Investigation of Medication Dosage Influences from Biological Weather

Kostas Karatzas[1], Marina Riga[1], Dimitris Voukantsis[1] and Åslög Dahl[2]

[1] ISAG – Informatics Systems and Applications Group, Department of Mechanical Engineering, Aristotle University of Thessaloniki, 54124, Thessaloniki, Greece
kkara@eng.auth.gr, {mriga, voukas}@isag.meng.auth.gr
[2] Gothenburg Atmospheric Centre/ Aerobiology, Department of Plant and Environmental Sciences, University of Gothenburg, SE 405 30, Gothenburg, Sweden
aslog.dahl@dpes.gu.se

Abstract. Airborne pollen has been associated with allergic symptoms in sensitized individuals, whereas atmospheric pollution indisputably aggravates the impact on the overall quality of life. Therefore, it is of major importance to correlate, forecast and disseminate information concerning high concentration levels of allergic pollen types and air pollutants to the public, in order to safeguard the quality of life of the population. In this study, we investigate the relationship between the Defined Daily Dose (DDD) given to patients in a triggered allergy reaction and the different levels of air pollutants and pollen types. By profiling specific atmospheric conditions, specialists may define the need for medication to individuals suffering from pollen allergy, not only according to their personal medical record but also to the existing air quality observations. Paper results indicate some interesting interrelationships between the use of medication and atmospheric quality conditions and shows that the forecasting of daily medication is possible with the aid of proper algorithms.

Keywords: Allergy, Pollen, Medication Dosage Forecasting, Information Gain Criterion, Self-Organizing Maps, Decision Trees.

1 Introduction

Allergies related to airborne pollen are very common in the general population and also act as a factor with a considerable impact on quality of life in urban areas [1]. Atmospheric pollen counts are positively correlated with symptoms of allergic rhinitis and/or conjunctivitis [2-3], emergency visits or hospitalization because of asthma [4-6] and stroke [7]. In addition, even a short term exposure to allergic pollen has been found to increase the prescribed anti-allergic medicines [8]. The dosage and duration of medication primarily depends on interaction between an individual's immune system and pollen allergens, but also, presumably, on environmental pressures, such as air pollution. Air pollutants can induce respiratory inflammation and affect the response to aeroallergens. The evidence of interaction between pollen and air pollution in their health impact comes from many laboratory and experimental studies, as well as from long-term epidemiological studies [9-11] (and references therein). The number of studies indicating short-term, acute effects are fewer but at

least ozone and aeroallergens have been found to interact in promoting asthma [12-15], and this stands also for the combination of particulates and pollen [16-17].

2 Materials and Methods

The aim of the current paper is to study the relationship between aeroallergens, air pollution and the use of anti-allergic medication (antihistamines) to citizens, during spring and summer seasons. More specifically, we investigate the relationship between the so called Defined Daily Dose (DDD) prescribed to allergic patients and the levels of air pollutants and pollen concentrations in the atmosphere. For this reason, a dataset of corresponding observation data was provided by the Gothenburg Atmospheric Centre (GAC), from the Department of Plant and Environmental Sciences at the University of Gothenburg in Sweden.

The available dataset is a time series (corresponding to middle-spring and summer seasons) for the years 2009 and 2010, consisting of heterogeneous data (air quality concentrations, pollen counts and medication dosage). Due to the heterogeneity of the dataset and the existence of some missing values in the time series, it became evident that a robust computational methodology was required, which may tolerate missing data and can also handle parameters of varying nature. For this reason, we adopted Self-Organizing Maps (SOMs) in the first phase of our study, as an unsupervised learning approach to extract hidden patterns and produce visualizations of interrelationships between the available parameters (data set attributes). In the second phase, we utilize the Information Gain criterion and produce forecasting models based on decision trees, for the prediction of the medication levels provided to patients on aggravated air quality conditions. The evaluation of the models' performance is made through the calculation of specific statistical indexes.

2.1 Study Area

Gothenburg is the second largest city in Sweden and the fifth largest in the Nordic countries, accounting for approximately 510,000 inhabitants in a total area of 1,029km^2. It is located on the west coast in South-West Sweden, at the mouth of the river Göta Älv. The city contains an adequate rate of green space, estimated at 175m^2 per citizen, while there are several parks and nature areas that reserve space ranging from tens of meters to hundreds of hectares. The climate of Gothenburg is characterized as suboceanic, with warm summers and cool, windy winters, pointing a narrow annual temperature range. During the summer, daylight extends 17 hours, but lasts only 7 hours in late December. The quality of the atmospheric environment is generally good in the city of Gothenburg, which is not extremely polluted due to its moderate size and location near the coast. Nevertheless, the climatic conditions of the last years accumulated the air pollutants emitted at the ground level and air quality became poorer than in the past [18].

2.2 Pollen Collection and Identification and Air Quality Monitoring

For pollen monitoring, Burkard 7-day volumetric spore traps were used. The Gothenburg trap is situated approx. 30m above ground at the roof-top of the Central Clinic at Östra sjukhuset, at the eastern border of Gothenburg and 20km from the sea (57˚72'N, 12˚05'E). In the microscope, pollen grains were counted along twelve latitudinal transects of the exposed tape, corresponding to an exposure to 1 cubic metre of air. The value used in the calculations is the number of pollen grains per cubic metre and 24 hours.

Gothenburg air pollution data originated from a monitoring station owned by the Environment Department of Gothenburg City, approx. 20m above ground at the roof top of the shopping centre "Femman" in the very city centre (57˚42'N 11˚58'E), and 8km away from the pollen trap. All pollution data used are 24-hour sums of hourly values.

2.3 Exploration of the Initial Dataset

A diverse collection of air quality, pollen and medication information was under investigation, forming a daily time series of observations from middle spring to summer season (1/4/2009 to 31/8/2009 and 1/4/2010 to 31/8/2010).

More specifically, the data used in this study consisted of daily pollen counts of 7 main pollen types (*Alnus*, *Betula*, *Corylus*, *Fagus*, *Quercus*, *Poaceae* and *Artemisia*) and 3 features concerning overall pollen related measurements (total birch related pollen, i.e. the sum of *Alnus*, *Betula*, *Corylus*, *Fagus* and *Quercus*, total allergenic pollen and total pollen counts). Daily sums of hourly concentrations of 9 air pollutants (SO_2, NO, NO_2, NO_x, CO, O_3, PM_{10}, $PM_{2.5}$ and Soot) and additional time information (year, month, day of the week and day of year) were taken into account. The DDD of medication given to allergic patients was used as a class attribute. This attribute is defined as the average "maintenance" dose per day for a drug used in adults. The available dataset was divided into 2 subsets and was investigated for the years 2009 and 2010 separately. Overall, 153 daily datasets per year were made available, each one including 24 different attributes.

2.4 Methods

The primary investigation of the relationships between air quality, pollen concentrations and medication dosage was performed using Self-Organizing Maps [19]. In addition, and in order to support patients in managing the quality of their everyday life, a number of forecasting models for the DDD was developed, by employing the well known *C4.5* Decision Tree algorithm [20]. In order to select the appropriate features from the dataset that will feed the prediction model with valuable information, the Information Gain criterion was adopted. It should be noted that Computational Intelligence methods have already been applied in the analysis of allergenic pollen types [21-22]. Moreover, a detailed analysis of interconnections

between pollen and air quality data has been performed in [23], by introducing a two-step clustering process consisting of the application of SOM and K-means algorithms.

In the current study, SOM visualizations have been deployed using the function package SOM Toolbox 2.0[1] in the Matlab environment, while J48 decision trees models have been developed using the open source Data Mining tool WEKA[2].

Self-Organizing Maps (SOMs). A self-organizing map is a form of an artificial neural network that uses a competitive, unsupervised learning algorithm, in order to model high-dimensional data into low-dimensional (usually 2-dimensional) visualizations, by also preserving constant their spatial/ topological interrelationships. SOM consists of neurons, each one of which is a set of coefficients corresponding to the variables of the data set. The number of neurons and the interconnection between them is determined by the user. Results of the method are visualized by the production of Kohonen maps, with specific characteristics, meaning: data with similar "behavior" are grouped together and displayed in the same 2-D space of maps [19]. It is a robust and computationally efficient method that not only compresses the dimensions of data but also explores and extracts hidden information of them. On this basis, SOMs are capable of identifying complex, non-linear relations within data, and representing them in a convenient and understandable way. In the current study, the map grid size of the topology structure was 24x16 (parameter size: 'normal'), arranged on a hexagonal lattice in a sheet shaped map.

Information Gain Criterion. In order to obtain an optimum performance of the modeling process, several parameters need to be evaluated, concerning the information gain they give. The method adopted in the current study for feature selection is the Information Gain (IG) criterion which evaluates the worth of an attribute by measuring the information gain with respect to the predicted class variable. The method is based on the notion of entropy that, in this context, characterizes the impurity of an arbitrary set of attributes.

In the current case, the target variable is the DDD. Since the IG criterion requires for the target variable to be nominal, the numeric values of DDD per 1000 inhabitants were transformed into nominal ones, according to the classification table bellow (Table 1).

Table 1. The transformation table of defined daily dose from numeric values to nominal.

DDD per 1000 inhabitants	Nominal value (class name)
<25	Class1: Very low
25-32.5	Class2: Low
32.5-40	Class3: Medium
40-47.5	Class4: High
>47.5	Class5: Very high

[1] Available at: http://www.cis.hut.fi/projects/somtoolbox/
[2] Available at: http://www.cs.waikato.ac.nz/ml/weka/

Continuing, by calculating IG for each variable through the WEKA software, a ranking of each parameter's impact became available, giving their relative importance in descending order (Table 2):

Table 2. Classification of each considering variable, according to their information gain value estimation (descending order).

Dataset of 2009		Dataset of 2010	
1. DayOfYear	13. Soot	1. DayOfYear	13. Year
2. Month	14. NO_2	2. Month	14. Day
3. NO	15. Day	3. Artemisia	15. SO_2
4. Artemisia	16. $PM_{2.5}$	4. O_3	16. NO_2
5. TotalBirchRelated	17. O_3	5. TotalBirchRelated	17. TotalPollen
6. TotalAllergenicPollen	18. SO_2	6. NO	18. NO_x
7. PM_{10}	19. Quercus	7. Betula	19. CO
8. Betula	20. Fagus	8. Corylus	20. Fagus
9. NO_x	21. Corylus	9. Alnus	21. TotalAllergenicPollen
10. TotalPollen	22. Poaceae	10. PM_{10}	22. Quercus
11. Alnus	-	11. $PM_{2.5}$	23. Poaceae
12. Year	-	12. Soot	-

The IG values of the parameters within the studied data set (Table 2) suggest that certain pollen types are the least important variables in the modelling process of DDD while air quality data (and specifically concentrations of O_3, PM_{10} and NO) seem to play a significant role in medication usage.

Decision Trees. Decision trees are a set of classifiers that produce a top-down structure of interconnected nodes (a tree), by selecting a root node and partitioning the data properly following "*if… else*" rules. The most significant variables (those with high information gain value) are located in the upper layers of the tree, while the least valuable are the leaf-nodes. They are considered especially attractive as classification techniques, due to their following characteristics:

- their intuitive representation allows for easy interpretation of the resulting classification model, as their structure "inherits" the basic characteristics of the knowledge domain they are mapping, and
- they are non-parametric, thus especially suited for exploring datasets where there is no prior knowledge about the probability distributions of the target class or data attributes.

In the current study, calculations were made with the J48 classifier (implementation of C4.5 algorithm in the WEKA data mining tool), which achieves fast execution times and adequate scale of large datasets. As this method requires the class attribute to be nominal, arithmetic values of DDD per 1000 inhabitants were transformed according to classification rules given in Table 1. In order to create the forecasting models, the input data were selected via a process by which certain data were excluded from the initial data set, in the frame of consequently executed data modification steps. After every step, a classification model was constructed and evaluated by using 10-fold cross validation.

The accuracy of the models was measured according to percentage (%) of correctly classified instances and Kappa-statistic. In more detail, the first measure is giving a

success rate by dividing the sum of correctly predicted instances to the total number of predicted instances, but this metric is not sensitive to class distribution. On the other hand, Kappa-statistic is a more robust statistical measure that is used to calculate the agreement between predicted and observed categorizations of a dataset, while correcting for agreement that occurs by chance. It estimates the error cost of a classification model, by being a more fair measure of the overall success of the model.

3 Results and Discussion

The aforementioned methodology was applied at both datasets of the years 2009 and 2010 separately. The objective was to identify interrelationships between pollen concentrations and DDD or air quality observations and DDD, by adopting the SOM method and then constructing efficient decision trees.

3.1 SOM Analysis of the Initial Dataset

High values of DDD seem to be a multi-parametric phenomenon, i.e. influenced by various parameters. On this basis, synergies can be expected between pollen and air quality. Figure 1 represents the map resulted using SOM on the overall dataset of 2009. The most important conclusions from the analysis are listed below:

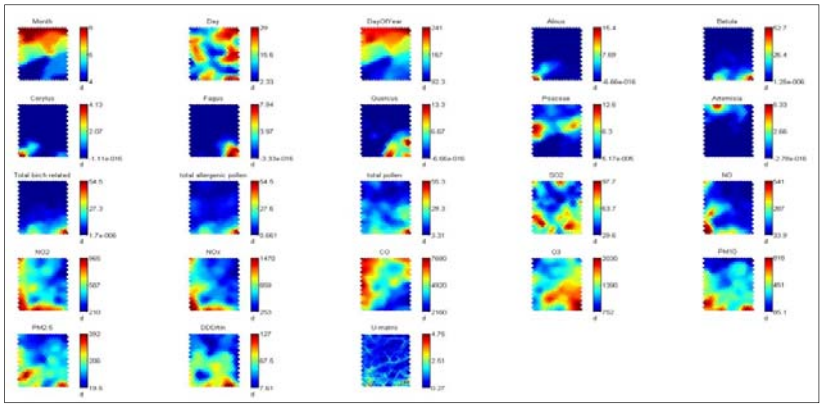

Fig. 1. Self-organizing maps of air quality, pollen and medication data for the year 2009

- The use of medication is lower when *Artemisia* is in its highest concentration levels, while *Poaceae* is correlated with medium levels of medication.
- *Alnus* and *Corylus* seem to be related with high levels of NO, NO_2 and PM_{10}, while additionally they are related with medium to high levels of medication.
- High levels of medication use is observed when air pollution is high, especially for PM and O_3 levels.

- The existence of high NO and NO_2 concentration levels is correlated to medium levels of medication.

No efficient conclusions could be derived as far as CO levels are concerned, since the corresponding levels of medication vary from low to high. Overall, it becomes evident that DDD is positively correlated with air quality and some specific pollen taxa.

By repeating the SOM analysis on the overall dataset of 2010 the following basic conclusions are derived:

- Medication usage is related to high levels of air pollution, and especially to PM_{10}, O_3 and Soot, with the latter revealing a very similar topological pattern with DDD in the derived SOM visualizations.
- NO_2 and NO_x seem to play an important role in high medication usage.
- High total pollen measurements increase medication.

3.2 Decision Trees Application

The development of models for the use of medication was performed by testing a number of scenarios concerning the types of data to be included. The results of this process in terms of model performance are described in Table 3.

Table 3. Evaluation of developed decision trees in modelling medication usage for 2009 and 2010 separately, according to two different dataset variations (case scenarios).

	Dataset 2009		**Dataset 2010**	
Data scenario	**Correctly classified %**	**Kappa Statistic**	**Correctly classified %**	**Kappa Statistic**
A	**56.86% (87/153)**	**0.3874**	54.90% (84/153)	0.3250
B	52.94% (81/153)	0.3271	**59.48% (91/153)**	**0.3991**

Data scenario A: Exclusion of the variables Year, Soot and *Fagus* (no measurements were available for both years).
Data scenario B: Removing all pollen data observations and total birch related variable and keeping only relative information about total allergenic pollen and total pollen counts.

The accuracy of developed decision trees models, concerning the best performing scenario per year, ranges from 56% to 59%, in terms of correctly classified instances. This method outperforms a completely random approach, where the probability of correct classifications, by choosing randomly one of the five pre-defined classes as a forecasting result, equals 20%.

While interpreting the developed decision trees for best case scenarios per each year (see Figure 2 for year 2009), it is interesting to note that the day of the year seems to be the most decisive parameter in the modelling process. Other parameters of importance include PM_{10} and NO_2 (for 2009) and PM_{10} and O_3 (for 2010), followed by pollen type parameters. Overall, there is evidence that air quality parameters play an important role in modelling medication use, for the aforementioned years.

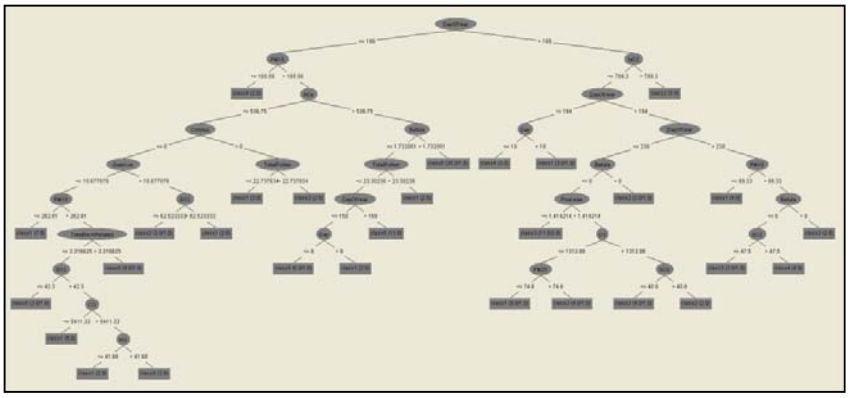

Fig. 2. J48 decision tree with the best performance (Case A) for the dataset of 2009

4 Conclusions

The motivation in the current study was to develop self-organizing maps that would exploit interrelationships between the use of medication and air quality conditions, concerning pollution and pollen concentrations, so as to evolve computationally effective decision trees for daily medication prediction. The preliminary analysis of the Gothenburg dataset indicates that there are some interesting relationships between the parameters under investigation. More specifically:

- Air quality seems to play a synergetic role in the use of medication, influencing the total DDD per 1000 inhabitants. Thus, an interaction between air quality levels and medication usage (i.e. symptom development) is suggested from the results of the preliminary analysis.
- The medication being used seems to be also influenced by the total allergenic pollen.
- The role of Particulate Matter seems to be away from neutral or independent, yet more research is certainly required in this subject.

Further investigation need to be done, in order to enhance the efficiency of prediction models. The inclusion of meteorological factors seems to be a promising task for research through, while datasets with more years of data, complete patient diary data or different health indicators (like hospital admissions, mortality, etc.) would reveal more complicated relationships and potentially result to more efficient solutions in the modelling process.

Acknowledgments. The authors would like to thank the Gothenburg Atmospheric Center/ Aerobiology Department of Plant and Environmental Sciences for providing the pollen and air quality data.

References

1. Cariñanos, P., Casares-Porcel, M.: Urban green zones and related pollen allergy: A review. Some guidelines for designing spaces with low allergy impact. Landscape and Urban Planning 101, Issue 3, pp. 205--214 (2011)
2. Breton, M.C., Garneau, M., Fortier, I., Guay, F., Louis, J.: Relationship between climate, pollen concentrations of Ambrosia and medical consultations for allergic rhinitis in Montreal, 1994-2002. Science of the Total Environment 370, pp. 39--50 (2006)
3. Takasaki, K., Enatsu, K., Kumagami, H., Takahashi, H.: Relationship between airborne pollen count and treatment outcome in Japanese cedar pollinosis patients. European Archives of Oto-Rhino-Laryngology 266, pp. 673--676 (2009)
4. Carracedo-Martinez, E., Sanchez, C., Taracido, M., Saez, M., Jato, V., Figueiras, A.: Effect of short-term exposure to air pollution and pollen on medical emergency calls: a case-crossover study in Spain. Allergy 63, pp. 347--353 (2008)
5. Tobias, A., Galan, I., Banegas, J.R. Aranguez, E.: Short term effects of airborne pollen concentrations on asthma epidemic. Thorax 58, pp. 708--710 (2003)
6. Jariwala, S.P., Kurada, S., Moday, H., Thanjan, A., Bastone, L., Khananashvili, M., Fodeman, J., Hudes, G., Rosenstreich, D.: Association between tree pollen counts and asthma ED visits in a high-density urban center. Journal of Asthma 48, pp. 442--448 (2011)
7. Low, R.B., Bielory, L., Qureshi, A.I., Dunn, V., Stuhlmiller, D.F., Dickey, D.A.: The relation of stroke admissions to recent weather, airborne allergens, air pollution, seasons, upper respiratory infections, and asthma incidence, September 11, 2001, and day of the week. Stroke 37, Issue 4, pp. 951--957 (2006)
8. Fuhrman, C., Sarter, H., Thibaudon, M., Delmas, M.C., Zeghnoun, A., Lecadet, J., Caillaud, D.: Short-term effect of pollen exposure on antiallergic drug consumption. Annals of Allergy, Asthma and Immunology 99, pp. 225--231 (2007)
9. D'Amato, G., Liccardi, G., D'Amato, M.: Environmental risk factors (outdoor air pollution and climatic changes) and increased trend of respiratory allergy. Journal of Investigational Allergology and Clinical Immunology 10, pp. 123--128 (2000)
10. Lubitz, S., Schober, W., Pusch, G., Effner, R., Klopp, N., Behrendt, H., Buters, J.T.: Polycyclic aromatic hydrocarbons from diesel emissions exert proallergic effects in birch pollen allergic individuals through enhanced mediator release from basophils. Environmental Toxicology 25, pp. 188--197 (2010)
11. Peden, D., Reed, C.E.: Environmental and occupational allergies. Journal of Allergy and Clinical Immunology 125, pp. 150--160 (2010)
12. Dales, R.E., Cakmak, S., Judek, S., Dann, T., Coates, F., Brook, J.R., Burnett, R.T.: Influence of outdoor aeroallergens on hospitalization for asthma in Canada. Journal Allergy and Clinical Immunology 113, pp. 303--306 (2004)
13. Higgins, B.G., Francis, H.C., Yates, C., Warburton, C.J., Fletcher, A.M., Pickering, C.A., Woodcock, A.A.: Environmental exposure to air pollution and allergens and peak flow changes. The European Respiratory Journal 16, pp. 61--66 (2000)
14. Lierl, M.B., Hornung, R.W.: Relationship of outdoor air quality to pediatric asthma exacerbations. Annals of Allergy. Asthma and Immunology 90, pp. 28--33 (2003)
15. Feo Brito, F., Mur Gimeno, P., Martínez, C., Tobías, A., Suárez, L., Guerra, F., Borja, J.M., Alonso, A.M.: Air pollution and seasonal asthma during the pollen season. A cohort study in Puertollano and Ciudad Real (Spain). Allergy 62, pp. 1152--1157 (2007)
16. Renzetti, G., Silvestre, G., D'Amario, C., Bottini, E., Gloria-Bottini, F., Bottini, N., Auais, A., Perez, M.K., Piedimonte, G.: Less Air Pollution Leads to Rapid Reduction of Airway Inflammation and Improved Airway Function in Asthmatic Children. Pediatrics 123, pp. 1051--1058 (2009)

17. Ghosh, D., Chakraborty, P., Gupta, J., Biswas, A., Gupta-Bhttacharya, S.: Asthma-related hospital admissions in an Indian megacity: role of ambient aeroallergens and inorganic pollutants. Allergy 65, pp. 795--796 (2010)
18. Grundstrom, M., Linderholm, H.W., Klingberg, J., Pleijel, H.: Urban NO_2 and NO pollution in relation to the North Atlantic Oscillation NAO. Atmospheric Environment 45, pp. 883--888 (2011)
19. Kohonen T.: Self-Organizing Maps. (2001)
20. Salzberg, L. S.: C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Machine Learning, Vol. 16, pp. 235--240 (1994)
21. Voukantsis, D., Niska, H., Karatzas, K., Riga, M., Damialis, A., Vokou, D.: Forecasting daily pollen concentrations using data-driven modeling methods in Thessaloniki, Greece. Atmospheric Environment 44, Issue 39, pp. 5101--5111 (2010)
22. Mitrakis, N., Karatzas, K., Jaeger, S.: Investigating pollen data with the aid of fuzzy methods. In: 20th International conference on Artificial Neural Networks: Part III, pp. 464--470. Springer-Verlag Berlin, Heidelberg (2010)
23. Voukantsis, D., Karatzas, K., Rantio-Lehtimaki, A., Sofiev, M.: Investigation of relationships and interconnections between Pollen and Air Quality data with the aid of Computational Intelligence Methods. In: 23rd Conference on Environmental Informatics and Industrial Environmental Protection: Concepts, Methods and Tools, pp. 189--198. Shaker Verlang, Aachen (2009)