Statistical methods towards more efficient infiltration measurements

T. Franz* and P. Krebs

Institute for Urban Water Management, Dresden University of Technology, D-01062 Dresden, Germany
**Corresponding author, e-mail Torsten.Franz2@mailbox.tu-dresden.de*

ABSTRACT
A comprehensive knowledge about the infiltration situation in a catchment is required for operation and maintenance. Due to the high expenditures an optimisation of necessary measurement campaigns is essential. Methods based on multivariate statistics were developed to improve the information yield of measurements by identifying appropriate gauge locations. The methods have a high degree of freedom against data needs. They were successfully tested on real and artificial data. For suitable catchments it is estimated, that the optimisation potential amounts up to 30 % accuracy improvement compared to non-optimised gauge distributions. Beside this, a correlation between independent reach parameters and dependent infiltration rates could be identified, which is not dominated by the groundwater head.

KEYWORDS
exploratory data analysis; extraneous water; infiltration measurements; location of measurements; statistical approaches

INTRODUCTION
Extraneous water in sewers might have serious consequences such as increased discharges of waste water to the receiving waters causing pollution and public health risks as well as increased pumping and treatment costs. Operating and maintenance strategies, which consider infiltration, require a comprehensive knowledge about the infiltration situation in each individual catchment.

Infiltration is basically driven by water head and leakage of the pipe. However, the knowledge about cause-effect relationships between independent pipe characteristics and dependent infiltration rates is fairly limited. There is a significant lack of data due to a high number of relevant processes and influencing factors (Davies *et al.*, 2001) as well as due to the general data situation of large sewer systems. Thus, both deterministic (e.g. Dupasquier, 1999) as well as empirical models (e.g. Gustafsson *et al.*, 1991) are difficult to apply and depend strongly on extensive measurements. Due to the high personnel and financial expenditures (APUSS, 2004) an optimisation of measurement campaigns regarding cost efficiency and information yield is essential. In this study, data driven models were developed using the similarity approach “similar pipe characteristics lead to similar infiltration rates”, i.e. similarities in and between groups of reaches are used to improve the information yield and to transfer available information.

METHODS

Available data

Information about 22 sub-catchments of the city of Dresden (Germany) and 5 sub-catchments of the Emscher catchment (Germany) with a mean sewer length of approx. 32 km were available for the verification of the similarity approach and the development of the optimisation methods. Besides the network topology the data sets contain a characterisation of the individual reaches by their attributes (Table 1). For every sub-catchment the infiltration rate was estimated by measurements, while the corresponding groundwater table was known in some sub-catchments.

Table 1. Independent Parameters.

Both catchments	Dresden catchment	Emscher catchment
Date of construction	Distance to river Elbe	Soil permeability
Function (regional, main, tributary)	Distance to storm sewers	Reduced area ratio
Material (stoneware, concrete, …)	Distance to drainage	Number of joints
Sewer system (foul, combined)	Thickness of cohesive layers	
Profile type (egg, circle, other)		
Profile circumference		
Reach length		
Population density		
Population-specific length		
Distance to surface water		
Distance to buildings		
Street type (main, residential, …)		
Distance to streets		
Slope		
Coverage		

Statistical methods

The characteristics of the sub-catchments were compared with analysis-of-variance methods (ANOVA). An “all-in-one” test was not available, because the parameters belong to different statistical scales. Therefore, every parameter was analysed individually with one-way ANOVA (metric scale), Kruskal-Wallis-ANOVA (ordinal scale), and contingency tables (nominal scale), respectively (Müller, 1991). With these statistical methods it can be investigated whether several grouped samples, i.e. reach groups, belong to one basic population. For that purpose they analyse variances to investigate whether a significant inter-group difference of mean values exists. The observed variances are separated into a component based on the coincidental error (i.e. sum of the squares within the groups) and into components based on different mean values. The latter are tested on statistic significance.

One necessity of the similarity approach is the description of similarities between reaches and of the homogeneity of reach groups, respectively. This problem is related to several multivariate methods of exploratory data analysis like cluster analysis, discriminant analysis or multidimensional scaling (Backhaus *et al.*, 1996). Using these methods reaches or sub-catchments can be seen as objects in an *n*-dimensional space where the dimensions represent relevant parameters. The distances between such objects are a measure for their similarity: the closer the objects the more similar they are.

The relationship between the independent parameters and the infiltration rates was analysed with multidimensional scaling (MDS). This method transforms efficiently a high dimensional arrangement of objects, so that a low-dimensional and interpretable configuration with the optimum approximation of the observed pattern is reached. Thus, any number of independent parameters can be represented by one or two figures. Compared with other multivariate methods the MDS has the advantage, that parameters of all scales can be handled together (Borg and Groenen, 1997).

Method to identify optimum location of measurements

An optimally placed measurement gauge is situated at the outlet of a more or less homogeneous sub-catchment. Thus, the reach-specific error of the measured infiltration rate is minimised and the information content gained is high. The developed method for determining the optimal arrangement of measuring points within a catchment consists of two steps: (1) describing the homogeneity of the upstream sub-catchment with a similarity figure φ for every potential placement and (2) determining the most homogenous sub-catchments with an optimisation algorithm.

The separating element between two reaches is a manhole. Therefore, the similarity figure is linked to manholes. A value $\varphi = 0$ stands for equality, i.e. the considered sub-catchment contains identical reaches, only. For manholes inside a mesh the similarity figure is not defined, because an infiltration rate measured in a mesh cannot be allocated to individual reaches. The calculation of the similarity figure was adapted from cluster analysis. The aim of clustering is the grouping of objects based on their attributes into (at the beginning) unknown classes. The objects in a class are to be similar, objects from different classes should differ significantly. For an overview see Kaufman and Rousseeuw (1990). But, a one-to-one application of cluster analysis is not feasible. Regarding infiltration the dimensional characteristics (network topology) and the anisotropy (flow direction) of sewer systems must be considered in as much as the reaches are not independent objects.

The input for calculating φ of one potential gauges placement is a data set containing all upstream reaches with their attributes. This data set should be pre-treated (outlier identification, parameter relevance etc.). Due to the scale invariance of most distance measures the parameters must be standardised. Otherwise some would have an unintentional weighting. On the other hand the parameters can be weighted by means of their relevance for infiltration. The distance or similarity between two reaches represented by their metric parameter vectors is calculated with the Euclidean and the Mahalanobis distance (Fahrmeir *et al.*, 1996). Compared to Minkowski-metrics, where the Euclidean distance is deduced from, the Mahalanobis distance has the advantage, that it is scale invariant and that it is calculated with uncorrelated parameters, even if the original parameters are correlated. Thus, the significance enhancement of some parameters due to correlation effects is prevented. Ordinal parameters are treated as metric ones. It is estimated, that the influence of this arbitrary information increase is not so profound. The generalised M-coefficient is proposed for nominal parameters (Bacher, 1996). The reaches can be weighted with the water head (extracted from Gustafsson, 2000). The calculated distances must be aggregated on sub-catchment level to determine one characteristic value. The mean distance between all reaches and the mean distance of all reaches to the groups centroid, i.e. the “mean reach”, are used. These aggregations are based on the idea of a more or less equal distribution of extraneous water in sewers.

An optimisation algorithm is run for determining the optimal placement. With an iterative procedure the catchment is divided until a given number of sub-catchments, i.e. measurement gauges, is reached (Figure 1). The most downstream manhole is related to the WWTP and is automatically a gauge. The conditions for the separating manhole are:

- The similarity figure φ is minimal: A minimal φ stands for a maximal possible homogeneity of the upstream reaches.
- Sum of the size measures of upstream reaches > critical value: The size measure can be the sewer length, the reach surface, the connected area, connected inhabitants etc. This constraint results from measurement requirements like minimal flow or length.
- The manhole is not within a mesh: An infiltration rate measured in a mesh cannot be allocated to reaches.

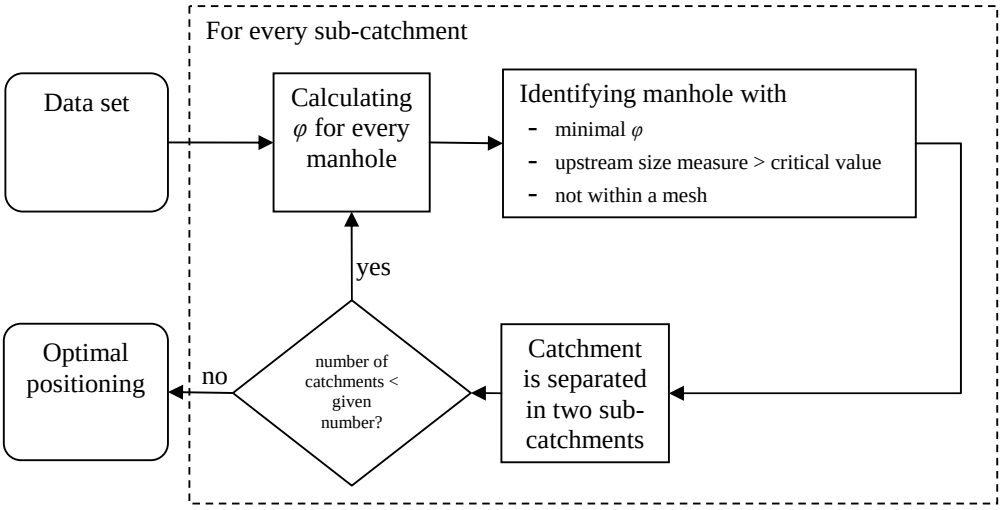

Figure 1. Optimisation algorithm.

Furthermore, the similarity figure can be used as measure for discriminatory power, i.e. the separation between sub-catchments with different φ-values is stronger than those with similar ones. Thus, neighbouring sub-catchments with a similar φ could be merged with a moderate information loss. For example adjacent sub-catchments with low φ might have different groups centroids, but their parameter distribution is narrow. Therefore, a high homogeneity remains while merging these sub-catchments. For that purpose the critical value is set to a relatively small value. Among the separating manholes the manholes with maximum $\Delta\varphi$ between up- and downstream are identified and used as measuring points.

RESULTS AND DISCUSSION

Similarity approach

By means of ANOVA methods significant dissimilarities were detected between sub-catchments and reach populations with different infiltration rates. Due to the uneven distribution and the wide range of observed infiltration rates (min..max = 0.18..2.95 L/[m²*h] = 1÷16) it seemed to be probable, that the dissimilarities between the reach populations are linked in some way with the infiltration rates.

A relation between the measured infiltration rates and the MDS results of the Dresden data set **without** groundwater information, i.e. the independent parameters, could not be found. But, a relation between sewer attributes and the urban quarter type (e.g. inner centre, broader centre, suburb) was identified. Due to this link between city history and network development it can be concluded that in principle a (larger) network consists of identifiable homogenous areas. The MDS results of the Dresden data set **with** groundwater information are shown in Figure 2.

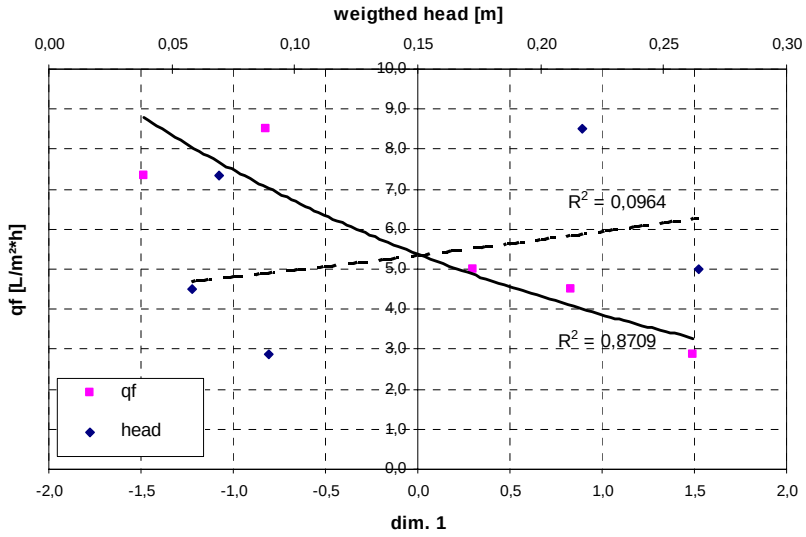

Figure 2. MDS result of Dresden data set with groundwater information and water head vs. reach-surface-specific infiltration rate q_f.

The parameter set was reduced to one dimension *dim. 1*. A good correlation between *dim. 1* and the reach-surface-specific infiltration rate q_f was found. A comparison with the time weighted head shows, that this correlation, i.e. the specific infiltration rate, is not dominated by the groundwater head. The analysis of a combination of the Dresden and the Emscher data set led to similar results. But, significant differences between the cities could be observed. Yet, it is not clear whether these differences are based on natural/regional factors like soil type (Lucas and Fuchs, 2003) or on artificial/local factors like building history. Thus, a transfer of analysis results between catchments does not seem to be reasonable.

Because of the reasonable correlation between *dim. 1* standing for the independent parameters and the infiltration rate (Figure 2) it can be concluded that (1) there is a recognisable relationship between the independent parameters and the infiltration rate and (2) that the parameters - or a part of them - are sufficient to describe infiltration. Thus, the similarity approach is applicable.

Optimised positioning

The optimisation method was tested at six sub-catchments both of the Dresden and the Emscher area. Reach related infiltration rates were modelled with a conceived infiltration model based on real independent parameters and a constant groundwater table on an arbitrary level:

$$Q_{\inf} = \frac{A_S * LENGTH}{\sqrt{DATE\_CONSTR} * (100 * e^{2*POP\_DENS}) * (10 * SLOPE)^2}$$

where Q_{inf} = infiltration rate per single reach
A_S = groundwater-influenced pipe surface
$LENGTH$ = reach length
$DATE\_CONSTR$ = date of construction
POP_DENS = population density
$SLOPE$ = slope

For every sub-catchment 50 random distributions of five gauges were generated and compared with the optimised positioning. Bacher (1996) recommends 20 classifications as minimal number for comparison purposes. As optimisation indicator the error reduction r_{rdm} was computed:

$$r_{rdm} = 1 - \frac{\sum_{j=1}^{N} \left| Q_{subc-opt,j} - Q_{\inf,j} \right|}{\sum_{j=1}^{N} \left| Q_{subc-rdm,j} - Q_{\inf,j} \right|} \quad \text{with} \quad Q_{subc-opt/rdm,j} = l_j \frac{\sum_{i=1}^{M} Q_{\inf,i}}{\sum_{i=1}^{M} l_i}$$

where $Q_{subc\text{-}opt}$ = infiltration rate of a single reach for optimised gauges
$Q_{subc\text{-}rdm}$ = infiltration rate of a single reach for random gauges
Q_{inf} = modelled infiltration rate of a single reach
l = length of a single reach
N = total number of reaches
M = total number of reaches within a certain sub-catchment

The results are given in Table 2. The comparison shows an information improvement and error reduction, respectively, of in mean 14 % to maximum 35 %. The main reason for these results which seem to be mediocre is the fact, that the procedure is not a classification of independent objects. The network topology, especially number and size of meshes, as well as the spatial distribution of attributes as a result of the building history have a strong influence on the results. The developed method is a linear optimisation. Due to the non-linearity of the infiltration model cases with $r_{rdm} < 0$ occur, i.e. the random arrangement of gauges yields better results than the optimised arrangement. With a better knowledge about infiltration processes, the optimisation results might be superior. Since these cases are a clear minority, the optimisation procedure proofs to be rather reliable.

Table 2. Error reduction r_{rdm} for optimised gauges vs. random gauges.

Catch-ment	Area	Total sewer length	Number of meshes	Error reduction r_{rdm}		
				Minimum	Mean	Maximum
01G145	Dresden	23 km	9	-5 %	3 %	8 %
04F79	Dresden	23 km	2	-17 %	22 %	27 %
16P46	Dresden	11 km	4	2 %	15 %	33 %
F42	Emscher	3 km	0	-23 %	1 %	11 %
F46	Emscher	8 km	1	2 %	27 %	35 %
F50	Emscher	10 km	2	-3 %	13 %	23 %
mean		13 km	3	-7 %	14 %	23 %

In order to compare the several options of the optimisation method (standardisation and aggregation methods, distance measures) parallel calculations were done. Except some cases there are no vast differences between the options. Since the Euclidean and the Mahalanobis distance lead to similar results, the parameters are nearly uncorrelated.

The optimisation method was tested in three smaller catchments with relatively detailed information about infiltration rates. The application of the method results in good coherence between the optimised arrangement of gauges and detected changes of measured infiltration rates. Two examples are shown in Figure 3. In cases of a distinct correlation between optimised gauges and a change of the observed infiltration rate (western half of both catchments) either the calculated similarity figures are relatively small or a significant change of the φ–value occurs. Thus, one gauge separates at least one homogenous sub-catchment. From investigating these three networks it can be concluded, that the proposed optimisation method is able to identify sub-catchments with different infiltration rates in practice.

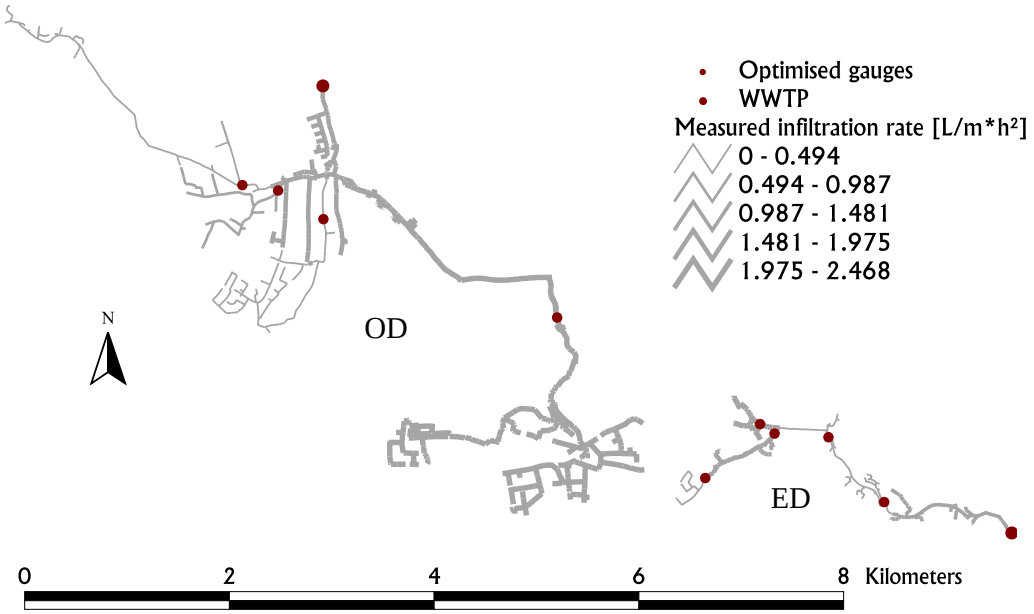

Figure 3. Optimised arrangement of gauges vs. detected changes of measured infiltration rates for the sub-catchments OD and ED.

Keeping in mind the unsatisfactory data situation as well as the strong influence of the artificial infiltration model and the network topology the verification of the optimisation method can be considered justified. For suitable catchments it is estimated, that the potential error reduction amounts up to 30 % compared to non-optimised gauge distributions. Suitable catchments where the optimisation potential is high are relatively well structured (number and size of meshes, distribution of reach properties) and have a distinct variation of infiltration dominated by diffuse sources. These constraints apply for larger catchments.

The proposed optimisation method is not based on typical input/output functions. It compares and classifies states. Therefore, the results are to be considered within the boundary conditions of the given data set. They will always have a significant uncertainty. Information about groundwater has an overwhelming relevance for their quality. Advantageously, a

definite parameter set is not necessary. The method can be adapted to nearly every data situation.

CONCLUSIONS

With comprehensive statistical analysis it could be shown, that (1) sub-catchments with different infiltration rates can be discriminated by means of structural data and groundwater information and that (2) larger catchments can be divided in more or less homogenous areas. Based on these findings a similarity approach "similar pipe characteristics lead to similar infiltration rates" is proposed. With the developed application of the similarity approach – the classification of reaches and sub-catchments and the optimisation method to identify gauge locations with a high information content – it is possible to improve the information about the infiltration status of sewer networks and to reduce the related uncertainty. The main advantage of this method is the very high degree of freedom against data needs.

ACKNOWLEDGEMENT

This study has been carried out within the framework of the European research project APUSS (Assessing Infiltration and Exfiltration on the Performance of Urban Sewer Systems) which partners are INSA de LYON (FR), EAWAG (CH), Dresden University of Technology (DE), Faculty of Civil Engineering at University of Prague (CZ), DHI Hydroinform a.s. (CZ), Hydroprojekt a.s. (CZ), Middlesex University (UK), LNEC (PT), Emschergenossenschaft (DE) and IRSA-CNR (IT). APUSS is supported by the European Commission under the 5th Framework Programme and contributes to the implementation of the Key Action "Sustainable Management and Quality of Water" within the Energy, Environment and Sustainable Development Contract n° EVK1-CT-2000-00072.

REFERENCES

APUSS (2004). Deliverable 11.3: Summarisation of Economic valuation. Project website, http://www.insa-lyon.fr/Laboratoires/URGC-HU/apuss/, visited 21 January 2005.

Bacher J. (1996). Clusteranalyse. (Cluster analysis): Oldenbourg: Muenchen, Wien. ISBN 3-486-23760-8.

Backhaus K., Erichson B., Plinke W. and Weiber R. (1996): Multivariate Analysemethoden (Methods of Multivariate Analysis.) Springer: Berlin, Heidelberg, New York. ISBN 3-540-60917-2.

Borg I. and Groenen P. (1997). Modern Multidimensional Scaling. Springer-Verlag New York Inc. ISBN 0-387-94845-7.

Davies J.P., Clarke B.A., Whiter J.T. and Cunningham R.J. (2001). Factors influencing the structural deterioration and collapse of rigid sewer pipes. *Urban Water*, **3**, 73-89.

Dupasquier B. (1999). Modélisation hydrologique et hydraulique des infiltrations d'eaux parasites dans les réseaux séparatifs d'eaux usées (Hydrological and hydraulic modelling of infiltration in separated sewers). PhD thesis, ENGREF Centre de Paris, 1999.

Fahrmeir L., Hamerle A. and Tutz G. (eds.) (1996). Multivariate statistische Verfahren. (Multivariate statistical methods.) Walter de Gruyter: Berlin, New York. ISBN 3-11-013806-9.

Gustafsson L.G., Lindenberg S. and Olsson R. (1991). Modelling of the indirect runoff components in urban areas. Proc. Int. Conf. on Urban Drainage and New Technologies, Dubrovnik, 1991, 127-133.

Gustafsson L.G. (2000). Alternative drainage schemes for reduction of inflow/infiltration – prediction and follow-up of effects with the aid of an integrated sewer/aquifer model. Proc. 1st Int. Conf. on Urban Drainage via Internet.

Kaufman L. and Rousseeuw P.J. (1990). Finding groups in data - an introduction to cluster analysis. Wiley: New York. ISBN 0-471-87876-6.

Lucas S. and Fuchs S. (2003). Regionalisierung von Fremdwasserproblemen (Regionalisation of extraneous water problems). *KA Abwasser Abfall*, **50**, 302-307.

Müller P.H. (ed.) (1991). Wahrscheinlichkeitsrechnung und mathematische Statistik. (Probability calculation and mathematical statistics.) Akademieverlag: Berlin. ISBN 3-05-500608-9.