Linking distribution system water quality issues to possible causes via hydraulic pathways

W.R. Furnass[*], S.R. Mounce, J.B. Boxall

Pennine Water Group, University of Sheffield, UK

Abstract

Our limited understanding and quantification of the variety and complexity of chemical, physical and biological reactions and interactions occurring within drinking water distribution systems currently prohibit the development of a deterministic model of water quality. The causes of known water quality anomalies can however be investigated through mining the large volumes of water quality, hydraulic and asset data currently being collected by utility companies.

The data-driven methodology described here permits historical cause-effect linkages to be identified in a scalable, largely automatable fashion. Under Distribution System Integrated Modelling (DSIM), spatio-temporal searches within the set of pipes that typically lie upstream of a known water quality anomaly are used to identify possible causes. Understanding of the flow paths that connect causes and effects are derived from the results of hydraulic network simulations.

DSIM was used to investigate contacts regarding discolouration and smell/-taste issues from customers within a Water Supply Zone in England, UK, over a six-year period. 17.6% of discolouration issues and 17.4% of smell/taste issues were linked to maintenance jobs using the methodology, much smaller proportions than were identified using radial cause searches. The DSIM search results contained a greater proportion of one-to-one linkages and so are less ambiguous than the results of the radial spatio-temporal searches. DSIM was found to be a useful and informative tool for data mining multiple

[*]Corresponding author. Pennine Water Group, Department of Civil and Structural Engineering, The University of Sheffield, Sir Frederick Mappin Building, Mappin Street, Sheffield, S1 3JD, UK.

Email address: `wrfurnass1@sheffield.ac.uk` (W.R. Furnass)

water quality related datasets.

Keywords: Distribution networks, water quality, data-driven modelling, GIS, PostGIS

1. Introduction and Previous Research

The quality of drinking water in the developed world is typically very high with regards to international guidelines and national regulations. In England and Wales the periodic sampling of potable water at treatment works, treated water reservoirs and customers' taps is mandated. In 2010 99.96% of the water samples taken to satisfy the water industry regulators' monitoring requirements were in compliance with regulations (DWI, 2011). There is however a need to ensure that the health impact and the aesthetics of water supplied by water providers continue to be satisfactory.

In many developed countries drinking water distribution system (DWDS) infrastructures are ageing and processes such as corrosion continue to impact on water quality in networks that contain older metal pipework (Kirmeyer, 2002, p. 89). As a result, regulators and water providers are now developing and implementing capital and operational measures for managing water quality in DWDS (DWI, 2002), in addition to the sound catchment management and treatment practices already in place.

The designs of these measures are in part informed by evidence of events and activities within DWDS that have previously had a detrimental effect on the quality of water at customers' taps. In addition, theoretical threats to water quality from intra-network causes are being explored through research. For example:

- 15% of the cases of diarrhoea in a UK study by Hunter et al. (2005) were thought to be associated with bursts and pressure losses within DWDS.
- Payment et al. (1997) considered the possible causes of pathogen ingress to include cross-connections, pipe replacements and the manipulation of valves and hydrants. Besner et al. (2007) then related variations in water quality to such maintenance activities.
- Research is being conducted into how transient pressure waves (resulting from valve closures, for example) can cause contaminant ingress

via back-siphonage (LeChevallier et al., 2003; Boyd et al., 2004; Collins et al., 2011).

- From one study it was calculated that on average 0.23 low pressure events, each of which had the potential to cause contaminant ingress, occurred per year per 1000 population served (WRc, 2008).

It should be noted that a) there is typically much uncertainty associated with the results of epidemiological studies such as Hunter et al. (2005), b) that intra-network water quality issues can have a variety of anthropogenic causes (e.g. power failures causing pump trips; maintenance activities conducted by water companies; structural failures, possibly due to a lack of maintenance/renewal; deliberate contamination) and c) several sets of industry guidelines have been developed in which methods are presented for greatly reducing the risk of operations and maintenance activities negatively impacting on water quality (e.g. Ainsworth and Holt (2004)).

Water providers are not only concerned with the chemical and microbiological quality of supplied water but its aesthetics too. In one study 41% of the complaints made by customers to English and Welsh water providers were regarding drinking water aesthetics, 37% of which pertained to discolouration (Vreeburg and Boxall, 2007).

A considerable proportion of discolouration incidents can be associated with 'known activities' within DWDSs. In the period 2006-2008 the Drinking Water Inspectorate for England and Wales attributed 47% of issues to planned works, 6% to pump failures, 16% to valve failures/replacements, 9% to mains damage, 10% to connections, 6% to reservoir issues and 6% to treatment works issues (Husband and Boxall, 2011).

A widely-accepted theory is that discolouration is due to cohesive layers of particulate matter being stripped from pipe walls due to a change in pipe wall shear stress (Husband et al., 2008). The manipultation of values during maintenance operations therefore has the potential to cause discolouration.

While combined compliance rates in the England and Wales are very high, unacceptable chemical, aesthetic and biological water quality failures do occur, a proportion of which are thought to occur within DWDSs. Given this, and the desire to operate DWDSs in a more proactive manner, there is a need to better understand water quality within DWDSs, particularly with respect to the number of water quality anomalies that may be due to known or planned activities.

Analytical modelling of the activities and events that can affect water quality in DWDSs is impeded by the latter's structural heterogeneity and by the complexity of the physical, chemical and microbiological processes involved. Data-driven investigations into the causes of water quality fluctuations can potentially elucidate causal relationships without the need for computationally-intensive analysis.

Jaeger et al. (2002) suggested that data regarding water quality incidents and possible causal events/activities could be mined for cause-effect relationships. They stated that such data mining needs to take into account the hydraulics of DWDS as flow paths determine if and how cause and effect are linked. However, the methodology proposed by Jaeger et al. (2002) for determining the causes of water quality issues quantifies the separation of cause and effect in Euclidean (planar) space. This is thought to be inappropriate for determining if and how an event at one location in a labyrinthine DWDS has influenced water quality at another as hydraulic connectivity and network path distances are not considered. Okabe et al. (2006) argued that planar spatial analysis techniques are inappropriate for studying network-constrained data (such as water quality incidents or mains repair data) as they can produce many false-positive results. Even so, planar/Euclidean searches continue to be the common method for determining the causes of water quality issues because of their simplicity and ease of implementation. Others have used GIS systems and either visual assessment of data (e.g. Eng et al. (1999)) or planar spatial searches (e.g. Kistemann et al. (2001)) to explore the causes of DWDS water quality issues.

The methodology of Jaeger et al. (2002) was developed further in its incorporation into the *IMADSIG* software (Trepanier et al., 2006; Besner et al., 2007). With this tool, flows can be traced back from the location of a water quality incident to upstream boundaries such as treatment works. This software allows the set of pipes identified through such 'back-tracing' to be visualised along with the results of spatio-temporal queries of the various data model datasets. Human judgement is then applied to associate 'qualitative probabilities' with cause-effect linkages.

The results of analysing historical data from DWDSs in seven cities were mixed (Besner et al., 2007): no causes could be identified for 0-38% of coliform-positive samples, for 36-53% of the customer complaints and for 8-83% of the heterotrophic plate counts. Besner et al. considered the efficacy of the methodology to be limited by the number of unrecorded events and activities, to inaccurate spatio-temporal references, to the low sampling

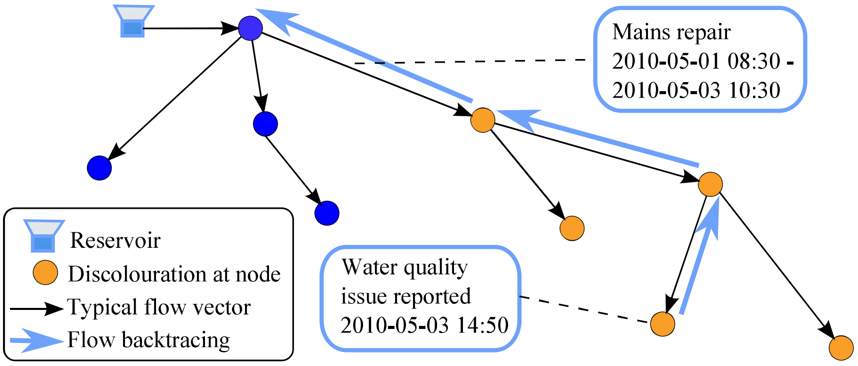

Figure 1: Searching within an integrated network model for upstream causes of a historical water quality issue.

frequency and to hydraulic models not being sufficiently representative.

2. Aims and objectives

A scalable, largely automatable methodology called Distribution System Integrated Modelling (DSIM) has been developed for linking historical water quality issues to known events by network topologies and hydraulics. The objectives of the research were to:

1. Develop a method for combining water quality, asset, topological and hydraulic data to form an integrated network data model, giving consideration to the quantity, quality and value of the datasets used.
2. Devise a method for determining the *region of influence* hydraulically upstream of a given water quality issue. This is the set of pipes that lies upstream of a particular network location, given typical directions of flow. Possible causes of that issue can be found by spatio-temporally searching the issue's region of influence.
3. Apply the developed methodology to datasets that describe a particular distribution system to assess its efficacy.

The application of the methodology for identifying a probable cause for a reported discolouration complaint is illustrated in figure 1.

3. Developing and analysing an integrated data model

The developed methodology is presented in this section and section 4 features a case study that demonstrates its application.

3.1. Dataset selection: water quality anomalies

To gain an understanding of the water quality events that occurred over a historical period one can refer to regulatory sampling and customer contact databases. In the future it may also be possible to study quality changes in DWDSs using on-line instrumentation that continuously samples multiple determinands. Such technology exists but it is currently unproven and has not yet been widely adopted (Boxall et al., 2010).

Regulatory sampling programmes in England and Wales provide valuable insight into the microbiological and chemical quality at treatment works, service reservoirs and customers' taps. However, these "*discrete, manually-collected samples produce a small amount of 'spot' data that is representative of the quality of the water at a single time and duration at a given location. These snapshots [...] provide no understanding of the system status prior to or after the data was recorded.*" (Boxall et al., 2010).

By contrast, customers themselves provide excellent network coverage and potentially continuous feedback on certain water quality parameters. Customer contacts are generally a utility company's first indication of low pressure or discolouration issues. However, customer contacts describe only highly subjective perceptions of water quality (Whelton et al., 2007) and may have inaccurate temporal references due to 'recognition' and 'reaction' delays. Customer behaviour, standards and expectations are also highly subjective; contacts are however often a good indicator of a change in water quality.

Only certain types of customer contacts are considered to be of interest as inputs to DSIM. Discolouration, measured as turbidity, is an important aesthetic parameter but can also be an indicator of the risk of the ingestion of dislodged biofilm material: Gauthier et al. (1999) found that $\leq 11\%$ of the particulate matter in a particular DWDS was organic carbon. Developing a better understanding of the relationship between discolouration incidents and activities such as mains repairs and valve operations could help inform maintenance strategies.

The odour and taste of supplied water can be affected by high residual disinfectant concentrations, precursor materials that have broken through treatment works interacting with disinfectants, trihalomethane concentrations and biological activity and decay, amongst other causes (Khiari et al., 2002). The ability to relate contacts regarding water that smells/tastes strongly of disinfectant to specific maintenance/rehabilitation activities could be of value to local asset management.

Contacts regarding no flow, low flows or low pressures are also of interest. Such observations can result from a burst, from pumping station issues or from system maintenance or operational activities including repairs. All these events/activities have the potential to impact on water quality through increasing flows and therefore shear stresses (see Boxall et al., 2001; Boxall and Saul, 2005; Boxall and Prince, 2006; Husband et al., 2008) or by inducing back-siphonage. However, contacts regarding water aesthetics and/or chemistry are of greater importance than contacts regarding hydraulic anomalies as they more directly relate to the issues of interest.

Contacts categorised as being related to illness are recorded in England and Wales. However, there is too much uncertainty associated with complaints regarding illness and the number of such contacts are too low to allow cause identification to be automated; the type, severity and onset delay can vary greatly between incidents and it can be difficult to prove that the consumption of water was the cause of an incident.

3.2. Dataset selection: the potential causes of water quality issues

Gray (2008) categorised the causes of drinking water quality issues as being the result of a) treatment works issues, b) raw water quality fluctuations and/or c) distribution network events/activities. Household plumbing/activities constitute a further category of possible causes. The effects of a) and b) above are minimised through the development and implementation of Water Safety Plans (WHO, 2005), diligent treatment practices and monitoring and feedback control. When drinking water quality issues do occur, cause and effect can be more easily correlated for issues resulting from a) or b) than c), due to the populations affected and the causes not lying within subterranean, pressurised distribution systems. This paper therefore focuses on the effects of c). The US Environment Protection Agency have stated that the "*age and complexity of distribution systems...has increased the likelihood for contamination events and water-borne disease not related to source water treatment deficiencies*" (USEPA, 2006). This highlights the importance of considering how the quality of treated water can be influenced by events/activities within distribution systems.

Under DSIM the intra-network causes of water quality issues can be described in terms of flow or pressure anomaly datasets (which can be indicative of the occurrence of bursts, exceptional demands and operational/maintenance activities) or maintenance job records. The latter are considered to be of greater value for DSIM due to the volume and resolution of available data and to the nature of the activities being known.

3.3. Data quality considerations

Before attempting to correlate datasets describing water quality issues and their possible causes one must first give consideration to their accuracy and whether they are sufficiently representative of reality. Means for ensuring satisfactory data quality are presented in this section.

Water quality issue records and possible cause records that do not have valid Geographical Information System (GIS) georeferences should be discarded. All records should be reprojected to a common spatial reference system. The customer relationship management systems and maintenance job management systems used by water providers typically feature georeferenced data, thus georeferences can be extracted directly from such databases. Polygonal GIS data denoting the area of interest (such as a Water Supply Zone (WSZ)[1] or DMA) can be used to cull irrelevant event records.

All event records require a commencement timestamp (date and time) and maintenance job records need both start-on-site and finish-on-site timestamps, with the latter occurring after the former and the separation of the two being plausible for the type of job. Customer contact timestamps will inevitably be associated with both recognition and reaction delays, particularly for the case of mailed letters.

The variable propagation delay associated with letters was thought to be a potential source of error in contact timestamps so such contacts are not ideally suited to automated cause identification via DSIM. However, only a very small proportion of the contacts received by water providers from their customers are letters: in England and Wales the ratio of written contacts to telephone contacts is approx. 1:225 (OFWAT, 2010). This suggests that the

[1] Distribution Metered Areas (DMAs) are semi-isolated DWDS regions that are used for leakage calculations and other management purposes, each of which has a small number of inlets and outlets; WSZs are comprised of multiple DMAs and have a population of no more than 100,000 (Tanyimboh and Key, 2010)

analysis of large volumes of water quality complaints for possible causes may not be overly affected by including written complaints.

The clustering of customer contacts is necessary to ensure that contact data are representative of water quality issues rather than individual phone calls/emails. A simple method is presented here for clustering contact records by discretising their timestamps.

In figure 3.3, each (georeferenced) customer contact is assigned an integer-valued *cell* number. This number is calculated by using ‘integer division’ to divide the magnitude of the timestamp (in seconds since a datum) by an empirical *cell size* parameter (also in seconds). The cell size parameter dictates the *maximum possible temporal separation* of two contacts in one cluster. The contact records are then grouped by cell number, georeference and complaint type (e.g. ‘discoloured water’). Each clustered contact grouping is assigned the earliest timestamp (a representative value) of its precursory records. The resulting dataset is henceforth referred to as *clustered contacts*. Sensitivity analysis is required to determine an appropriate duration in seconds for the *cell size* parameter.

Customer contacts should only be clustered in space (as well as time) if it can be assured that both the clustering algorithm and the method used to link water quality issues to possible causes use the same measure of proximity (Euclidean or network distance).

3.4. Implementation

To implement DSIM on large datasets a storage, processing and analysis system is required that is flexible, scalable and efficient. The PostgreSQL relational database management system (RDBMS) along with the PostGIS geo-spatial extensions (Obe and Hsu, 2010) were selected for this purpose as: file-based storage is much more costly to search than indexed database tables; data-type integrity is ensured by relational databases and PostgreSQL/-PostGIS provides an array of spatio-temporal analysis functions from which arbitrarily-complex Structured Query Language (SQL) statements can be constructed.

The presented methodology was initially developed through manipulating PostGIS databases using SQL statements and custom functions, with Python scripts being used for parameter sensitivity analysis. The functionality was then encapsulated in an Application Programmers’ Interface (API) written in Python.

```
min_timestamp = min(contacts_dataset.all_timestamps)

foreach c in contacts_dataset:
   c.cell = floor((c.timestamp - min_timestamp) / cell_size)

foreach group in contacts_dataset.group_by(cell,
                                          georeference,
                                          complaint_type):
   i = new_quality_issue(min(group.timestamps),
                         georeference,
                         complaint_type)
   clustered_customer_issues.append(i)
```

Figure 2: Customer contacts clustering algorithm. Sensitivity analysis and/or engineering judgement must be used to determine a value (in seconds) for the empirical parameter `cell_size`.

A flowchart that illustrates the various components of the developed methodology is shown in figure 3. The purpose of these components, along with further basic implementation details, are provided in this section (§3).

Datasets describing water quality issues and their potential causes are imported into PostGIS from sources for which there are ODBC drivers available (e.g. MDB, XLS). The OGR (Sherman, 2008) tools are used to import polygonal GIS data that define the area of interest and internal partitions (e.g. DMA boundaries), the former being used for data filtering and the latter for query scoping.

Data stored within a PostGIS database can be visualised using a graphical GIS application such as Quantum GIS (Sherman, 2008).

3.5. DSIM: Linking issues to possible causes via hydraulic pathways

The water quality issue and possible cause datasets used as inputs to DSIM are network-constrained. It is possible to associate issues with possible causes using pathways derived from corporate GIS asset data but these linkages do not predicate hydraulic connectivity.

Under DSIM a water quality issue is associated with possible causes through searching the Region of Influence (RoI) upstream of that issue. It was thought that searching upstream for the causes of a set of effects would

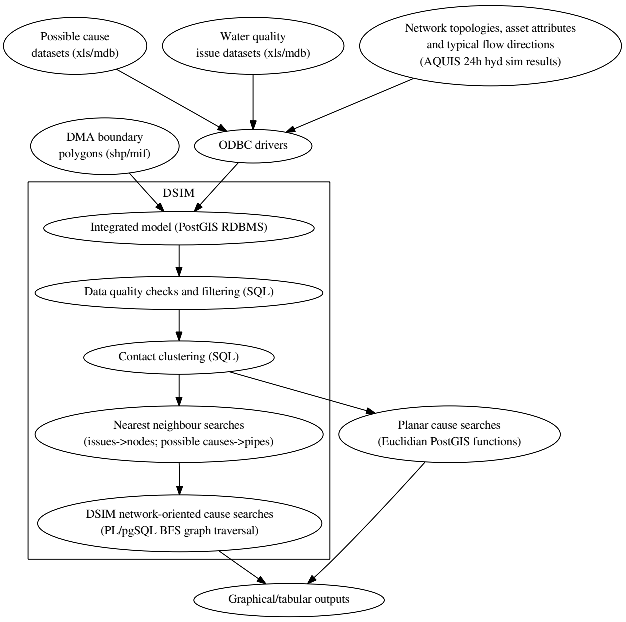
Possible cause datasets (xls/mdb)
Water quality issue datasets (xls/mdb)
Network topologies, asset attributes and typical flow directions (AQUIS 24h hyd sim results)
DMA boundary polygons (shp/mif)
ODBC drivers
DSIM
Integrated model (PostGIS RDBMS)
Data quality checks and filtering (SQL)
Contact clustering (SQL)
Nearest neighbour searches (issues->nodes; possible causes->pipes)
Planar cause searches (Euclidian PostGIS functions)
DSIM network-oriented cause searches (PL/pgSQL BFS graph traversal)
Graphical/tabular outputs

Figure 3: DSIM overview.

lead to a less ambiguous understanding of causal relationships than the inverse (see figure 4).

Historic flow data captured via SCADA systems is of insufficient spatial resolution to determine the pipe flow vectors required for constructing a RoI. Instead, typical flow directions can be determined per pipe from the results of an extended time-series hydraulic network simulation (Walski et al., 2003).

An idealised extended time series simulation (24h) is conducted using the AQUIS modelling software (7-Technologies, 2009) and a hydraulic model of the DWDSs of interest. Node, pipe and valve attributes, geometries and state information are then migrated from AQUIS to PostGIS via ODBC along with pipe bulk velocity vectors per 15min simulation time-step. The mean daily velocity vector is then calculated per pipe to give a typical direction of flow.

Note that hydraulic modelling packages other than AQUIS (e.g. EPANET (USEPA, 2008), SynerGEE[2] or Infoworks[3]) could be used in place

[2] http://www.gl-group.com/en/water/SynerGEEWater.php

[3] http://www.innovyze.com/products/infoworks_ws

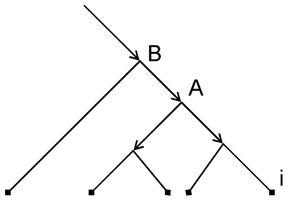

Figure 4: Causality between event A and downstream event i can be less ambiguously determined if one searches for the possible causes of i (those being A and B) rather than the possible effects of A.

of AQUIS to solve the DWDS of interest for flows and pressures as long as the data types listed previously could be migrated to PostGIS from the results of a 24h network simulation.

Before RoI can be constructed it is necessary to first associate water quality issues and possible causes with network model elements. In hydraulic DWDS models, customer demands are allocated at nodes. Clustered customer contacts and regulatory tap sampling records are therefore also assigned to nodes in the presented method. Possible cause datasets such as maintenance job records are associated with pipes. These associations are achieved through storing the unique ID of the nearest node with each clustered customer contact and regulatory tap sampling record and storing the ID of the nearest pipe with each maintenance job record.

The *nearest neighbour* searches used to make such associations take the form of a SQL update statement that features PostGIS' radial (Euclidean) separation test function (`St_DWithin`), a maximum separation and SQL `DISTINCT ON` and `LEFT JOIN` constructs. The statement caches the unique ID of the closest network model entity for each issue or possible effect record. GIS 'linestrings' can be produced to allow the nearest neighbour search results to be visualised in Quantum GIS.

The determination of the extent of a RoI can be interpreted as a graph theory problem (Cormen, 2005). The network topology and time-averaged velocity vectors contained within PostGIS together form a directed, cyclic graph that can be methodically explored to identify the set of pipes that typically lie upstream of a node. To this end a breadth-first upstream traversal algorithm was implemented for efficiency as a non-recursive (*double-ended queue*-based) stored procedure in PostgreSQL's own PL/pgSQL language. A list of visited nodes were maintained to avoid infinite cycling around network loops. Valves $< 1\%$ open are considered hydraulically impassible. A Quantum GIS plug-in was developed for visualising the RoI upstream of a clustered customer contact (figure 5).

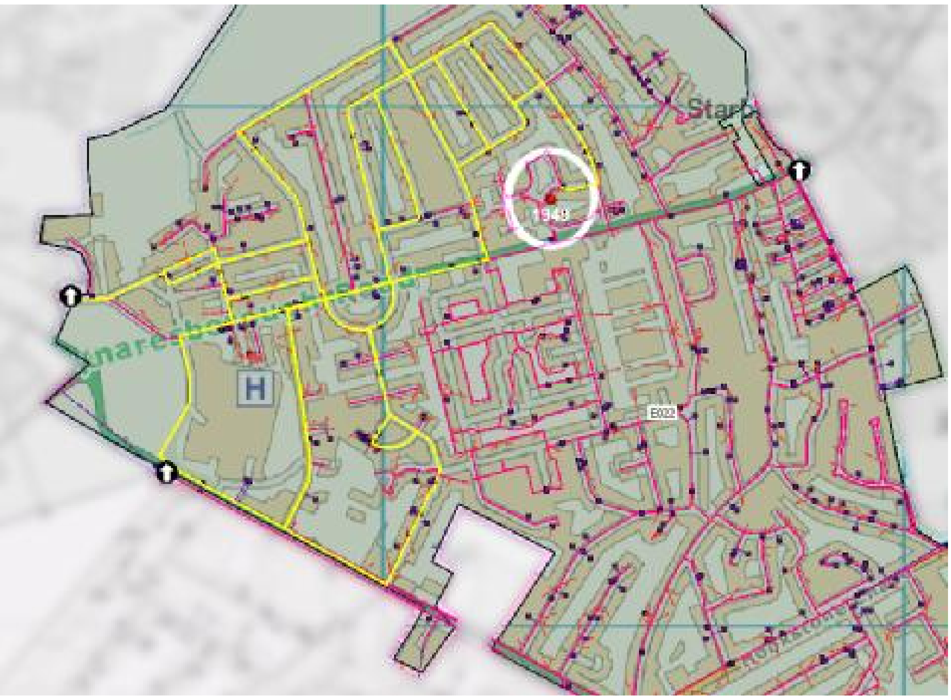

Figure 5: Region of Influence upstream of a specified (circled) customer contact (Quantum GIS plug-in). The upstream pipes are depicted using a lighter colour. The boundaries to upstream traversal are the two DMA inlets, shown on the left as arrows inside black circles (the arrow on the right is an export to another DMA).

The algorithm pseudocode in figure 6 shows how DSIM uses RoI to determine possible causes for water quality issues once nearest neighbour searches have been performed.

3.6. Planar analysis of the integrated network model

The efficacy of DSIM can be contrasted to a network-agnostic, more GIS-like approach to causal linkage identification through applying both DSIM and the algorithm shown in figure 7 to a particular developed integrated data model. r_{win} and t_{win} in figure 7 are correlation factors that account for

Figure 6: DSIM causal linkage identification. t_{win} is a correlation factor that accounts for propagation delay plus customer recognition/reaction delay.

```
foreach i in water_quality_issues_dataset:
   foreach c in possible_cause_dataset:
      if c.nearest_pipe in RoI(i.nearest_node)
      and i.start_timestamp > c.start_timestamp
      and i.end_timestamp < c.end_timestamp + t_win then:
         causal_linkages.append([c,i])
```

Figure 7: Planar linkage identification. r_{win} and t_{win} are correlation factors that account for spatial advection and propagation delay plus customer recognition/reaction delay respectively.

```
foreach i in water_quality_issues_dataset:
   foreach c in possible_cause_dataset:
      if planar_separation(i,c) < r_win
      and i.start_timestamp > c.start_timestamp
      and i.end_timestamp < c.end_timestamp + t_win then:
         causal_linkages.append([c,i])
```

spatial advection and propagation delay plus customer recognition/reaction delay, respectively.

The number and types of causal linkages identified by the planar method are in part dependent on the values of t_{win} and r_{win}; sensitivity analysis should therefore be conducted for m distinct values of r_{win} between $0m$ and the length of the longest lateral axis of the analysed region and for n distinct values of t_{win} between $0h$ and the maximum propagation delay in the studied distribution network (the upper bound of the temporal separation of cause and effect). Through varying t_{win} whilst using a representative value of r_{win} and vice-versa, the number of distinct applications of the planar method required for bi-variate sensitivity analysis is reduced from $n \times m$ to $n+m-1$.

The mean DMA radius serves as r_{win_repr}, a representative value of r_{win}, as the majority of cause-effect linkages pertaining to water quality are thought to be confined by DMA boundaries due to DMAs typically having few inlets and outlets. A representative value of t_{win}, t_{win_repr}, is defined as the typical propagation delay to all points in a network and is calculated from the result of water quality simulations as the mean of the maximum age at each node (ignoring dead ends) once steady-state has been reached (Machell et al., 2009) plus a term to account for typical customer reaction delay if analysing customer contact records. t_{win_repr} can be used to compare the efficacy of the DSIM and planar methodologies.

4. Case study

The efficacy of DSIM was tested through its application to datasets pertaining to a WSZ in northern England. The $101km^2$ study region, depicted in figure 8, covers both urban and rural areas and features 14 DMAs and $368km$ of mains that supply approximately $19,000$ properties.

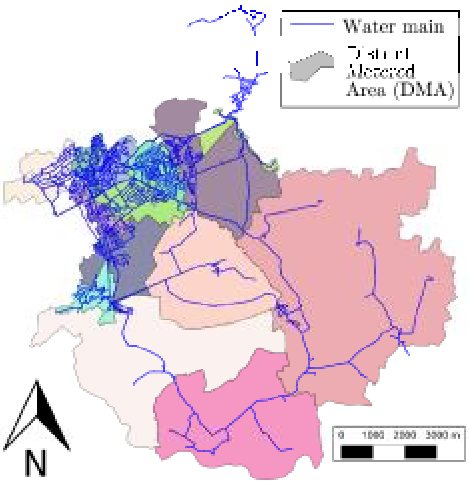

Figure 8: Case study area showing pipes and DMA boundaries.

A substantial dataset pertaining to water quality issues in the study area (customer contact records; regulatory samples) and the potential causes of those issues (maintenance job records) from October 2001 to July 2007 were collated within a PostGIS database. Only seven regulatory sample failures were identified for the study period so this data was not analysed further.

Written contacts could not be distinguished from telephone calls/emails so were not discarded. However, this was not thought to be particularly significant for ‘bulk’ cause searches given the low ratio of calls to letters received by the utility company in 2009-2010 (see §3.3).

Customer contact records regarding the quality (and quantity, for reference purposes) of received water were clustered in time, as per algorithm 3.3. Customer contacts were not clustered in space in this study.

Sensitivity analysis showed the relationship between the number of distinct customer issues and the algorithm cell size being linear between 0h and 12h and logarithmic above the latter value (figure 9); the customer issue dataset used in subsequent analysis (see table 1) was therefore produced using a 12h cell size.

Each contact/cluster had been associated with a property ‘seed point’ within the utility company’s Customer Relationship Management (CRM) database. The geocoding of the customer issues was visually verified by overlaying land use data with customer contacts using the Quantum GIS software (Sherman, 2008).

50% of maintenance job records were lacking timestamps for both the arrival on, and departure from, the incident site or had departure timestamps that preceded the arrival timestamp; these records were therfore purged.

Questions regarding the fidelity of GIS data and of the company’s all-mains hydraulic model of the WSZ were raised when the model and customer contacts were graphically superimposed over polygonal DMA boundary data: certain mains appear to have been omitted from the model and the model does not reflect rezoning over the period of analysis. The latter is unavoidable given that both GIS systems and hydraulic models typically only provide a snapshot view of DMA boundaries and the state of hydraulic devices.

Dataset	Class	Records
Clustered customer contacts	Discolouration	1413
	Smell/taste	51
	Low flow/pressure; no water	765
Maintenance job records	Service pipe	303
	Repair main	276
	Main fitting	255
	Supply pipe	198
	Stop Tap	145
	Other	288

Table 1: Case study water quality issue and possible cause datasets grouped by classification

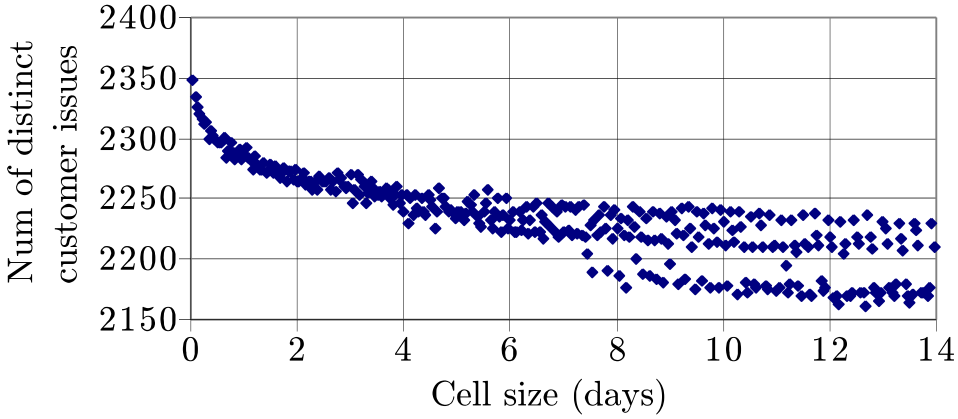

Figure 9: Sensitivity of customer contact dataset size to clustering algorithm cell size.

5. Case study results

Nearest neighbour searches were conducted to associate clustered contacts with nodes and maintenance jobs with pipes. It was anticipated that the results of applying algorithm 6 to the integrated WSZ model would be highly dependent on the value of t_{win}. The sensitivity to t_{win} of the number of contacts for which possible causes could be found is shown in figure 10 and table 2. Here the maximum value of this parameter, 480h, corresponds to the maximum propagation time calculated for the WSZ and is an upper bound to the temporal separation of cause and effect. Results are also discussed in terms of a representative value for t_{win} of 40h, shown in figure 10. This representative t_{win} is comprised of a conservative measure of water age (the mean of the maximum water age at each node as calculated using AQUIS; nodes at ‘dead ends’ with ever-increasing water age were ignored) plus 10h to allow for customer reaction delay.

Using DSIM with a t_{win} of 40h, possible causes were identified for 17.6% of discolouration issues and 17.4% of smell/taste issues, with 82.9% of the explainable discolouration incidents being associated with mains repair, reconditioning and renewal. Only 11.5% of explainable discolouration issues and 11.4% of smell/taste issues were associated with ≥ 1 possible cause. As mentioned in §4, 50% of maintenance records were discarded for not meeting the data quality criteria; one or more water quality issues could be associated with 125 of the 1465 remaining records ($t_{win} = 40h$).

The planar search method was also applied to the case study model. With $r_{win} = 1600m$ (mean DMA radius) and $t_{win} = 40h$ possible causes were identified for 74.9% of all discolouration issues and 86.3% of all smell/taste issues (figure 11). However, these causal linkages were much more ambiguous than those found using DSIM: 53.4% of the explainable discolouration issues and 42.4% of the explainable smell/taste issues had ≥ 1 possible cause. If the planar cause searches are restricted by DMA boundaries then 49.6% of discolouration issues and 51.0% of all smell/taste issues can be associated with ≥ 0 causes and 31.0% of discolouration issues and 15.3% of all smell/taste issues can be associated with ≥ 1 causes (figure 12).

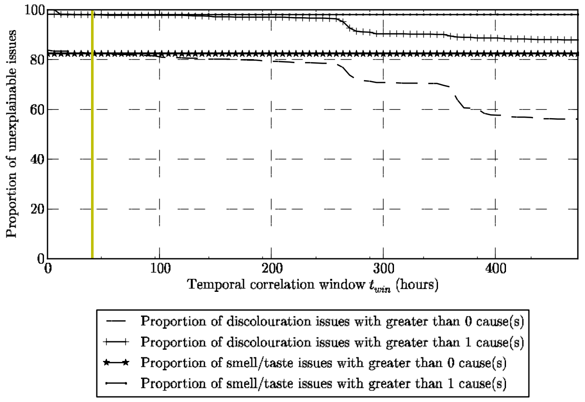

Figure 10: Customer contacts for which causal linkages not identified by DSIM.

6. Discussion

A much smaller proportion of the water quality issues investigated through the study could be attributed to maintenance jobs when using the network-orientated DSIM searches than the Euclidean searches. One might therefore conclude that, using DSIM, fewer water quality issues can be said to be due to known activities than if cause-effect relationships were explored through conducting planar proximity searches using a typical GIS application.

The planar method returned more ambiguous results: a greater proportion of the explainable water quality issues were associated with more than

	Cause category	Cause subcategory	Issues per issue category	$t_{win} = 0h$	$t_{win} = 40h$	$t_{win} = 480h$
Discolouration issues	Repair Main	Repair Main using dowel piece.	1416	124	125	126
	Repair Main	Repair Main other methods (not Dowel Piece)	1416	53	56	250
	Service Pipe	Renew Existing Service Pipe	1416	21	23	27
	Main Fitting	Excavate to Boundary Stop Tap & Carry out Flow Testing	1416	12	16	70
	Main Fitting	Repack or Refurbish Sluice Valve	1416	16	16	16
	Service Pipe	Repair Existing Service Pipe	1416	12	12	33
	Other	Install Stop Tap to Existing Service Pipe	1416	5	10	35
	Stop Tap	Renew Stop Tap	1416	7	7	210
	Stop Tap	Repair Leak in Chamber or Existing Excavation	1416	1	2	15
	Other	Lead Renewal - Shortside	1416	0	2	15
	Supply pipe	Section 75 Repair leaking supply pipe	1416	0	2	6
	Other	Excavate and Lay Main	1416	2	2	3
	Main Fitting	Remove Ferrule & Stop Tap or Meter	1416	1	1	4
	Other	Install Bulk Meter (in line)	1416	0	1	2
	Other	Install Non-standard Meter	1416	0	1	2
	Supply pipe	Repair leaking supply pipe	1416	0	0	23
	Service Pipe	Leak on or near meter	1416	0	0	12
	Other	Reinstate Consequential Damage	1416	0	0	3
	Supply pipe	Rechargeable Section 75 Repair leaking supply pipe	1416	0	0	1
	Main Fitting	Repair/refurbish Sluice Valve.	1416	0	0	1
Smell/taste issues	Other	Install Stop Tap to Existing Service Pipe.	51	6	6	6
	Repair Main	Repair Main other methods (not Dowel Piece)	51	3	3	4
	Service Pipe	Renew Existing Service Pipe.	51	1	2	2
	Main Fitting	Excavate to Boundary Stop Tap & Carry out Flow Testing	51	1	1	1
	Repair Main	Repair Main using dowel piece.	51	1	1	1
	Other	Lead Renewal - Shortside	51	0	0	1

Table 2: Quantification of types of causal linkage identified by DSIM

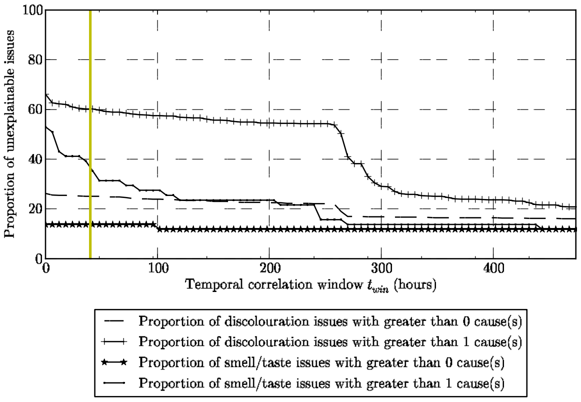

Figure 11: Customer contacts for which causal linkages not identified by planar method (r_{win}=1600m).

one cause. This was anticipated as network-orientated methods can potentially identify fewer false positives than planar methods when analysing network-constrained data (Okabe et al., 2006).

Under DSIM the number of explainable discolouration issues increased by 7% when t_{win} was increased from 225h to 275h (figure 10). This is due to 158 discolouration issues occurring on one particular day in one DMA but probable causes (two mains repairs in that DMA) only being found for those issues if one searches approximately ten days back in time. The mean of the maximum water age at each node is just 30h and the step-increases in explainable issues are also seen when conducting sensitivity analysis on the t_{win} parameter of the planar method. It is therefore thought that the anomaly is due to the true cause(s) corresponding to events/activities not featuring in the analysed possible cause dataset. The efficacy of the DSIM and planar methodologies could therefore be improved if a) a jobs dataset with fewer missing timestamp pairs had been used and b) other possible cause datasets were incorporated into the model such as flow anomalies identified using machine learning or statistical techniques (e.g. Mounce et al. (2010)) and/or anomalous samples from treatment works outflows. The latter would allow bursts to be considered as possible causes.

Under DSIM, a RoI was defined as the set of pipes that lie upstream of a quality issue, i, given that time-averaged velocity vectors from an idealised 24h hydraulic DWDS simulation. RoI could alternatively have been defined

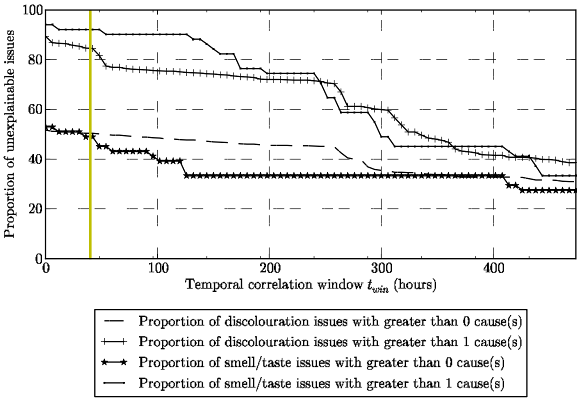

Figure 12: Customer contacts for which causal linkages not identified by planar method (r_{win}=1600m; cause and effect in same DMA).

for i, which occurred at time t_0, as follows: given the results of an idealised 24h simulation the pipes that lie upstream of i and the corresponding propagation delays relative to t_0 could recursively be identified. An event/activity, e, could then be said to be a possible cause of i if e is associated with a pipe p in the set $RoI(i)$ and occurs at a time between the minimum propagation delay (t_{prop}) between p and i and $t_{prop} + t_{win}$, where t_{win} is a time window to account for error. Causal linkages identified using this methodology would be more accurate were accurate data and models available. Also, such a methodology would better recognise changes in the direction of flow along a pipe due to times of high or low customer demand. However, it was decided that the overheads of such an approach could not be justified given a) the perceived data and hydraulic model quality issues, b) the low numbers of flow reversals identified during a 24h simulation by AQUIS and c) uncertainty and reaction delays associated with customer contact timestamps.

The assumption made in this investigation that causes lie upstream of effects may not always be true in reality: a burst along one branch of a fork may cause an increase in shear stress upstream of the junction and mobilised discolouration material may then reach customers supplied via the other branch. In May 2009 28,000 customers in the vicinity of Ashington, UK, were affected when the flow at the downstream extremity of a trunk system was elevated to increase levels in a service reservoir (DWI, 2010). Although the flow velocity increase at the upstream end was small, the resulting change in wall shear stress was sufficient to exceed its 'conditioning levels' (Husband et al., 2010), leading to mobilisation of discolouration material. This material was then advected into DMAs supplied from a take-off mid-way along the trunk and led to a significant number of customer complaints regarding discoloured water. The aforementioned assumption could therefore be relaxed in future work.

The method used to cluster customer contacts could be improved were it to cluster in space as well as time to provide a street-level rather than property-level view of water quality issues. This would require that contacts be grouped by both network distance (Okabe et al., 2006) and timestamp.

The hydraulic models used with DSIM provide good spatial resolution of typical flow vectors but do not reflect temporary network reconfigurations (during burst repairs etc.), nor more permanent network changes such as rezoning. DMA-level demand scaling or 'on-line' modelling strategies/software (e.g. Machell et al. (2010); Hatchett et al. (2011)) could provide more accurate flow vectors for a point in time. However, both are limited by valve state change data not being readily available: when addressing structural failures the primary objective of network field teams is to ensure continuity of supply so temporary and permanent changes in asset configurations are not always recorded. Also, GIS and hydraulic modelling software packages do not typically allow for the storing of long-term time series state changes for an asset, thus requiring that utility companies define a further data store should they wish to capture such data. Were water companies to record valve state changes more accurately in real time then DSIM could be integrated with an on-line DWDS model; machine learning techniques could then be used to identify water quality anomalies and possible causes in real time and could 'learn' the typical impact of certain events/activities on water quality.

7. Conclusion

A data-driven, highly-scalable, and largely automatable method known as Distribution System Integrated Modelling (DSIM) has been developed for linking historical water quality issues in drinking water distribution networks to possible causes via hydraulic pathways. This methodology was implemented using software and modelling components including: AQUIS, PostgreSQL/PostGIS (including PL/pgSQL queries and Quantum GIS visualisation) and Python scripts.

When applied to customer contact data and maintenance job records pertaining to a six year period for a Water Supply Zone (WSZ), far fewer causal linkages were identified than were found using a Euclidean radial search method. A greater proportion of the linkages identified by DSIM featured one-to-one relationships and so those results were deemed less ambiguous. This suggests that using planar proximity searches for such analysis within a typical GIS application could lead to an overestimate of the number of customer complaints regarding discolouration and odour/taste issues that are due to known or planned activities.

DSIM provides useful insight into trends relating to water quality issues and their causes.

Acknowledgements

The authors would like to thank Yorkshire Water Services Ltd for supplying the hydraulic models, customer contact data and maintenance job records analysed in this research.

This research was supported by Engineering and Physical Sciences Research Council grant EP/G029946/1.

References

7-Technologies, 2009. AQUIS software v1.47.
URL http://www.7t.dk/products/aquis/

Ainsworth, R., Holt, D., 2004. Safe Piped Water: Managing Microbial Water Quality in Piped Distribution Systems. World Health Organization.

Besner, M., Gauthier, V., Trepanier, M., Martel, K., Prevost, M., 2007. Assessing the effect of distribution system O&M on water quality. J. American Water Works Assoc. 99 (11), 77–91.

Boxall, J., Dewis, N., Machell, J., Gedman, K., Saul, A., 2010. Operation, maintenance and performance. In: Savic, D., Banyard, J. (Eds.), Water Distribution Systems. ICE Publishing, Ch. 8.

Boxall, J., Saul, A., 2005. Modelling discoloration in potable water distribution systems. J. Environ. Eng. 131, 716.

Boxall, J., Skipworth, P., Saul, A., September 2001. A novel approach to modelling sediment movement in distribution mains based on particle characteristics. In: Proc. International CCWI conference.

Boxall, J. B., Prince, R. A., 2006. Modelling discolouration in a Melbourne (Australia) potable water distribution system. J. Water SRT - Aqua 55 (3), 207–219.

Boyd, G., Wang, H., Britton, M., Howie, M., Wood, D., Funk, J., Friedman, M., 2004. Intrusion within a simulated water distribution system due to hydraulic transients. ii: Volumetric method and comparison of results. J. Environ. Eng. 130 (7), 778783.

Collins, R., Beck, S., Boxall, J., 2011. Intrusion into water distribution systems through leaks and orifices: initial experimental results. In: Savic, D., Kapelan, Z., Butler, D. (Eds.), CCWI 2011 - Urban Water Management: Challenges and Opportunities.

Cormen, T., 2005. Introduction to algorithms, 2nd Edition. The MIT press.

DWI, 2002. Distribution operation and maintenance strategies (DOMS) - DWI requirements and expectations. Information Letter 15/2002, Drinking Water Inspectorate, London, UK.
URL http://dwi.defra.gov.uk/stakeholders/information-letters/

DWI, 2010. Drinking water 2009: Northern region of England. PDF, last accessed: 2011-10-03.
URL http://dwi.defra.gov.uk/about/annual-report/2009/cir09northern.pdf

DWI, 2011. Drinking water 2010: Letter to minister - England. PDF, last accessed: 2011-07-24.
URL http://dwi.defra.gov.uk/about/annual-report/2010/letter-england.pdf

Eng, S. B., Werker, D. H., King, A. S., Marion, S. A., Bell, A., Isaac-Renton, J. L., Irwin, G. S., Bowies, W. R., 1999. Computer-Generated dot maps as an epidemiologic tool: Investigating an outbreak of toxoplasmosis. Emerg. Infect. Dis. 5 (6), 815–819.

Gauthier, V., Gérard, B., Portal, J., Block, J., Gatel, D., 1999. Organic matter as loose deposits in a drinking water distribution system. Water Res. 33 (4), 1014–1026.

Gray, N., 2008. Drinking water quality: problems and solutions. Cambridge Univ. Pr.

Hatchett, S., Boccelli, D., Haxton, T., Janke, R., Kramer, A., Matracia, A., Panguluri, S., 2011. Real-time distribution system modeling: development, application and insights. In: Urban Water Management: Challenges and Opportunities. Exeter, UK, pp. 743–748.

Hunter, P., Chalmers, R., Hughes, S., Syed, Q., 2005. Self-Reported Diarrhea in a Control Group: A Strong Association with Reporting of Low-Pressure Events in Tap Water. Clin. Infect. Dis. 40, e32–e34.

Husband, P., Boxall, J., 2011. Asset deterioration and discolouration in water distribution systems. Water Res. 45 (1), 113–124.

Husband, P., Boxall, J., Saul, A., 2008. Laboratory studies investigating the processes leading to discolouration in water distribution networks. Water Res. 42 (16), 4309–4318.

Husband, P., Whitehead, J., Boxall, J., 2010. The role of trunk mains in discolouration. Water Manag. 163 (8), 397–406.

Jaeger, Y., Gauthier, V., Besner, M., Viret, B., Toulorge, R., Lemaire, E., de Roubin, M., Gagon, J., 2002. An integrated approach to assess the causes of water quality failures in the distribution system of Caen. Water Supply 2 (3), 243–250.

Khiari, D., Chinn, R., Barrett, S., Matia, L., Ventura, F., Suffet, I., Gittelman, T., Leutweiler, P., 2002. Distribution generated taste-and-odor phenomena. American Water Works Association.

Kirmeyer, G., 2002. Guidance manual for monitoring distribution system water quality. American Water Works Association.

Kistemann, T., Herbst, S., Dangendorf, F., Exner, M., 2001. GIS-based analysis of drinking-water supply structures: a module for microbial risk assessment. Int. J. Hyg. and Environ. Health 203 (4), 301–310.

LeChevallier, M., Gullick, R., Karim, M., Friedman, M., Funk, J., 2003. The potential for health risks from intrusion of contaminants into the distribution system from pressure transients. J. Water and Health 1 (1), 3–14.

Machell, J., Boxall, J., Saul, A., Bramley, D., 2009. Improved representation of water age in distribution networks to inform water quality. J. Water Resour. Plan. and Manag. 135 (5), 382–391.

Machell, J., Mounce, S., Boxall, J., 2010. Online modelling of water distribution systems: a UK case study. Drink. Water Eng. and Sci. 3, 21–27.

Mounce, S. R., Boxall, J. B., Machell, J., 2010. Development and verification of an online artificial intelligence system for detection of bursts and other abnormal flows. J. Water Resour. Plan. and Manag. 136, 309.

Obe, R. O., Hsu, L. S., 2010. PostGIS in Action. Manning.

OFWAT, 2010. Service and delivery - performance of the water companies in England and Wales 2009-10 report; Supporting information: Customer issues. Tech. rep., OFWAT, last accessed: 2010-10-29.
URL http://www.ofwat.gov.uk/publications/los/rpt_los_2009-10/

Okabe, A., Okunuki, K., Shiode, S., 2006. SANET: A toolbox for spatial analysis on a network. Geogr. Anal. 38 (1), 57–66.

Payment, P., Siemiatycki, J., Richardson, L., Renaud, G., Franco, E., Prevost, M., 1997. A prospective epidemiological study of gastrointestinal health effects due to the consumption of drinking water. Int. J. of Environ. Health Res. 7 (1), 5–31.

Sherman, G., 2008. Desktop GIS: Mapping the Planet with Open Source Tools.

Tanyimboh, T., Key, M., 2010. Distribution network elements. In: Savic, D., Banyard, J. (Eds.), Water Distribution Systems. ICE Publishing, Ch. 5.

Trepanier, M., Gauthier, V., Besner, M., Prevost, M., 2006. A GIS-based tool for distribution system data integration and analysis. J. Hydroinformatics 8 (1), 13–24.

USEPA, 2006. Distribution system indicators of drinking water quality. Tech. rep., US Environment Protection Agency, last accessed: 2011-07-23.
URL http://www.epa.gov/ogwdw000/disinfection/tcr/pdfs/issuepaper_tcr_indicators.pdf

USEPA, 2008. EPANET: Software that models the hydraulic and water quality behavior of water distribution piping systems. PDF, last accessed: 2012-07-05.
URL http://www.epa.gov/nrmrl/wswrd/dw/epanet.html

Vreeburg, J. H. G., Boxall, J. B., 2007. Discolouration in potable water distribution systems: A review. Water Res. 41 (3), 519–529.

Walski, T. M., Chase, D. V., Savic, D. A., Grayman, W. M., Beckwith, S., Koelle, E., 2003. Advanced water distribution modeling and management. Haestead Press.

Whelton, A. J., Dietrich, A. M., Gallagher, D. L., Roberson, J. A., 2007. Using customer feedback for improved water quality and infrastructure monitoring. J. American Water Works Assoc. 99 (11), 62–76.

WHO, 2005. Water Safety Plans: Managing drinking-water quality from catchment to consumer. Tech. rep., World Health Organisation, last accessed: 2011-11-14.
URL http://www.who.int/water_sanitation_health/dwq/wsp170805.pdf

WRc, 2008. A review of research on pressure fluctuations in drinking water distribution systems and consideration and identification of potential risks of such events occurring in UK distribution systems (WT1205 / DWI 70/2/220). Tech. rep., WRc plc, last accessed: 2011-11-14.
URL http://www.dwi.gov.uk/RESEARCH/COMPLETED-RESEARCH/reports/DWI70_2_220.pdf