Assignment of positionally uncertain point events to regions

Tonio Fincke, Jochen Schiewe, and Karl-Peter Traub
Lab for Geoinformatics and Geovisualization (g²lab)
HafenCity University Hamburg
Hebebrandstraße 1, 22297 Hamburg, Germany
{tonio.fincke | jochen.schiewe | karl-peter.traub} @hcu-hamburg.de
http://www.hcu-hamburg.de/geomatik/g2lab

1. INTRODUCTION

Although it is often not considered explicitly, positional uncertainty is a frequent issue in spatial analysis. It denotes the phenomenon that the spatial coordinates of an object do not indicate, but only approximate the actual position. As with other kinds of uncertainties, positional uncertainty can be introduced through every step of working with data. Uncertainty can be described or modeled with confidence intervals or probability density functions (Morgan, 1990; MacEachren, 2005). For modeling spatial uncertainty, buffered regions or fuzzy borders can be draped around the spatial feature. An example is the 'Egg-Yolk' model, where regions with indeterminate boundaries are represented by concentric regions (Cohn, 1996).

Due to their special characteristics, spatial uncertainties and methods for their consideration continue to receive specific attention (Zhao, 2009; Navratil, 2008; Fisher, 2000; Shokri, 2006; Hession, 2006). For example, the 9-intersection model (Egenhofer, 1991) has been extended to uncertain geographic objects (Clementini, 1996). Locational or positional uncertainties pose special challenges (Murray, 2003; Grubesic, 2004).

In a specific case, we want to count the number of emergency rescues which have occurred in several distinct city districts. The fire department of Hamburg provided data about emergencies that happened between 2005 and 2008 around a rescue service station (for former investigations of emergency rescues in Hamburg see Traub, 2003). Due to data privacy we were only given the name of the street in which the emergency occurred, not the exact geographical location. Thus, we have to deal with several events that have happened along a line feature. The positional uncertainty consists in that a road is not necessarily fully contained within a district, but instead may run through various districts. Hence, the amount of emergencies which have occurred in a district cannot be determined unambiguously. The overall goal of this paper is to find a method to estimate the total number of emergency events per district and describe the uncertainties of the estimated values.

Prior investigations about uncertainties concerning line features have been performed by Wu, 2008, who examined the characteristics of lines with uncertain endpoints, and by Reis, 2006, who presented two different approaches to expand the 9-intersection model to uncertain line relationships.

This work is related to the topic of line clipping. In line clipping, line features are clipped by reference to the intersection of another line feature or a polygon feature's boundary (e.g., Cyrus, 1978). The results of such an operation are feature segments. The issue here is the treatment of the thematic attributes after clipping. Some attributes, e.g., the name of a road, remain the same for each resulting feature segment. Others can be derived from the new geometrical or topological relations, e.g., the length of the road segment. However, some attributes cannot be easily assigned to the segments. The occurrence of an emergency is such an attribute. It is clear that, if the emergency happened along a road, it must have happened along exactly one of its segments. The emergency counter for one of the segments would need to be increased by 1 for one segment, while the counters for the other segments would not be modified. In this paper, we present an approach where we use not only one, but several counters per district, which are increased according to different rules.

Another related research field is raster coding. In raster coding, the values of the raster cells are assigned during a scanning process by a vector to raster conversion. There exist several methods to pick the value which then would represent the cell, e.g., choosing the most dominant value or the one which is situated at the center (Davis, 2001). Each of these methods causes a generalization and thus introduces uncertainty into the model.

2. METHODOLOGY

The data set under analysis consists of a total of 32,984 emergency cases that occurred between 2005 and 2008 around the rescue station Bergedorf in Hamburg. This area extends over ten districts (Lohbrügge, Bergedorf, Curslack, Altengamme, Neuengamme, Kirchwerder, Ochsenwerder, Reitbrook, Allermöhe, and Billwerder), which vary significantly in their demographic characteristics.

To eventually count the emergencies with various counters, we first had to prepare the data. We intersected the road features with the shape of the districts to receive information about which roads run through which districts. After that we calculated the length of the individual road segments per district. In order to come up with an assignment model that also serves for the consideration of uncertainties, we use a set of five counters, each with different counting rules. For every district there is one such counter set. Every emergency event can possibly affect the counters of various districts. Some of these counters do not necessarily increase the counter by 1, but by a value between 0 and 1.

The first counter is named *peC*, which is short for pessimistic counter. It is increased by 1 for a district if and only if the line feature is completely contained in the polygon feature. If the road runs through more than one district, no *peC* of any district is increased. Let *i* be one specific district, *x* a specific road, and *p(x,i)* the percentage of *x* contained in *i* with *p(x,i)* ∈ [0,1]. Then *peC* is updated by

$$peC_i^{new} = peC_i^{old} + \begin{cases} 1 & if\, p(x,i) = 1 \\ 0 & else \end{cases}$$

The term *wC* stands for winner counter. It is increased by 1 for the district in which the major part of a road is situated. It is not increased for any other district. Let *j* be another district and *n* be the number of districts. Then *wC* is updated by

$$wC_i^{new} = wC_i^{old} + \begin{cases} 1 & if\, p(x,i) > p(x,j)\, \forall j \in (1,\dots n),\, i \neq j \\ 0 & else \end{cases}$$

The third counter is called equal counter (*eC*). It is increased for every district the road crosses by the value of 1 divided by the number of districts the road crossed. This means that the count is divided between all concerned districts. For example, if the road runs through four districts, the *eC* values for these districts are increased by 0.25. Let *m* be the number of roads with *p(x,i)>0. eC* is then updated by

$$eC_i^{new} = eC_i^{old} + \frac{1}{m}$$

The fourth counter, *paC*, or partial counter, is increased by the degree to which a road is located in a district. If, for example, a road runs for twenty meters through district *i* and for thirty meters through district *j*, paC_i is increased by 0.4 and paC_j by 0.6.

$$paC_i^{new} = paC_i^{old} + p(x, i)$$

The fifth and final counter is named *oC*, which is short for optimistic counter. It is increased by 1 for every district the road runs through, regardless of how much of the road is actually situated in the district.

$$oC_i^{new} = oC_i^{old} + \begin{cases} 1 \; if \; p(x,i) > 0 \\ 0 \; else \end{cases}$$

Obviously, *peC* and *oC* are very unlikely to hold the true value. Note that if *peC* or *oC* hold the true value for district *i*, for any district *j* with which *i* shares a road, peC_j or oC_j cannot hold the true values. However, *peC* and *oC* form the lower and upper borders of the interval in which the true value definitely lies. In the following this interval will be called the value interval. The counters *wC*, *eC,* and *paC* are all situated in this interval. We then counted the road encoded emergency events according to above formulae. Note that this methodology could also be adapted to the case that the events are not given as line but as polygon features, so that various polygons would need to be examined.

In order to describe the uncertainties, we employed some additional values. The first of these values is *peORatio*, which gives the ratio between *peC* and *oC*. The closer it is to 1, the narrower the value interval is.

$$peORatio = \frac{peC}{oC}$$

peORatio correlates heavily with the ratio between roads which are only partially contained in a district and roads which are fully contained. A high value indicates a dense interval and thus narrows down the true value.

The second additional value is

$$unc_i = \sum^{o} unc_{i,x}$$

where

$$unc_{i,x} = 1 - 2 * [0.5 - p(x,i)]$$

and o is the total number of emergency events. unc_i is an important value for the rating of *wC* and *paC*, since it stresses the degree of containment of a road in a district. It indicates how many of the events could be unambiguously assigned. The closer unc_i is to 1, the more uncertain the values of *wC* and *paC* are.

Another value which tells about this relationship is

$$\sigma_i^2 = \frac{1}{\sqrt{o}} \sum^{o} \sigma_{x,i}^2$$

where

$$\sigma_{x,i}^2 = p(x,i) * (1 - p(x,i))$$

This value treats each event as a Bernoulli trial, as it holds the possibility of a specific event occurring in a specific district against the possibility of the event to happen outside of the district. It has the advantage that it does not only consider the uncertain positions of the events, but also the absolute number of uncertain events. It thereby introduces a further quality of uncertainty consideration to the analysis. Such an approach is necessary in order to underline the importance of the consideration of the absolute values.

All these values highlight different aspects of the uncertainty and emphasize aspects which would not be considered by the other values alone. Note that the outcomes of both unc_i and σ_i^2 are not influenced in any way by *peORatio*. Their values correspond only to the uncertainty within the value interval.

3. RESULTS

The outcome of the emergency counts for the study area of Bergedorf is presented in figure 1 and in table 1.

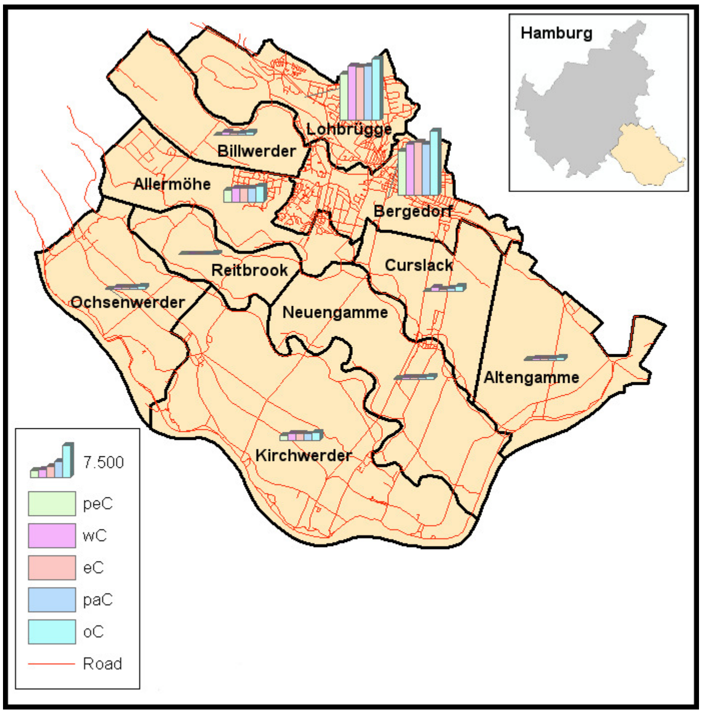

Figure 1: The counter values for the study area of Bergedorf

As can be seen from table 2, the values describing uncertainties complement each other. Since the values emphasize different aspects of uncertainty, the distinct districts do not necessarily receive values indicating certainty or uncertainty for each measure. This is most notable for the districts of Reitbrook on the one hand and Lohbrügge and Bergedorf on the other. Reitbrook has the lowest values for unc_i and σ_i^2, but one of the lowest for *peORatio*. This implies that most roads intersecting Reitbrook are not contained fully, but to a high degree. Lohbrügge and Bergedorf have high *peORatio* values, but also very high values for σ_i^2. These high values can be explained by the fact that Lohbrügge and Bergedorf have by far the highest population. Reitbrook has clearly the lowest. It is possible that in highly populated districts most people live at roads which are fully contained within a district.

	peC	wC	eC	paC	oC
Allermöhe	2,801	3,257	3,336.5	3,378.33	3,872
Altengamme	221	351	440	352.9	707
Bergedorf	10,376	12,102	12,436.64	12,113.36	15,06
Billwerder	171	804	432.83	652.79	809
Curslack	95	957	611	809.75	1,175
Kirchwerder	1,207	1,553	1,539.67	1,517.36	1,873
Lohbrügge	10,905	12,795	12,526.31	12,791.34	14,548
Neuengamme	254	589	550.67	626.29	848
Ochsenwerder	160	487	494.67	481.15	849
Reitbrook	84	89	218	95.61	352

Table 1: The counter values for the study area of Bergedorf

	peORatio	unc_i	σ_i^2
Allermöhe	0.72	0.04	0.36
Altengamme	0.31	0.07	0.82
Bergedorf	0.69	0.17	6.68
Billwerder	0.21	0.38	3.46
Curslack	0.08	0.45	4.45
Kirchwerder	0.64	0.1	1.53
Lohbrügge	0.75	0.16	6.21
Neuengamme	0.3	0.16	1.62
Ochsenwerder	0.19	0.17	2
Reitbrook	0.24	0.04	0.36

Table 2: The uncertainty measures for the study area of Bergedorf

Still, some districts have values indicating high or low uncertainty for all measures. The values for Billwerder and most notably Curslack all show a high degree of uncertainty. Values indicating low uncertainty are derived for Altengamme and Kirchwerder. These cases highlight the need for uncertainty descriptions that take into account more than one characteristic.

Since the true values are unknown, we cannot compare the results of the counters to the real values but only amongst themselves. To do so, we used three ratios that would help to understand the outcome of the counters. The *wRatio* shows the relative location of *wC* in the value interval. When *wRatio* is close to 1, *wC* is close to *oC*. Vice versa, when *wRatio* is close to 0, *wC* is close to *peC*. Likewise, *eRatio* and *paRatio* show the relative locations of *eC* and *paC* in the value interval.

$$wRatio = 1 - \frac{oC - wC}{oC - peC}$$

$$eRatio = 1 - \frac{oC - eC}{oC - peC}$$

$$paRatio = 1 - \frac{oC - paC}{oC - peC}$$

As can be seen from table 3 and figure 2, the values of *eC* tend to be closely below the value interval middle. They differ strongly from the values for *wC* and *paC* and are always situated in the same region of the value interval. The values of *wC* and *paC* are distributed over the whole range of the value interval. The most frequent case is that when *wC* is situated between the interval mean and either *peC* or *oC*, *paC* is situated between the interval middle and *wC*. In order to examine the common behaviour of *wC* and *paC*, we included a fifth value *wPDiff*, which indicates the difference between *wRatio* and *paRatio*.

$$wPDiff = [wRatio - paRatio]$$

The distinct values of *wPDiff* can also be found in table 3. Additionally, we also included data about the population of the districts. This data refers to the year 2008 and stems from the local office for statistics, Statistikamt Nord.

	Population	wRatio	eRatio	paRatio	wPDiff
Allermöhe	15,347	0.426	0.5	0.539	0.113
Altengamme	2,194	0.267	0.451	0.271	0.004
Bergedorf	40,521	0.368	0.44	0.371	0.002
Billwerder	1,301	0.992	0.41	0.755	0.237
Curslack	3,743	0.798	0.478	0.662	0.136
Kirchwerder	9,012	0.52	0,5	0.466	0.054
Lohbrügge	38,442	0.519	0.445	0.518	0.001
Neuengamme	3,453	0.564	0.499	0.627	0.063
Ochsenwerder	2,295	0.475	0.486	0.466	0.008
Reitbrook	480	0.019	0.5	0.043	0.025

Table 3: The ratio values and the population for the districts of the study area

The ratios of the district counters vary strongly—which is not surprising, since the districts are diverse with respect to population. One thing that becomes apparent from table 3 is that when *peORatio* is high, *wC* and *paC* are more likely to show a different behavior than explained above. This happens because the number of roads which make up for the difference between *wC* and *paC* is small when the value interval is narrow. Therefore, districts with unusual road distributions can lead to unusual results. Examples are given below.

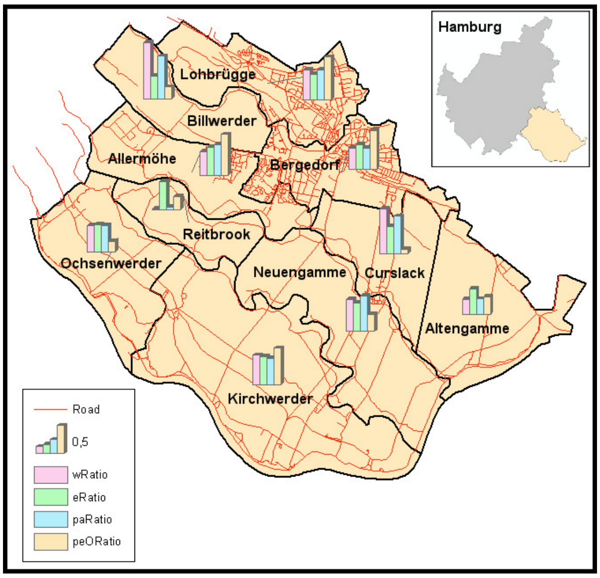

Figure 2: The ratio values for the study area of Bergedorf

A large variance between *wC* and *paC* occurred in Billwerder, where *wPDiff* has an overall value of 0.237. *wRatio* is 0.992, which is much larger than the *paRatio* of 0.755. An examination of this district revealed that 76.74% of the emergencies occurred on roads which are included to a degree above 0.5, but below 1. This is emphasized by a low *peORatio* of 0.21 and a rather high unc_i value of 0.38. Thus the difference between *wC* and *paC* was stressed. See figure 3 for details.

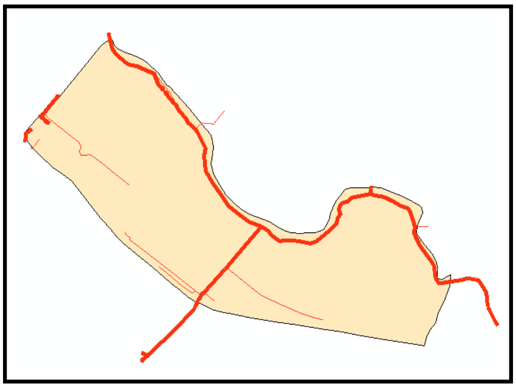

Figure 3: The district of Billwerder and the roads which intersect it. Along the marked roads, 76.74% of the emergencies occurred.

In Billwerder, relatively little emergencies occurred in the corresponding districts. This is not the case for Allermöhe, where *paRatio* is 0.539, but *wRatio* is only 0.426. One part of the explanation of this phenomenon is that the value interval is very small, with a *peORatio* of 0.72. Therefore small differences in the values of *wC* and *paC* can have a large impact on the ratios. The other part of the explanation is that many roads are contained only to an amount between 0.35 and 0.5. Thus, for an emergency that occurred in them, *paC* is increased by a not very small value, while *wC* is not increased at all. See figure 4 for details.

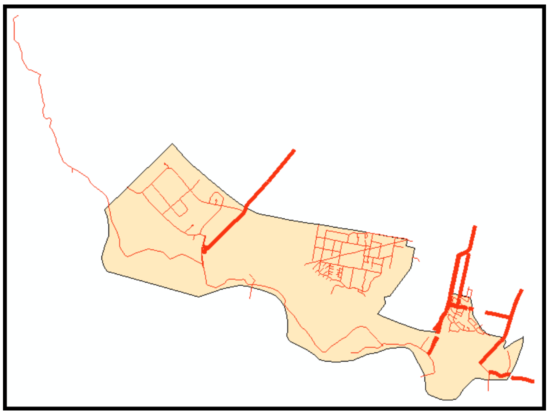

Figure 4: The district of Allermöhe and the roads which intersect it. The marked roads are contained to a degree between 0.35 and 0.5. Along these roads 13.54% of the emergencies occurred.

4. CONCLUSION

In this paper, we have shown a way to count positionally uncertain events for discrete areas with explicitly considering and describing the uncertainty. In order to count these uncertainties, we increased five counters according to different rules. We compared the various counters to each other. We found that the equal counter *eC* was always located below the interval mean and therefore did not seem to be a good approximation of the true value. The values of the winner counter *wC* and the partial counter *paC* in most cases took similar values. Unlike the values for *eC* they were distributed over the whole interval. *wC* usually was oriented towards the interval borders, whilst *paC* was rather oriented towards the interval middle. Since for *paC* more information about the spatial properties of the roads is considered, it is probable that it is better suited to estimate the true value than the winner counter. It is also important to point out that the counters might form a different distribution pattern in other applications.

We also showed three measures of uncertainty and how they correspond to these counters. The pessimistic counter *peC* and the optimistic counter *oC* span a confidence interval. The relationship between *peC* and *oC* can be described by their ratio *peORatio*. The counter values could also be used to describe uncertainties with fuzzy boundaries or probability density functions. For this purpose, *wC* or *paC* could be used as the estimations of the true values. The uncertainty measures unc_i and σ_i^2 serve to describe the uncertainty of these counters.

Future work could be the test of the counters in another application, probably in an environment where the positional uncertainty of an event does not refer to a line but to an area. An optimization of this approach would be the comparison of the counted values to the real values. With regard to the topic of line clipping, it could be examined what to do with other thematic data, e.g., pollution measures. Our approach could be extended in such a way that sophisticated interpolation methods could be considered.

Our future focus will aim at examining the interdependencies between the amount of emergencies that occurred in a district and the socio-demographic characteristics. We will not only examine the relation between population and the amount of emergencies but also the influence of other socio-demographic factors, like age distribution, average income, or unemployment. Previous work has shown that relations between the amount of emergencies and population (Krisp and Karasová, 2005) or age distribution (Špatenková, 2007) do exist, so we want to test further socio-demographic parameters. When these interdependencies are known and when the information about the distribution of these parameters along a street is available, they could be used to build a weighted partial counter. Also tests could be performed in order to examine which parameters influence specific types of emergency events.

5. ACKNOWLEDGEMENT

The authors would like to thank the fire department of Hamburg for the kind cooperation.

BIBLIOGRAPHY

Clementini, E. and Di Felice, I., An Algebraic Model for Spatial Objects with Indeterminate Boundaries. In P.A. Burrough and A.U. Frank (eds.). Geographic Objects with Indeterminate Boundaries. London: Taylor & Francis: 139-170, 1996.

Cohn, A.G. and N.M. Gotts, The 'Egg-Yolk' Representation of Regions with Indeterminate Boundaries. In. P.A. Burrough and A.U. Frank (eds.). Geographic Objects with Indeterminate Boundaries. London: Taylor and Francis: 171-187, 1996.

Cyrus, M. and Beck, J., Generalized two- and three-dimensional clipping. Computer & Graphics Vol. 3 (1), 23-28, 1978.

Davis, B.E., GIS: A visual approach. 2nd Edition. Onword Press. 2001.

Egenhofer, M. and Herring, J., Categorizing Binary Topological Relations Between Regions, Lines, and Points in Geographic Databases. Technical Report, Department of Surveying Engineering, University of Maine, Orono, ME, 1991.

Fisher, P.: Sorites paradox and vague geographies. Special Issue on Uncertainty in geographic information systems and spatial data. Vol. 113 (1), 7-18, 2000.

Grubesic, T.H. and Murray, A.T., Assessing positional uncertainty in geocoded data. In Proceedings of the 24th Urban Data Management Symposium. 2004.

Hession, S.L., Shortridge, A.M. and Torbick, N.M., Categorical models for spatial data uncertainty. In Proceedings of the 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, 386-395, 2006.

Krisp, J.M. and Karasová, V., The relation between population density and fire / rescue service incidents in urban areas. In Proceedings of the 10th Scandinavian Research Conference on Geographical Information Science (ScanGIS), 237-246, 2005.

MacEachren, A.M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M., and Hetzler, E., Visualizing Geospatial Information Uncertainty: What We Know and What We Need to Know. Cartography and Geographic Information Science. Vol. 32 (3), 139-160, 2005.

Morgan, M.G. and Henrion, M., Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge University Press, Cambridge, United Kingdom, and New York. 1990.

Murray, A.T., Site placement uncertainty in location analysis. Computers, Environment and Urban Systems. Vol 27 (2), 205-221, 2003.

Navratil, G., Karimipour, F., and Frank, A.U., Lifting Imprecise Values. In: L.Bernard, A. Friis-Christensen, and H. Pundt (eds.). The European Information Society: Taking Geoinformation One Step Further. Springer Berlin Heidelberg. 79-94, 2008.

Reis, R., Egenhofer, M., and Matos, J., Topological Relations using two models of uncertainty for lines. In Proceedings of the 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, 286-295, 2006.

Shokri, T., Delavar, M., Malek, M., and Frank, A., Modeling Uncertainty in Spatiotemporal objects. In Proceedings of Accuracy 2006. 469-478.

Špatenková, O., Demšar, U., and Krisp, J.M., Self-organising maps for exploration of spatio-temporal emergency response data. In: Proceedings of Geocomputation 2007, 2007.

Traub, K.-P., Geomedical Information Systems. The Application of GIS to Health Mapping – Case Hamburg. Geoinformatics, Vol. 6. 2003.

Wu, H.S. and Liu, Z.L., Simulation and Model Validation of Positional Uncertainty of Line Feature on Manual Digitizing a Map. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XXXVII, Part B2. 2008.

Zhao, X., Chen, X., and Stein, A., A Random Sets Model for Spatial Objects with Uncertain Boundaries. 12th AGILE International Conference on Geographic Information Science, Hannover, Germany, 2009.