APUSS

Assessing Infiltration and Exfiltration on the Performance of Urban Sewer Systems

Contract number : EVK1-CT-2000-00072 APUSS
Project homepage : http://www.insa-lyon.fr/Laboratoires/URGC-HU/apuss

DELIVERABLE 5.1

Classification method to identify sub-catchments for an efficient measurement procedure and transfer of measuring results

June 2004

Torsten Franz, Christian Karpf and Peter Krebs
Dresden University of Technology, Institute for Urban Water Management, 01062 Dresden
mail: Torsten.Franz2@mailbox.tu-dresden.de

INSA LYON
EAWAG
Swiss Federal Institute for Environmental Science and Technology
TECHNISCHE UNIVERSITÄT DRESDEN
Czech Technical University in Prague
Faculty of Civil Engineering
DHI WATER & ENVIRONMENT
CZ HYDROPROJEKT
MIDDLESEX UNIVERSITY
EMSCHER GENOSSENSCHAFT
IRSA CNR

Content

1 Objective

The goal of deliverable 5.1 is the development of a classification method for sewer systems, which identifies homogeneous sub-catchments.

One application of the method is the determination of an optimal positioning of infiltration and exfiltration measurement gauges. Due to mostly financial reasons the number of measuring points within a catchment is limited. It is necessary to find an optimal positioning in order to maximise the knowledge about the catchment.

Furthermore the classification should be used for the transfer of measurement results within a catchment. According to the definition of milestones (Franz and Krebs, 2003), this application is part of deliverable 5.2.

As a result of the APUSS management committee meeting held in London (APUSS, 2004) the leakage approach determining the infiltration situation on a small scale was included in our work.

2 Classification method

2.1 Similarity approach

2.1.1 Basic assumption

The knowledge about cause-effect relationships between pipe characteristics, pipe environment and structural state of pipes on the one hand and the structural state and I/E rates on the other hand is fairly limited. Only knowledge about qualitative relationships is available (Rutsch, 2003). Due to the enormous number of relevant processes and influencing factors and due to the general data situation of large sewer systems there is a significant lack of data. Thus, both deterministic models as well as statistical models do not seem to be feasible to determine the infiltration and exfiltration situation in an urban catchment. A different approach is necessary.

With the assumption “similar pipe conditions lead to similar infiltration/exfiltration rates” it is possible to look for and work with “similarities” within a sewer system. With a procedure containing classification and generalisation of homogeneous areas and homogeneous groups of reaches, respectively, these groups are comparable and information should be transferable.

Models and procedures based on this approach do not have typical input/output functions. They compare and classify states. Therefore, the results are to be considered within the boundary conditions of the date set. They will always have a significant uncertainty. A definite parameter set is advantageously not necessary. The models can be adapted to nearly every data situation.

2.1.2 Verification of the basic assumption

The first step of the implementation of the similarity approach was the verification of the basic assumption “similar pipe conditions lead to similar infiltration/exfiltration rates”. By means of typical data sets and methods of classical statistics and exploratory data analysis similarities and dissimilarities of sub-catchments with different infiltration rates were studied.

The following data sets were available:

- 22 sub-catchments of Dresden catchment without groundwater information
- 6 sub-catchments of Dresden catchment with groundwater information
- 5 sub-catchments of Emscher catchment with groundwater information.

The parameters considered are listed in Table 1. For every sub-catchment the infiltration rate was known by measurements.

Table 1: Independent Parameters

abbreviation	parameter
both catchments	
DATE_CONSTR	date of construction
FUNCTION	function (regional, main, tributary)
MATERIAL	material
SEW_SYSTEM	sewer system (foul, combined)
PROF_TYP	profile type (egg, circle, other)
PROF_CIRC	profile circumference
LENGTH	reach length
POP_DENS	population density
POP_LENGTH	population-spec. length
DIST_WATER	distance to surface water
DIST_BUILD	distance to buildings
STREET_TYP	street type (main, residential, …)
DIST_STREET	distance to streets
SLOPE	slope
COVERAGE	coverage[1]
Dresden catchment only	
DIST_ELBE	distance to river Elbe
DIST_STORM	distance to storm sewers
DIST_DRAIN	distance to drainage
THICK_COHSV	thickness of cohesive layers
GW_LEVEL	distance to groundwater[2]
Emscher catchment only	
K	soil permeability
A_IAREA	reduced area ratio
NO_JOINTS	number of joints

[1] incl. wall thickness
[2] based on only one measurement campaign

First, the characteristics of the sub-catchments were compared. An “all-in-one” test was not available, because the considered parameters belong to different statistical scales (these are: metric (numbers like SLOPE), ordinal (ranks like STREET_TYP), nominal (names like MATERIAL), q.v. Walford, 1995). Therefore, every parameter was analysed individually with one-way ANOVA (metric scale), Kruskal-Wallis-ANOVA (ordinal scale), and contingency tables (nominal scale), respectively (Mueller, 1991). These analysis-of-variance methods compare mean values of groups to verify a significant inter-group difference. For every parameter it could be shown that the reach populations of the sub-catchments differ significantly from each other in their characteristics. Due to the wide range of observed infiltration rates (minimum:maximum = 1:16) it seemed to be probable, that the dissimilarities between the reach populations are linked in some way with the infiltration rates.

Second, 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. The distances between the objects are a measure for dissimilarity. The sub-catchments can be seen as objects in an

n-dimensional space with n as number of parameters. Compared with other multivariate methods the MDS has the advantage, that parameters of all scales can be handled together (see Borg and Groenen, 1997, Fahrmeir *et al.*, 1996).

The MDS result of the Dresden data set without groundwater is shown in Figure 1. Due to the relatively high number of objects the parameter set was reduced to two dimensions. A relation between these dimensions and the infiltration rate could not be found. But, a relation between the parameter set and the quarter type (e.g. inner centre, broader centre, suburb) could be found. Furthermore, the parameter set was reduced to one dimension. Significant correlations between this dimension and metric parameters are shown in Table 2. It can be concluded, that the parameters of the sewer system – esp. DATE_CONSTR, PROF_CIRC, POP_DENS, POP_LENGTH and DIST_STORM – can be used as an indicator for the degree of urbanisation and urban development, respectively.

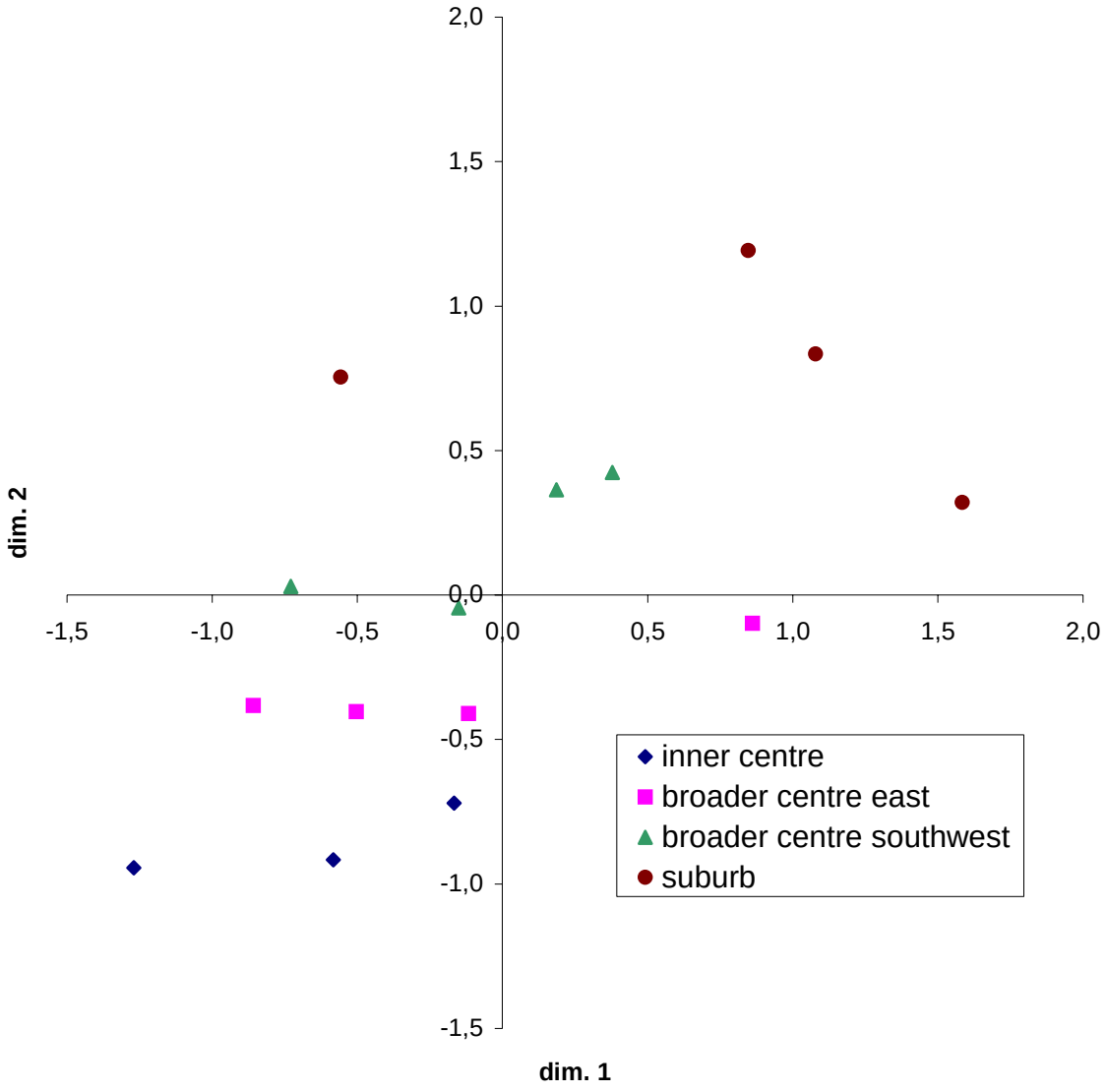

Figure 1: MDS result of Dresden data set without groundwater information

Table 2: Significant correlations between MDS results and parameters

parameter	correlation r
DATE_CONSTR	0.74
PROF_CIRC	-0.92
POP_DENS	-0,88
POP_LENGTH	0,73
DIST_WATER	-0,67
DIST_STORM	-0,83
THICK_COHSV	0,56
SLOPE	0,65
COVERAGE	0,71

dim. 1(centre) < dim. 1(suburb)

The MDS result of the Dresden data set with groundwater information is shown in Figure 2. Only reaches influenced by groundwater were considered. The parameter set was reduced to one dimension. A good correlation between reach-surface-specific infiltration rate q_f and the parameter set was found. A comparison with the time weighted head shows, that this correlation, i.e. the specific infiltration rate is not dominated by groundwater (Figure 3).

An MDS of the Emscher data was not reasonable due to the low number of available catchments. The analysis of a combination of the Dresden and the Emscher data set is shown in Figure 4. Since more data are included into the analysis the different pattern leads to different values of dim. 1 compared to Figure 2. The MDS does not show better or different results. Thus, a transfer of analysis results between catchments seems not be reasonable (see Franz, 2004).

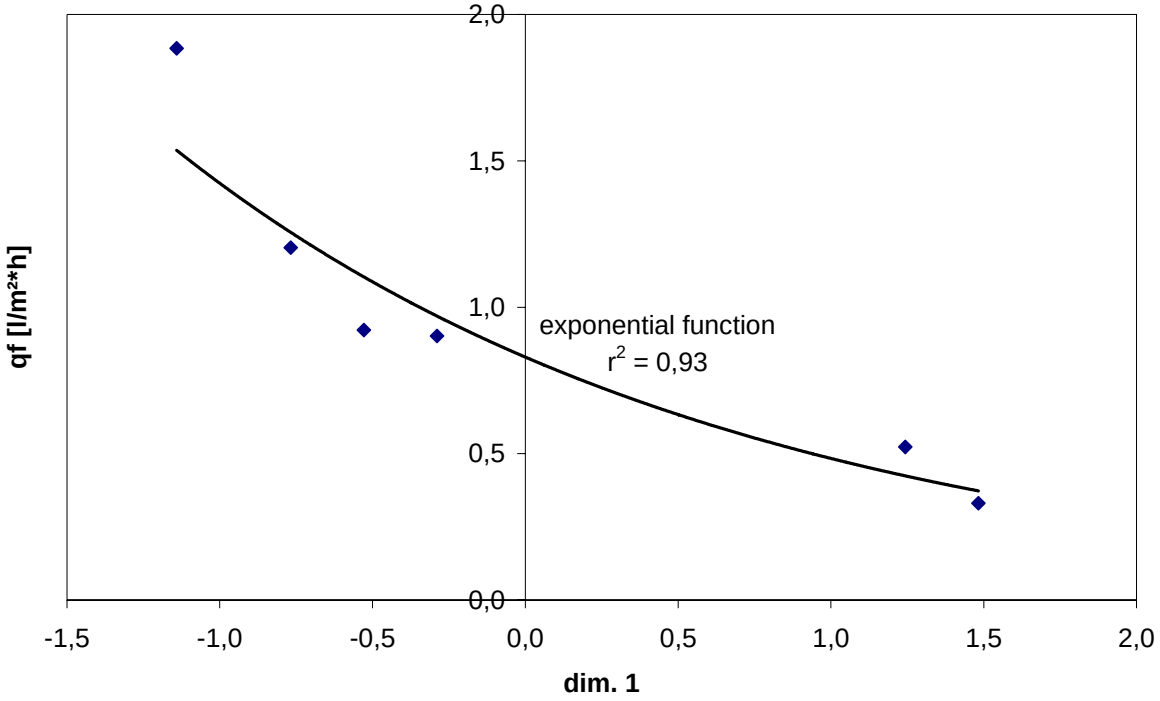

Figure 2: MDS result of Dresden data set with groundwater information vs. infiltration rate

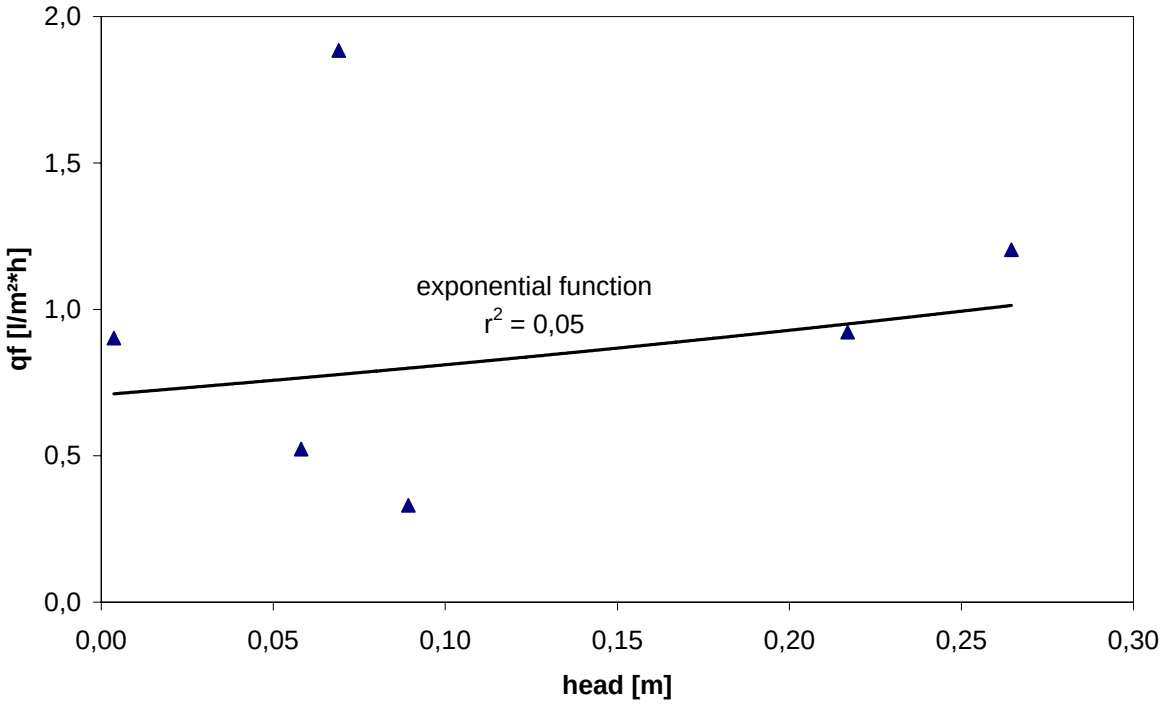

Figure 3: Head vs. infiltration rate

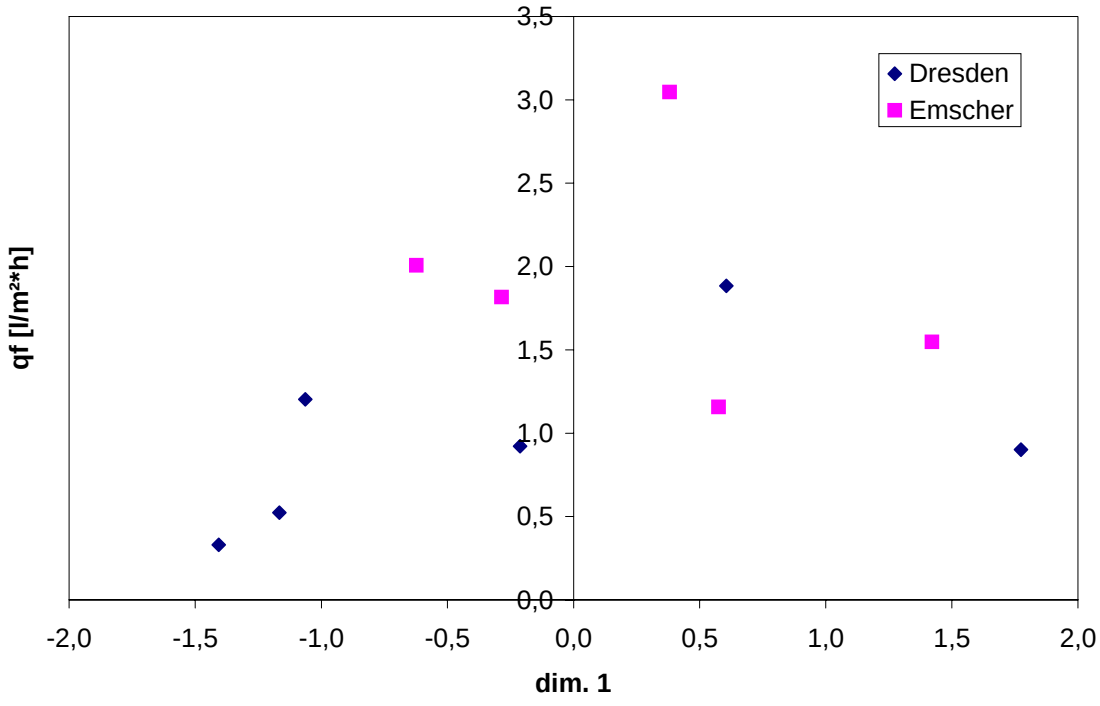

Figure 4: MDS result of a Dresden/Emscher data set with groundwater information vs. infiltration rates

2.1.3 Conclusion

Because of the reasonable correlation between the dimension 1 standing for the independent parameters and the infiltration rate (Figure 2) it can be concluded that (i) there is a recognisable relationship between the independent parameters and the infiltration rate and

(ii) that the parameters - or a part of them - are sufficient to describe infiltration. Thus, the basic assumption of the similarity approach "similar pipe conditions lead to similar infiltration/exfiltration rates" can be the fundament for further research.

2.2 Method development

2.2.1 Cluster analysis as related mathematical method

2.2.1.1 Overview

One necessity of the similarity approach is the identification of homogeneous reach groups, i.e. the description of the homogeneity of reach groups. This is related to several multivariate statistical methods like cluster analysis, discriminant analysis, factor analysis or multidimensional scaling.

Potential procedures and tools can be adapted esp. 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 distinctly differ. For an overview see Kaufman and Rousseeuw (1990) or Bacher (1996). A simplified algorithm of the method is shown in Figure 5.

A one-to-one application of cluster analysis is not feasible. Regarding infiltration/exfiltration the dimensional characteristics (network structure) and the anisotropy (flow direction) of sewer systems must be considered. Thus, the objects (reaches) are not independent from each other. For the classification procedure standardisation, the determination of distance and proximity measures, respectively, and homogeneity measures are of outstanding interest.

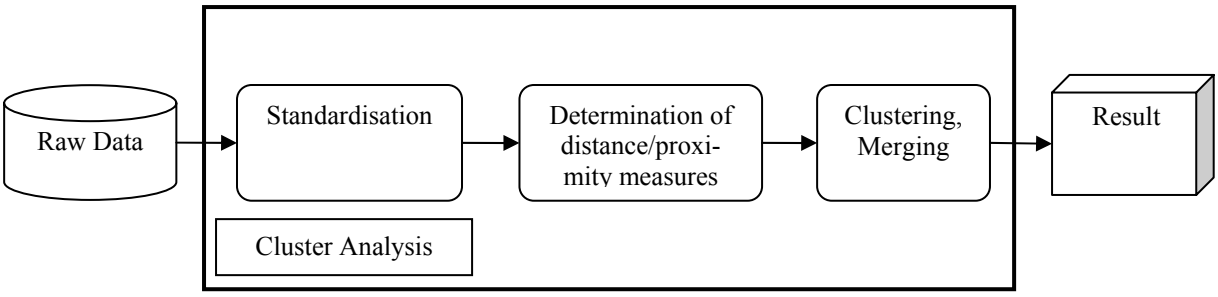

Figure 5: Algorithm of Cluster Analysis (Kurth, 2004)

2.2.1.2 Standardisation

Distance measures for metric parameters are often not scale invariant. Therefore, all parameters must be "comparable" and have a similar range. Otherwise the parameters would have an unintentional weighting. The problem of different ranges is solved with standardisation.

The following options are common, where

x_i = raw value
z_i = standardised value
s_i = empirical standard deviation

1. Standardisation to characteristic values like mean, median, maximum

$$z_i = \frac{x_i}{\bar{x}} \qquad \text{Equation 1}$$

2. Standardisation to extreme values

$$z_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \qquad \text{Equation 2}$$

3. Standardisation to variances (z-transformation)

$$z_i = \frac{x_i - \bar{x}}{s_i} \qquad \text{Equation 3}$$

4. Standardisation to relevance (weighting)

$$z_i = w_i * x_i \qquad \text{Equation 4}$$

Options 1 and 2 standardise the range only, the z-transformation standardises range and distribution. The weighting is only used in combination with other options. All options can be done with theoretical or empirical values, i.e. possible or measured values.

2.2.1.3 Distance and proximity measures

In most cases similarities of objects are determined indirectly by calculating distance and proximity measures, respectively. With N objects and d as distance and s as proximity measure, the following applies:

$$\left.\begin{aligned} &d(n,n) = 0 \\ &d(n,m) \geq 0 \\ &d(n,m) = d(m,n) \end{aligned}\right\} \text{ for } n, m = 1, \ldots, N$$

$$\left.\begin{aligned} &s(n,m) \leq s(n,n) \\ &s(n,m) = s(m,n) \end{aligned}\right\} \text{ for } n, m = 1, \ldots, N$$

Distance measures can be transformed in proximity measures and vice versa (Fahrmeir *et al.*, 1996). The approach for the measures differs depending on the statistical scale.

In general, distance measures are appropriate when the absolute distance of objects is relevant, i.e. the dissimilarity between the objects. Proximity measures are appropriate, when the shape of parameter profile is of interest rather than the level, i.e. the correlation between

the objects (Backhaus, 1994). Therefore, only distance measures are relevant for our purpose and described in the following. For a comprehensive overview see Bacher (1996).

Nominal parameters

If the nominal parameter has just two possible values (binary: 0, 1) there are four possible situations for two objects (Table 3).

Table 3: Contingency table for two objects n, m with nominal binary parameter *p*

	p(n) = 1	*p(n)* = 0
p(m) = 1	*a(n,m)*	*c(n,m)*
p(m) = 0	*b(n,m)*	*e(n,m)*

The distance is then calculated with

$$d(n,m) = 1 - \frac{a + \delta * e}{a + \delta * e + \lambda(b + c)} \qquad \text{Equation 5}$$

The weighting parameters δ and λ characterise a multitude of distance measures, e.g. simple matching for δ=λ=1. For a comprehensive list see Steinhausen and Langer (1977).

Nominal Parameters p_i with more than two values can be binary coded (Table 4) or treated with the generalised M-coefficient:

$$d(n,m) = 1 - \frac{a}{\sum p_i} \qquad \text{Equation 6}$$

where a = number of matching components.

Ordinal parameters

Ordinal parameters can be binary coded (Table 4) and treated as nominal binary parameters, or they can be treated as metric parameters. In that case they are normally standardised on an interval [0, 1].

Table 4: Binary coding of nominal and ordinal parameters

parameter type	**values**	**Coding**
nominal	$x_1 \neq x_2 \neq x_3$	$x_1 = (1,0,0)$
		$x_2 = (0,1,0)$
		$x_3 = (0,0,1)$
ordinal	$x_1 < x_2 < x_3$	$x_1 = (1,0,0)$
		$x_2 = (1,1,0)$
		$x_3 = (1,1,1)$

Metric parameters

Distance measures for metric parameters are deduced from the generalised Minkowki-metrics (Fahrmeir *et al.*, 1996):

$$d(n,m) = \left(\sum_{i=1}^{p} \left| x_i(n) - x_i(m) \right|^r \right)^{1/q} \qquad \text{Equation 7}$$

where p = number of parameters
$r \geq 1, q \geq 1$

The use of city-block-metrics (r = 1, q = 1), Euclidean distance (r = 2, q = 2) and squared Euclidean distance (r = 2, q = 1) is common.

Another measure is the Mahalanobis distance (Fahrmeir *et al.*, 1996):

$$d_{nm} = \sqrt{(n-m)^T S^{-1} (n-m)} \qquad \text{Equation 8}$$

where S = empirical covariance matrix

Compared to Minkowki-metrics 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. The transformed vectors $S^{-1}(n-m)$ are empirically uncorrelated. The use of this distance measure is equivalent to a transformation of the characteristics to uncorrelated ones (e.g. with principal components analysis) followed by the calculation of the Euclidean distance. Thus, the significance enhancement of some parameters due to correlation effects is prevented. The Mahalanobis distance cannot be calculated in case of standard deviation s = 0.

Parameters with different scales

There are several possibilities to combine parameters with quantitative and qualitative scales. First, the parameters with higher scales can be transformed to the lowest one (level regression, e.g. binary coding). That means a loss of (mostly relevant) information. Second, the parameters with lower scales can be transformed to higher ones (level progression, e.g. ordinal parameters defined as metric). It is to prove carefully, whether this is acceptable.

Another possibility is the calculation of a weighted mean of all distances:

- ordinary mean:

$$d(n,m) = \frac{1}{p} \left(a_N d^N(n,m) + a_O d^O(n,m) + a_M d^M(n,m) \right) \qquad \text{Equation 9}$$

where p = number of parameters
a = number of parameters of a certain type

- Gower coefficient

$$d(n,m) = \frac{\sum_{i=1}^{p} \delta_k(n,m) * d_k(n,m)}{\sum_{i=1}^{p} \delta_k(n,m)} \qquad \text{Equation 10}$$

where p = number of parameters
δ = weighting factor, see Gower (1971)

2.2.1.4 Homogeneity measures

The existence of group homogeneity can be proved by testing expected probability distributions (Bacher, 1996). For comparing the homogeneities of groups G_i a measure $h(G_i) > 0$ must be defined, which is the smaller the more homogeneous the group is.

Hartung and Elpelt (1995) propose

- mean distance between all objects

$$h(G_i) = \frac{1}{c} \sum_{\substack{n<m \\ n,m \in G_i}} d(n,m) \text{ with } c = |G_i| \text{ or } c = |G_i| * (|G_i| - 1) \qquad \text{Equation 11}$$

- mean distance to the centroid
- maximal observed distance between two objects
- minimal observed distance between two objects
- sum of parameter variances

$$h(G_i) = \sum_{j=1}^{p} s_j^2 \qquad \text{Equation 12}$$

2.2.2 Similarity Figure φ

2.2.2.1 Algorithm

Due to the basic assumption “similar pipe conditions lead to similar infiltration/exfiltration rates” the classification method should identify similar groups of reaches, i.e. the method is based on a similarity measure – with consideration of the nature of sewer systems and focused on in- and exfiltration. It follows that one measure describing the homogeneity for every possible group of reaches must be calculated. Sub-catchments or parts of the sewer network can be separated by means of this similarity figure φ.

The determination of group members depends on the object of investigation. For infiltration all reaches of the upstream sub-catchments are relevant, whereas for exfiltration it is a number of connected reaches.

In the following, an algorithm (Figure 6) to calculate the similarity figure is described. For every step several options are proposed. Thus, the calculation can be adapted to any individual data set.

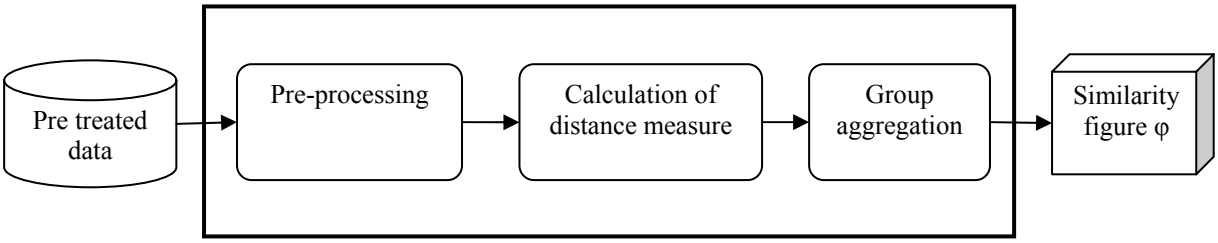

Figure 6: Algorithm for determining the similarity figure

2.2.2.2 Pre-processing

The input is a data set containing all relevant reaches with their attributes. This data set should be pre-treated (outlier identification, parameter relevance etc.).

The proposed pre-processing procedures are shown in Table 5. Two options were selected, which standardise range on the one hand and range and spreading on the other. Ordinal parameters are treated as metric ones. It is estimated, that the influence of this arbitrary information increase is not so profoundly compared with the information loss of level regression. Ordinal parameters with N values must be coded in classes c with c = 1, 2, … N.

It is possible to weight parameters and objects. Parameters are weighted due to their relevance for the object of investigation, e.g. infiltration. Because of the current knowledge (Franz, 2004) the weighting of parameters for in-/exfiltration purposes is not recommended.

Using the in-/exfiltration model (APUSS, 2003)

$$Q_{in/ex} = k_{L,in/ex} * A_S * \Delta h \qquad \text{Equation 13}$$

where Qin/ex = in-/exfiltration flow
$k_{L,in/ex}$ = leakage factor
Δh = head between water level in sewer pipes and groundwater level

k_L can be seen as expression of the parameters described in Table 1. Because of $Q_{in/ex} \sim \Delta h$ the reaches should be weighted due to the influence of water head Δh, where it is available. Further weights like damage number (Rutsch and Uibrig, 2003) or classes are possible.

Table 5: Pre-processing

standardisation	nominal p.	ordinal p.	metric p.
none	X		
standardisation to extreme values		X	X
standardisation to variances		(X)	X
weighting	X	X	X

2.2.2.3 Distance measures

The generalised M-coefficient is proposed for nominal parameters. For metric and ordinal parameters the Euclidean and the Mahalanobis distance are proposed. While using the Mahalanobis measure it has to be proved whether the constraint standard deviation = 0 for all parameters is fulfilled.

2.2.2.4 Aggregation

The following aggregations are proposed:

- mean distance between all reaches
- mean distance to the centroid
- mean distance to the centroid of the 5 % most distant reaches

The first aggregations emanate from an equal distribution of in-/exfiltration rates. The last aggregation implies, that a few reaches, e.g. the most damaged, have an outstanding importance for in-/exfiltration rates. The percentage of 5 % is arbitrary. Whether it should be higher will be dealt in further investigations.

2.2.2.5 Derived properties of φ

The similarity figure is calculated for a group of reaches. The separating element between two reaches is a manhole, therefore φ should be linked to manholes. The uppermost and the second uppermost manhole of a network have a value $\varphi = 0$. For manholes inside a mesh the similarity figure is not defined, because an infiltration rate measured in a mesh cannot be allocated definitely to reaches.

3 Optimal positioning of measurement gauges

3.1 Target functions

Measurement gauges that are positioned optimally are associated with – to the greatest possible extent – homogeneous reach populations. Therefore, the optimal position can be identified by finding the minimal similarity figure. Besides this, further boundary conditions have to be considered:

- Sensors require a minimal flow or sewer length to measure reliably.
- The number of gauges is limited due to financial restrictions.
- The distribution of gauges within a catchment can be regular or not.

In the following an optimisation algorithm is explained and discussed. Due to the data situation the focus lies on infiltration. The optimisation for exfiltration measurements differs only in the selection of relevant reaches and parameters.

3.2 Catchment-wide optimisation

3.2.1 Algorithm

With an iterative procedure the catchment is divided until a given number of sub-catchments, i.e. measurement gauges is reached (Figure 7). The most downstream manhole (standing for a WWTP) is automatically a gauge. For every separation step one manhole is identified, which divides the (sub-)catchment into two sub-catchments. The conditions for the separating manhole are:

1. The similarity figure φ is minimal: A minimal φ stands for a maximal possible homogeneity of the upstream reaches.
2. 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.
3. The manhole is not within a mesh: An infiltration rate measured in a mesh cannot be definitely allocated to reaches.

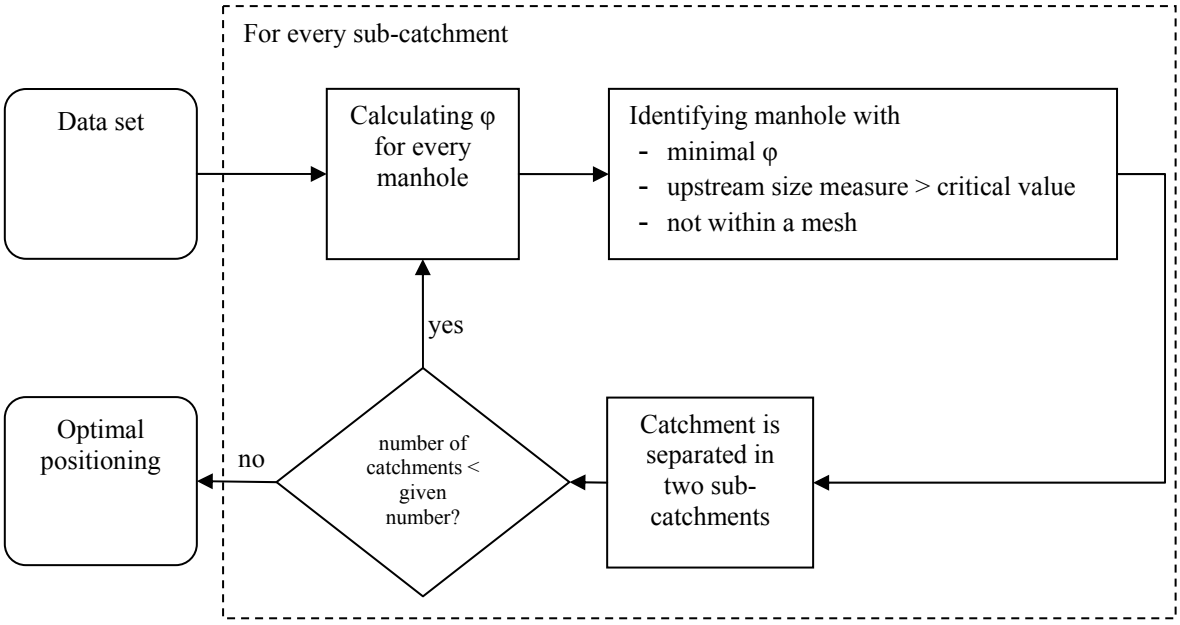

Figure 7: Algorithm of the Simple optimisation

3.2.2 Verification

3.2.2.1 Framework

Investigated cases

The following cases were considered:

1. Theoretical cases (chapter 3.2.2.2): Theoretical cases have either randomized infiltration rates or artificial parameters. They were used to investigate extreme states.
2. Cases with artificial infiltration rates (chapter 3.2.2.3): Infiltration data with a high spatial resolution were not available. By using (arbitrary) infiltration models it was possible to verify the optimisation procedure on reach scale. The question remains how far away from reality these models are.
3. Cases with real infiltration rates (chapter 3.2.2.4): Relatively detailed infiltration rates where available for a few catchments.

Available data

For six real data sets containing the parameters date of construction, profile circumference, length, population density, slope and head, three infiltration models were implemented:

- model "random": randomized rates
- model "simple": $Q_{\inf} = \prod p_i$ with head = const.
- model "not so simple": $Q_{\inf} = \frac{A_S * LENGTH}{\sqrt{DATE\_CONSTR} * (100 * e^{2*POP\_DENS}) * (10 * SLOPE)^2}$

  where A_S = groundwater-influenced pipe surface

Furthermore three catchments with real infiltration rates were considered as well as two networks filled with artificial independent parameters and infiltration rates (Table 6).

Table 6: Data sets

catchment	part of	length [km]	artificial parameters	artificial Q_{inf}	real Q_{inf}
01G145	Dresden	22.9		X	
04F79	Dresden	22.6	X	X	
16P46	Dresden	10.9	X	X	
ED	Dresden	9.3			X
OD	Dresden	37.4			X
SF	Dresden	6.9			X
F42	Emscher	2.7		X	
F46	Emscher	7.7	X	X	
F50	Emscher	9.8		X	

Target functions

For every catchment the best combination of options for calculating φ was applied. The option "mean distance between all reaches" was not considered, because the difference to "mean distance to centroid" should be minimal. The head was used as weighting factor where it was available. In the following the chosen combination is marked with a four letter code:

- standardisation: e – to extreme values, v – to variances
- weighting: n – no weigthing, f – as factor
- distance measure: e – Euclidean, m – Mahalanobis
- aggregation: m – mean distance to centroid, x – maximal mean distance to centroid.

The number of gauges was set to five. As size measure the sewer length was chosen. The optimisation was done for equally und unequally sized sub-catchments. i.e. either the critical value was set as a constant to 2 % of the total length or it was variable due to the given number of gauges.

Measure of optimisation quality

The error reduction compared to the mean error was chosen as indicator of optimisation quality. With

$$Q_{WWTP,i} = l_i \frac{\sum_{j=1}^{N} Q_{\inf,j}}{\sum_{j=1}^{N} l_j} \qquad \text{Equation 14}$$

and

$$Q_{subc,i} = l_i \frac{\sum_{j=1}^{M} Q_{\inf,j}}{\sum_{j=1}^{M} l_j}$$ Equation 15

where $Q_{WWTP,i}$ = infiltration rate for reach *i* based on mean catchment rate
$Q_{subc,i}$ = infiltration rate for reach *i* based on mean sub-catchment rate
$Q_{inf,j}$ = modelled infiltration rate for reach *j*
l_i = length of reach *i*
N = total number of reaches
M = number of reaches within a certain sub-catchment

the optimisation quality indicator *r* is calculated to

$$r = 1 - \frac{\sum_{j=1}^{N} \left| Q_{subc,j} - Q_{\inf,j} \right|}{\sum_{j=1}^{N} \left| Q_{WWTP,j} - Q_{\inf,j} \right|}$$ Equation 16

3.2.2.2 Results for theoretical cases

The infiltration model "random" causes no links between independent parameters and infiltration rates. The optimisation quality should be minimal. For the catchments 04F79, 16P46 and F46 with the infiltration model "random" (see chapter 3.2.2.1) the error reduction was $r < 1$ % (Equation 16).

The networks of the catchments 01G145, 16P46 and F46 were filled with two parameters each with two to three values. The infiltration was calculated with model "simple". The result was a very structured catchment, i.e. the specific infiltration rates were distributed without extreme discontinuities (Figure 8). The optimisation quality should be good. In fact, the results were widely spread (Table 7). Considering the uniform distribution of infiltration rates, the network structure seems to be very important for the optimisation result.

Table 7: Results for completely artificial data sets

catchment	error reduction *r*
01G145 (enem)	19 %
16P46 (enex)	95 %
F46 (enem)	45 %

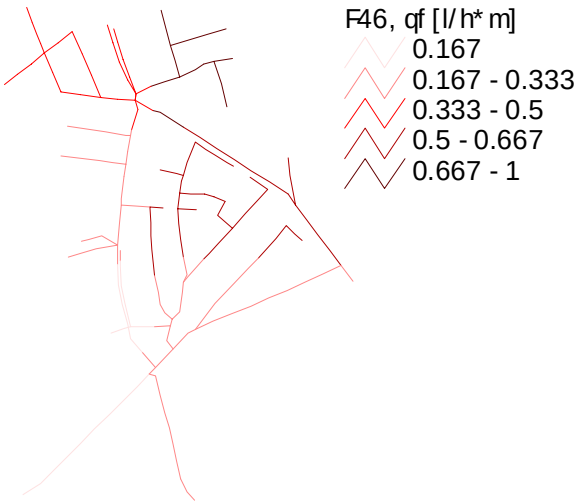

Figure 8: Length-specific infiltration rates for F46

3.2.2.3 Results for cases with artificial infiltration rates

Table 8 shows the error reduction *r* for the target functions for equally and unequally sized sub-catchments and the infiltration models "simple" and "not so simple" (see chapter 3.2.2.1). With a mean value of $r = 22\ \%$ and a maximal value of $r = 39\ \%$ the results of the optimisation are not as good as expected.

The main reason for these results is the fact, that the procedure is not a classification of independent objects. The net structure, esp. kind and number of meshes, has overwhelming influence on the optimisation, because manholes within a mesh cannot be used as separating structures. In Figure 9 the net structures of catchments 01G145 (with bad results) and 04F79 (with better results) are compared. Catchment 01G145 has much more meshes spread in a wider area than catchment 04F79.

Furthermore, the infiltration models have an influence on the results. The relative difference between the optimisation based on the two models is approx. 20 %. With other models the results of optimisation might be superior.

The target function on the sub-catchment size has also a distinct influence. The unequal distribution of gauges leads to worse results.

Table 8: error reduction *r* for optimised gauges vs. WWTP

catchment	equal sewer length		unequal sewer length	
	model "simple"	model "not so simple"	model "simple"	model "not so simple"
01G145 (efmm)	10 %	12 %	5 %	12 %
04F79 (efem)	38 %	30 %	38 %	26 %
16P46 (efmm)	31 %	18 %	4 %	17 %
F42 (efem)	11 %	39 %	catchments too small length < 10,000 m.	
F46 (vfem)	24 %	32 %		
F50 (efmm)	16 %	34 %		

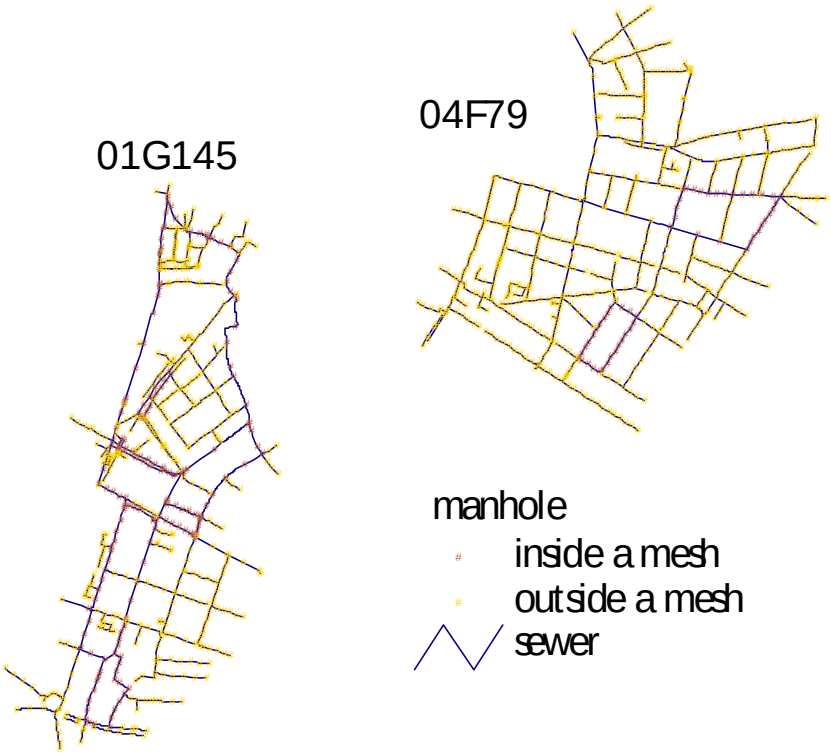

Figure 9: Network structure of catchments 01G145 and 04F79

The quality indicator r (Equation 16) does not reflect the practical situation that gauges already exist. The comparison between error of optimised gauges (Equation 15) and error of WWTP (Equation 14) neglects the additional information gained by the pure number of gauges. Therefore, 50 random distributions of gauges were generated for every catchment except those with an unfavourable net structure (too small, too much meshes). Bacher (1996) recommends 20 classifications as minimal number for comparing purposes.

With the random gauges the error reduction r_{rdm} is calculated

$$r_{rdm} = 1 - \frac{\sum_{j=1}^{N} \left| Q_{subc,j} - Q_{\inf,j} \right|}{\sum_{j=1}^{N} \left| Q_{subc-random,j} - Q_{\inf,j} \right|} \qquad \text{Equation 17}$$

where $Q_{subc\text{-}random,j}$ = infiltration rate for reach j based on random gauges

The optimisation results are given in Table 9. Obviously $r_{rdm} < r$ applies. Inconsistencies of boundary conditions between optimised and random gauges are the reason for cases with $r_{rdm} < 0$, i.e. the random arrangement of gauges yields better results than the optimised arrangement. Since these cases are a clear minority, the optimisation procedure proofs to be rather reliable. The influence of the net structure is indicated by the range of the results.

Table 9: error reduction r_{rdm} for optimised gauges vs. random gauges

catchment	minimal r_{rdm}	mean r_{rdm}	maximal r_{rdm}
01G145		unfavourable net structure	
04F79 (efem)	-10 %	5 %	16 %
16P46 (efmm)	-3 %	4 %	11 %
F42		unfavourable net structure	
F46 (vfem)	4 %	15 %	22 %
F50 (efmm)	0 %	29 %	34 %

Based on the numerous calculations the following recommendations for calculating the similarity figure φ can be given:

- standardisation: to extreme values
- weighting: when available
- distance measure: Mahalanobis distance when possible
- aggregation: mean distance to centroid.

Except some cases there are no vast differences between the options.

3.2.2.4 Results for cases with real infiltration rates

The main disadvantage of artificial infiltration rates are their unpredictable influence on the optimisation quality indicator *r* and therefore on the evaluation of the optimisation procedure. In order to gain an impression of the applicability of the optimisation in reality, several catchments with a relatively high number of inside infiltration gauges and unequal infiltration rates were investigated. From the available data three catchments of Dresden could be used, only.

Due to the different dependent parameter (infiltration rate for a sub-catchment, not per reach) another kind of verification was used: For every catchment an optimisation was performed generating approx. 50 % more sub-catchments than measuring gauges. Then, the separating similarity figure (minimal φ–value for the individual sub-catchment) was compared with the length-specific infiltration rate.

With this procedure it was possible not just to determine the optimal positioning of gauges (recognisable by changing of the φ-value), but also to compare the results with the measurements. The similarity figure is used as measure for discriminatory power, i.e. the separation between sub-catchments with different φ-value is stronger than with similar ones. Thus, sub-catchments with a similar φ-value could be merged with a moderate information loss.

The results for the catchments SF and OD are shown in Figure 10 and Figure 11. The marked sub-catchments are determined by a measurement gauge and have got a different infiltration rate than the surrounding sub-catchments. They contain several optimised sub-catchments, i.e. gauges, all of them with a similar low φ-value. That means, that their centroid might be different, but their parameter distribution is low. Therefore, a high homogeneity remains

while merging the optimised sub-catchments to the measured sub-catchment. The remaining sub-catchments are characterised by high similarity figures φ, i.e. by a relative significant heterogeneity among those larger sub-catchments. Partly they contain areas with low φ-values.

For the marked cases it can be concluded, that with the change of the infiltration rate there is concurrently a significant change of the similarity figure and the homogeneity of the reaches, respectively.

In the western part of the catchment ED the optimised gauges are nearly identical with the measuring gauges, which represent sub-catchments with a wide range of different infiltration rates (Figure 12). Detailed conclusions for the eastern part cannot be drawn. These results are affirmed by the comparison between infiltration rates and similarity figures (Figure 13). The similarity figures are low in the western part and high in the eastern part of the catchment.

From investigating these three real networks it can be concluded, that the proposed optimisation method is able to identify sub-catchments with different infiltration rates.

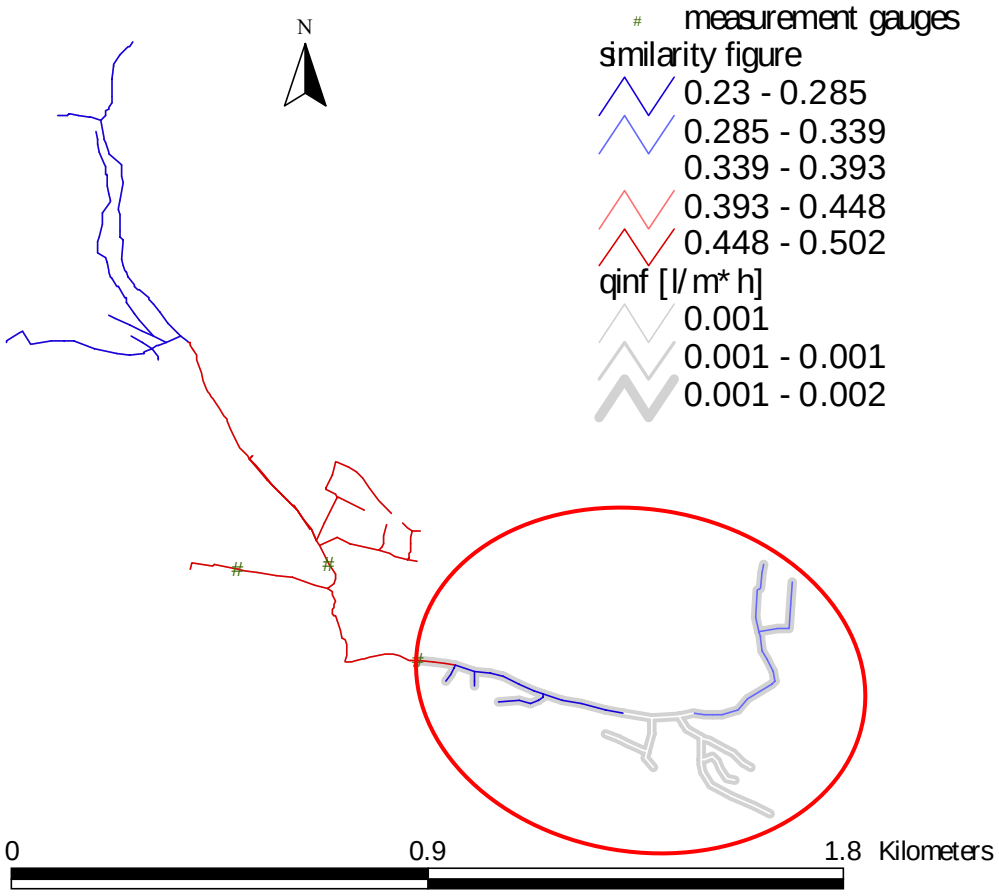

Figure 10: Infiltration and optimisation for catchment SF

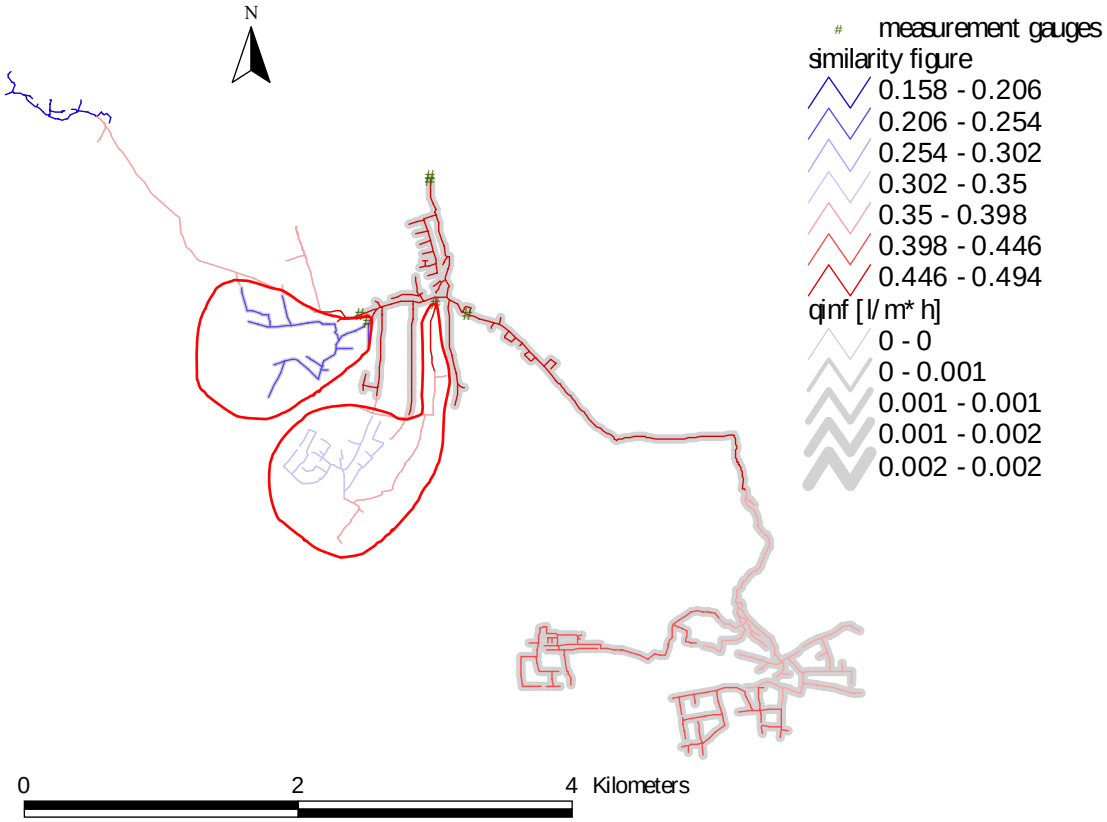

Figure 11: Infiltration and optimisation for catchment OD

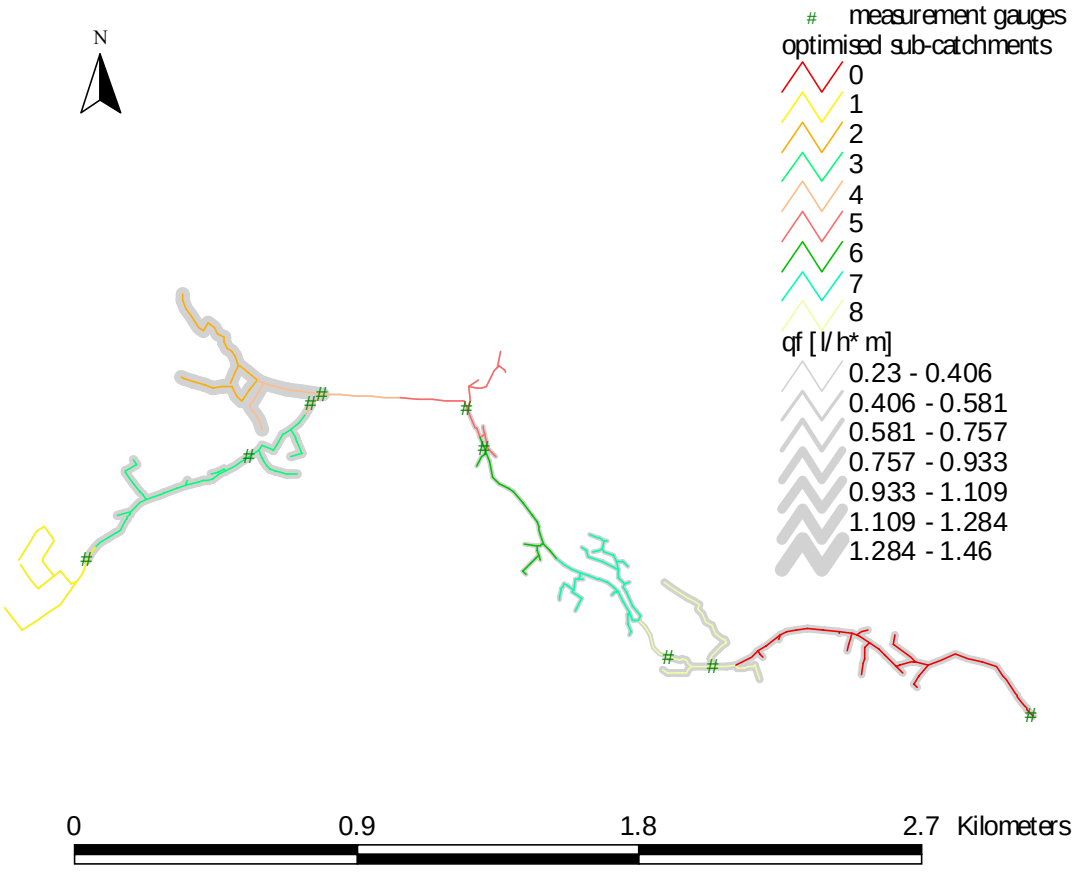

Figure 12: Infiltration and optimised sub-catchments of catchment ED

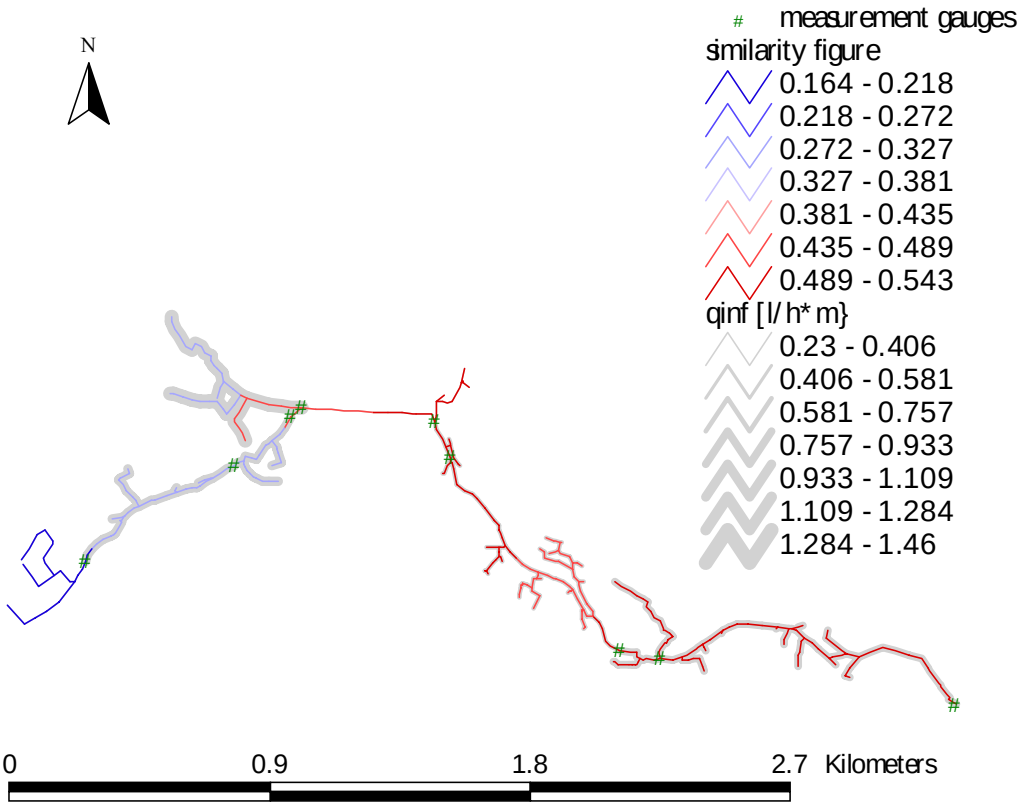

Figure 13: Infiltration and similarity figures of catchment ED

3.3 Conclusion

With the proposed application of the similarity approach – the classification and optimisation method due to the similarity figure φ – it is possible to improve the information about the infiltration status of sewer networks. For suitable catchments the error reduction amounts up to 30 % compared to arbitrary gauge distributions. Suitable catchments are relatively well structured (degree of meshes and ramification, distribution of reach properties) and have infiltration dominated by diffuse sources. These constraints should apply for very large catchments (compare results of chapter 2.1.2).

Keeping in mind the influence of the artificial infiltration models and the net structure as well as the numerous options and target functions the positive verification of the methods is justifiable but not completely satisfying.

The results point to other imaginable optimisation approaches, e.g. reducing the number of gauges and comparing the information loss with financial benefits. This will be dealt with in further investigations.

4 Blind alleys

During developing the procedures of chapter 2.2 and 3.2 several methods were investigated which do not lead to results or were not applicable.

Geostatistics

Geostatistical methods were developed by geosciences. Typical applications are the investigation of mining or contaminated sites. Under the conditions of natural investigation objects geostatistics lead to better results than classical statistics, e.g. the use of semi-variograms and kriging compared to calculating the mean. For an overview see Wackernagel (1998), Armstrong (1998) or Walford (1995).

Geostatistical methods are conditionally suitable for quasi-one-dimensional objects like sewer networks. Only one example for a more or less comparable object could be found: an investigation of river sediments (Stoyan *et al.*, 1997).

In artificial networks the discontinuities of properties are too extreme to be handled by geostatistical methods. Thus, these methods were not applicable.

Classical classification

Classical classification methods as well as data mining and pattern recognition (see e.g. Kiers *et al.*, 2000) were not applicable, because the reaches could not be considered as independent from each other.

Neural networks

According to Haykin (1994) a neural network (NN) is defined as a "massively parallel distributed processor that has a natural propensity for storing experiential knowledge and making it available for use. It resembles the brain in two respects:

1. Knowledge is acquired by the network through a learning process.
2. Interneuron connection strengths known as synaptic weights are used to store the knowledge."

A very comprehensive review about NNs can be found in Sarle (1997).

Neural networks need a lot of detailed data for the training, i.e. data on reach scale. The dependent parameter infiltration rate was not available on that scale. Therefore, NNs could not be applied. But, they were used for an extension of the leakage approach (see chapter 5.2.3).

Grids

The breakdown of very large investigation areas into grids is very common, e.g. for balancing river basins compounds (e.g. Biegel *et al.*, 2004). For the investigated catchments such a breakdown while keeping the information loss small was not necessary.

Options for φ

Based on pre-calculations, several options to deal with the similarity figure φ other than the implemented (see chapter 2.2.2) were tested and discarded: use of raw values, standardisation to characteristic values, sum of parameter variances and other.

5 Leakage Approach

5.1 Background

The leakage approach can be traced back to the Darcy equation. It is widely used for modelling hydrological interactions between aquifer and surface water and is related to the wetted area of the river bed, the difference between groundwater and surface water level, and a specific leakage factor representing the ability of the reach to infiltrate groundwater. As an integrative parameter the leakage factor describes various attributes of the soil layer of the river bed and thereby the potential of exchange between the compartments ground- and surface water.

According to Gustafsson (2000) and Karpf and Krebs (2004) the leakage approach can be modified for the simulation of groundwater infiltration into sewer systems:

$$Q_{Infiltrati\,on} = k_L \cdot A_S \cdot (h_G - h_S) \qquad \text{Equation 18}$$

requirement: $h_G > h_S$
where:

$Q_{Infiltration}$	infiltration of groundwater (m^3/s)
A_S	groundwater-influenced pipe surface (m^2)
h_S	water level in sewer pipes (m)
h_G	groundwater level (m)
k_L	leakage factor (s^{-1})

An extended model, which takes into consideration regional differences and house connections was proposed in APUSS (2003) for further development and implementation:

$$Q_{\inf} = (h_{GWL} - h_w) \cdot P_{wl} \cdot L \cdot K_l \cdot K_r + q_{0\,\inf} + \overline{q_{\inf\,HC}} \cdot N_{HC} \qquad \text{Equation 19}$$

requirement: $h_{GWL} > h_w$
where:

Q_{inf}	infiltration flow (m^3/d)
h_w	water level in the pipe (m)
h_{GWL}	groundwater level around the pipe (m)
P_{WI}	external wet perimeter (m)
L	length of the pipe (m)
K_l	local coefficient (d^{-1})
K_r	regional coefficient (-)
$q_{0\,inf}$	infiltration flow from other infiltration sources (m^3/d)
$q_{inf\,HC}$	mean infiltration flow for a single house connection (m^3/d)
N_{HC}	number of house connections to the pipe

The achievable spatiotemporal resolution of the approach is relatively high but depends on the data situation. It is (theoretically) possible to determine infiltration at single pipe level and at daily scale. But, only little information about processes can be expected. The quality of modelling and predicting infiltration rates is high for larger catchments (Karpf and Krebs, subm.).

5.2 Application

5.2.1 Data needs

The data needs for the application are listed in Table 10.

Table 10:data needs

Type	Data	Source
structural data	L	The structural data are contained in the sewer data base. Profile shape, diameter and slope are necessary to calculate P_{WI}.
	N_{HC}	
	profile shape	
	diameter	
	slope	
water levels	h_w	The water level in the sewer pipe can be modelled. However, it is also feasible to estimate it by measurements of water levels in the system at known conditions.
	h_{GWL}	The estimation of groundwater level at each pipe is based on the interpolation of groundwater measurements.
coefficients	K_l	The coefficients have to be calibrated (see chapter 5.2.2).
	K_r	
specific infiltration rates	$q_{0\,inf}$	The constants must be determined by separate investigations (e.g. by balancing). At least $q_{inf\,HC}$ seems to be not essential for a successful application.
	$q_{inf\,HC}$	

For the calibration of the coefficients time series of groundwater levels and infiltration rates Q_{inf} at the end of the catchment and the WWTP, respectively, must be available. To cover various groundwater conditions, a long time period and a high temporal resolution of data is necessary. Monthly values over a period of several years are recommended.

The infiltration rates can be balanced by calculating the difference of wastewater flow and the average consumption of drinking water. Due to uncertainties associated to balanced drainage rates (Karpf and Krebs, 2003) the variation of infiltration rates may be smoothened.

5.2.2 Calibration

The regional coefficient K_r is determined for a group of pipes according to field measurements. Unless external inputs from measurements are available it should be set to a default value equal to 1.

The calibration of the local coefficient K_l is based on the following equation.

$$K_{l,T} = \frac{Q_{\inf,T}}{\sum_{i=1}^{n}\left[\left(h_{GWL,i,T} - h_{w,i,T}\right)\cdot A_{i,T}\right]} \qquad \text{Equation 20}$$

requirement: for all reaches with $h_{GWL,i,T} > h_{w,i,T}$

where:

$K_{l,T}$	integral leakage factor at time T
$Q_{inf,T}$	balanced infiltration in the catchment at time T, without Q_{infHC} and Q_{0inf}
$h_{w,i,T}$	water level in the sewer pipe i at time T
$h_{GWL,i,T}$	groundwater level at the sewer pipe i at time T
$A_{i,T}$	groundwater-influenced pipe surface of pipe i at time T

The calculated leakage factor represents an integral parameter for all groundwater-influenced pipes at a certain time T. In order to refer individual leakage factors to pipes the calculation has to be carried out for a number of time spots. Thereby, the leakage factor of each groundwater-influenced pipe can be estimated to a weighted average of all calibration cases:

$$K_{l,i} = \frac{\sum_{T_i} K_{l,T}}{n_i} \qquad \text{Equation 21}$$

requirement: equidistant time steps

where:

$K_{l,i}$	calibrated leakage factor for pipe i
$K_{l,T}$	integral leakage factor at time T
T_i	time spot when $h_{GWL,i,T} > h_{w,i,T}$
n_i	number of time spots when $h_{GWL,i,T} > h_{w,i,T}$

A simplified example for the calibration procedure is shown in Figure 14. For every time T (march and october) one leakage factor was calculated. The weighting for every pipe is done by averaging these $K_{l,T}$-values. Thus, the calibrated leakage factor of pipe 1, which is not groundwater-influenced in march, is equal to $K_{l,october}$, the factor of pipe 2 is equal to the average of $K_{l,october}$ and $K_{l,march}$.

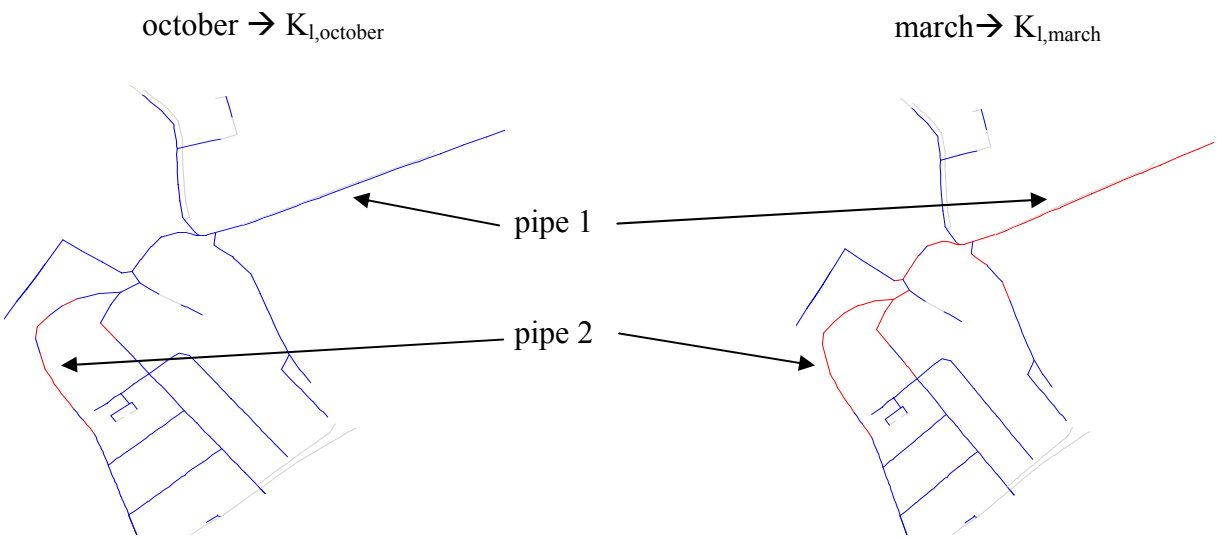

Figure 14: example for calibration

5.2.3 Extrapolation of K_l

In the case that a calibration of K_l is not possible for all reaches, the coefficient can be externally determined for the remaining reaches by using a neural network (Franz, 2004).

With a data set of the Dresden catchment containing the parameters coverage, date of construction, distance to buildings, distance to streets, distance to surface water, population density, reach surface (length and circumference were combined for simplification purposes) and street type it was possible to model the coefficient with an average error of 9 %.

6 References

APUSS (2003): Minutes of the APUSS management committee meeting held in Rome, Italy, 18-19 September 2003.

APUSS (2004): Minutes of the APUSS management committee meeting held in London, UK, 01-02 April 2004.

Armstrong M. (1998): Basic linear geostatistics. Springer: Berlin, Heidelberg. ISBN 3-540-61845-7.

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

Backhaus K. (1994): Multivariate Analysemethoden. (Multivariate Methods.) Springer: Berlin; Heidelberg. ISBN 3-540-56908-1.

Biegel M., Schanze J. and Krebs P. (2004): Improved resolution in urban water modelling for large river basins. *Proceedings of the 19th European Junior Scientist Workshop*. Lyon, 2004.

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

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.

Franz T. (2004): Reduction of Parameter Number. Internal report for the APUSS project.

Franz T. and Krebs P. (2003): Modifications of Work Package 5. Internal report for the APUSS project.

Gower J. C. (1971): A general coefficient of similarity and some of its properties. *Biometrics*, **27**, 857-872.

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. Proceedings 1st International Conference on Urban Drainage via Internet.

Hartung J. and Elpelt B. (1995): Multivariate Statistik. (Multivariate Statistics.) Oldenbourg: Muenchen, Wien. ISBN 3-486-23500-1.

Haykin S. (1994): Neural Networks: A Comprehensive Foundation, NY: Macmillan.

Karpf C. and Krebs P. (2003). Definition und Bilanzierung von Fremdwasser. (Definition and Balancing of Parasiting Water.) Umweltpraxis (10/2003), 35-38, ISSN 1616-5829.

Karpf C. and Krebs P. (2004). Sewers as drainage systems – quantification of groundwater infiltration. *Proceedings Novatech 2004, Lyon*.

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

Kiers, Rasson, Groenen and Schader (eds.) (2000): Data Analysis, Classification and Related Methods. Springer: Berlin, Heidelberg. ISBN 3-540-67521-3

Kurth W (2004): Clusterbildung und Klassifikation. (Clustering and Classification.) BTU Cottbus. http://www-gs.informatik.tu-cottbus.de/~wwwgs/bia2_v09.pdf. 28.06.04.

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

Rutsch M. (2003): Factors expected to influence the tightness of sewers. Internal report for the APUSS project.

Rutsch M. and Uibrig H. (2003): Classification System to Estimate the Leakage of Sewers. In: Baur R., Kropp I. and Herz R. (eds.): Rehabilitation Management of Urban Infrastructure Networks. *Proceedings of the 17th European Junior Scientist Workshop*. pp. 153 – 158.

Sarle W.S. (ed.) (1997): Neural Network FAQ, periodic posting to the Usenet newsgroup comp.ai.neural-nets, ftp://ftp.sas.com/pub/neural/FAQ.html. 30.12.2002.

Steinhausen D. and Langer K. (1977): Clusteranalyse. (Cluster Analysis.) de Gruyter: Berlin. ISBN 3-11-007054-5.

Stoyan D., Stoyan H. and Jansen U. (1997): Umweltstatistik. (Environmental statistics.) Teubner: Stuttgart, Leipzig. ISBN 3-8154-3526-9.

Wackernagel H. (1998): Multivariate geostatistics. Springer: Berlin, New York. ISBN 3-540-64721-X.

Walford N. (1995): Geographical Data Analysis. John Wiley & Sons Ltd: Chichester. ISBN 0-471-94162-X.