 1,2 | 0,5 | 1,6 | 7,8 |
| T111 | 0,9 | 0,2 | 0,8 | 2,1 | 0,5 | 0,0 | 4,4 |
| T213 | 1,0 | 1,5 | 0,2 | 1,5 | 2,1 | 0,2 | 6,4 |
| T221 | 0,7 | 1,7 | 0,6 | 1,5 | 2,3 | 0,3 | 7,1 |
| T320 | 0,8 | 2,9 | 1,1 | 3,2 | 0,2 | 0,1 | 8,2 |
| T430 | 0,4 | 0,0 | 0,0 | 0,5 | 0,2 | 0,3 | 1,3 |
| T490 | 0,0 | 0,5 | 0,5 | 1,1 | 1,8 | 0,2 | 4,1 |
| T491 | 2,8 | 0,0 | 2,6 | 1,2 | 0,3 | 0,1 | 7,0 |
| T621 | 0,4 | 0,2 | 0,0 | 1,2 | 2,1 | 0,1 | 4,0 |
| T711 | 0,0 | 0,1 | 0,8 | 0,1 | 0,8 | 0,0 | 1,9 |
| T811 | 1,2 | 0,0 | 1,8 | 0,0 | 1,4 | 0,0 | 4,6 |
| T821 | 0,7 | 0,8 | 3,5 | 5,1 | 3,6 | 2,1 | 15,7 |
| $\sum$ | 32,6 | 12,1 | 23,0 | 33,5 | 30,1 | 18,7 | 150,0 |

Table 4: Calculation of the $\chi^2$ for incidents and average age of population

mal accident) and the youngest population group, perhaps young families with children. High value of $\chi^2$ for the same incident can be also found in relation with 53-63 age group. By comparing the observed and expected frequencies, we can also find whether the particular combination happens more or less often than we would expect. In both of the cases above, the observed pattern is higher than expected values, thus the incidents happen more often in reality. Similarly, we can observe that the oldest population group needs more often urgent assistance (P760), or surprisingly that a high amount of car accidents happens in proximity to the youngest population group.

## 5. DISCUSSION

The analysis using contingency tables provides a simple and fast insight into the dataset from a point of view of two particular variables in question. This method is also relatively easy to interpret, which minimizes misunderstandings in the communication between the involved parties that have often very different background. Further, the method is based on solid statistical grounds assuring the credibility of the results.

A disadvantage within the methods applied and the resulting outcome is the input data. Since we did use the input data, which was provided by the participating partners (Fire&Rescue services), we did not have sufficient information about its quality. Besides the incompleteness, in some individual cases the locations for the incidents seem to have an error in the geocoding process. If this error is too large it will influence the proximity calculations to the housing points and give an error to our results. One way to deal with this problem is to use density surfaces, which would eliminate the error to a certain degree. In the future research, we will need to replace these data points and perhaps use a different incident dataset.

Other potential variables for the new risk model should be analyzed in the same way. Apparently, we might put an emphasis on the importance of the temporal aspect. Since every incident data is recorded with a time stamp, it might result in significant relationships to other spatial variables that change over time (e.g. population distribution). Ongoing research investigates the significance of a changing spatio-temporal population model to support risk assessment and damage analysis for decision-making (Ahola et al., 2007).

## 6. CONCLUSIONS

Contingency tables and $\chi^2$-based measures of association are well established statistical methods for analyzing relationships in the data, suitable for categorical variables. Applied to the risk model, the proposed approach has been found useful in finding new relevant variables.

However, to draw reliable conclusions, the method requires a quality input data. Incompleteness and geocoding problems of the incident records, which have arisen during the analysis, have to be solved in the future perhaps by using different dataset. In addition, the available census data should be updated.

Assuming that the census data is accurate, the age of population can be a relevant variable in the analysis for the cause of certain accident types. Nevertheless, our results are in need of further interpretation by the Helsinki Fire&Rescue services before they can be used in a risk model. This analysis is only one step in a process that is often very complex.

The next research has to consider the temporal aspect. Further integration of a time variable (time of the year, special holidays, etc.) seems to be essential when relating the population distribution with regard to age to specific incident types. Furthermore we should aim to integrate more data from different resources and perhaps from a longer period of time.

## ACKNOWLEDGEMENTS

The authors would like to thank the Finnish Fire&Rescue services for cooperation and supporting the research.

## REFERENCES

Agresti, A., 2002. *Categorical Data Analysis*. John Wiley & Sons, Inc.

Ahola, T., Virrantaus, K., Krisp, J. M. and Hunter, G., 2007. A spatio-temporal population model to support risk assessment and damage analysis for decision-making. *International Journal of Geographical Information Science*, in review.

Anděl, J., 2002. *Základy matematické statistiky*. MFF UK Praha, pp. 283–317.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W., 1975. *Discrete Multivariate Analysis: Theory and Practice*. The MIT Press, Cambridge, pp. 385–389.

Everitt, B. S., 1992. *The Analysis of Contingency Tables*. Chapman&Hall/CRC.

Jolma, A., Virrantaus, K. and Krisp, J. M., 2004. Review of geoinformatical methods for a city safety information system. In: *The 9th International Symposium on Environmental Software Systems, Harrisonburg, VA, USA.*

Pearson, K., 1904. On the theory of contingency and its relation to association and normal correlationthe analysis of contingency tables. Foreword to Drapers Company research memoirs: Biometric series I.

YTV, 1999. Establishments in the Metropolitan Area 1999. Helsinki Metropolitan Area Council (YTV), Helsinki, Finland.