Data validation and data quality assessment

Clemens, F.H.L.R.; Lepot, Mathieu; Blumensaat, Frank ; Leutnant, Dominik ; Gruber, Guenter

DOI
10.2166/9781789060119_0327

Publication date
2021

Document Version
Final published version

Published in
Metrology in Urban Drainage and Stormwater Management

Chapter 9

Data validation and data quality assessment

Francois H. L. R. Clemens-Meyer[1,2,3], Mathieu Lepot[1,4], Frank Blumensaat[5], Dominik Leutnant[6] and Guenter Gruber[7]

[1]Delft University of Technology, Faculty of Civil Engineering and Geosciences, Water Management Department, Delft, The Netherlands
[2]Norwegian University of Science & Technology, Faculty of Engineering, Dept. Civil & Environmental Engineering, Trondheim, Norway
[3]Deltares, Unit Hydraulic Engineering, Delft, The Netherlands
[4]Un poids une mesure, Lyon, France
[5]Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland
[6]Emschergenossenschaft, Essen, Germany
[7]Graz University of Technology, Institut für Siedlungwasserwirtschaft und Landschaftwasserbau, Graz, Austria

ABSTRACT

Once data have been recorded, data validation procedures have to be conducted to assess the quality of the data, i.e. give a confidence grade. Furthermore, gaps may occur in time series and, depending on the purposes, these can be given values by application of e.g. interpolation. Since both aspects are strongly correlated, this chapter gives an overview on the main data validation and data curation/imputation methods. Instead of offering exhaustive details on existing methods, this chapter aims at providing concepts for most popular techniques, a discussion of their advantages and disadvantages in the light of different cases of application, and some thoughts on potential impacts of the choices that must be made. Despite involving mathematical methods, data validation remains a largely subjective process: every data user must be aware of those subjectivities.

Keywords: Data curation/imputation, data quality assessment, data validation, interpolation.

doi: 10.2166/9781789060119_0327

SYMBOLS

(Some symbols are used for different parameters; it should be clear from the context what is meant in a specific case.)

a	fitted parameter in the linear regression
A	cross-sectional area (m^2)
b	fitted parameter in the linear regression
c	constant in an ARMA (Auto Regressive Moving Average) model
d	duration (month)
d_{RC}	maximum delay recommended between two verifications or calibrations (month)
d_{RM}	recommended duration between two maintenance procedures (month)
D	sewer pipe diameter (m)
D_j	Cook's distance for point j
G	test value for the Grubbs test
$G_{Q,t+\Delta t}$	gradient between two discharge values ($m^3/s/min$)
$G_{v,t+\Delta t}$	gradient between two velocity values (m/s/min)
$G_{V1,t+\Delta t}$	gradient between two values (V_1) (various units)
$G_{WL,t+\Delta t}$	gradient between two water level values (m/min)
$Gradient_{min}$	minimal gradient (various units)
$Gradient_{max}$	maximal gradient (various units)
h	water level (m)
h_c	hydraulic gradient (m/m)
i	counter
I	slope of a sewer pipe (m/m)
k_{st}	Manning-Strickler roughness coefficient ($m^{1/3}/s$)
K	quantity in the Mann-Whitney test
l_u	length of the wetted perimeter (m)
m	number of elements in a time series
MSE	mean squared error
n	number of elements in a time series
N	number of data points in a time series for the trend test
$N(\Delta t,T)$	number of observations in T
N_A	number of data points available in a data set
N_D	number of data points labelled as 'Doubtful'
$N_{D\text{-}D}$	number of data points labelled as 'Doubtful' after the final validation
$N_{D\text{-}G}$	number of data points labelled as 'Good' after the final validation
$N_{D\text{-}U}$	number of data points labelled as 'Unsuitable' after the final validation
$N_{equi}(\Delta t,T)$	equivalent number of observations in T, eliminating redundant information
N_E	expected number of data points in a data set
N_G	number of data points labelled as 'Good'
N_M	number of measured data points in a data set
N_T	number of tests applied to a data set
N_U	number of data points labelled as 'Unsuitable'
p	probability value (p-value) or ARMA model first parameter

q	ARMA model second parameter
Q	discharge (m^3/s)
Q_t	discharge value recorded at the date t (m^3/s)
$Q_{t+\Delta t}$	discharge value recorded at the date $t + \Delta t$ (m^3/s)
r	residues in the linear regression
R_{hyd}	hydraulic radius (m), defined as A/l_u
$R(y)$	rank of element y in a series (Mann-Whitney test)
s	standard deviation of V_t during a time window w (various units)
t	time (min, s) or Student t value
t_i	the number of subjects having the rank i (Mann-Whitney test)
T	time series, i.e. pairs of (t_i, x_i) in a time window
T_r	magnitude of the trend
$u(V_1)$	standard uncertainty of the value V_1 (various units)
$u(V_2)$	standard uncertainty of the value V_2 (various units)
$u(V_3)$	standard uncertainty of the value V_3 (various units)
$u(V_{1,t})$	standard uncertainty of the value V_1 recorded at the date t (various units)
u_{MAX}	maximal acceptable standard uncertainty (various units)
v_t	velocity recorded at the date t (m/s)
$v_{t+\Delta t}$	velocity recorded at the date $t + \Delta t$ (m/s)
$\bar{V}$	mean value of V_t during a time window w (various units)
V_1	value 1 (various units)
$V_{1,t}$	value 1 recorded at the date t (various units)
$V_{1,t+\Delta t}$	value 1 recorded at the date $t + \Delta t$ (various units)
V_2	value 2 (various units)
V_3	value 3 (various units)
$V_{I,i}$	interpolated value at the step i (various units)
$V_{LL,CR}$	lower limit for the calibration range test (various units)
$V_{LL,ER}$	lower limit for the expertise range test (various units)
$V_{LL,MR}$	lower limit for the measuring range test (various units)
$V_{LL,PR}$	lower limit for the physical range test (various units)
V_{max}	maximum value of V_t in a time window w (various units)
V_{min}	minimum value of V_t in a time window w (various units)
V_t	value recorded at the date t (various units)
$V_{t+\Delta t}$	value recorded at the date $t + \Delta t$ (various units)
$V_{UL,CR}$	upper limit for the calibration range test (various units)
$V_{UL,ER}$	upper limit for the expertise range test (various units)
$V_{UL,MR}$	upper limit for the measuring range test (various units)
$V_{UL,PR}$	upper limit for the physical range test (various units)
w	time window
WL_t	water level recorded at the date t (m)
$WL_{t+\Delta t}$	water level recorded at the date $t + \Delta t$ (m)
$\bar{x}$	mean value of x_i
x_i	observed values in the linear regression
X_k	element number k in a time series X

$\hat{y}_i$	i^{th} value of y for the fitted linear function
$\hat{y}_{i(j)}$	i^{th} value of y for the fitted linear function leaving out the j^{th} observation in the regression
z	a time series or test value in the Mann-Whitney test
z_q	quantiles of the time series z
z_{max}	maximum threshold value in the Mann-Whitney test
z_{min}	minimum threshold value in the Mann-Whitney test
Z	Z-value in the Z-test for outliers
Z_{max}	threshold in the Z-test for outliers
α	p-value for Type I error, level of confidence
β	p-value for Type II error
γ_i	polynomial coefficient in the AutoRegressive part of an ARMA model
γ_x	weighing function for the autocorrelation function
Δt	time step between two consecutive measurements (min)
$\varepsilon(i)$	residuals at the step i (various units)
ε_i	noise term at step i in an ARMA model
θ_i	polynomial coefficient in the Moving Average part of an ARMA model
ρ	autocorrelation function, Spearman's test value, density (kg/m^3)
ρ_p	autocorrelation function for the process (for the window T)
σ_a	standard deviation of a
σ_b	standard deviation of b
σ_m	standard deviation in the measuring data (various units)
σ_P	standard deviation of the process (various units)
σ_r	standard deviation of the residues (various units)
$\xi(\alpha/2)$	quantile for $\alpha/2$
$\psi(\Delta t, T, T_r)$	quantile value as defined by Equations (9.23) and (9.24)

Motivation anecdote 'Disturbing lamppost'

After installing a Doppler flow meter, on some days a very clear signal was produced and on some days very regular outliers occurred. After analysing a few weeks of data, it became apparent that the outliers only occurred during working hours. This led to the discovery of some industrial discharge of wastewater that interfered with the measuring equipment. Once this was acknowledged the outliers could be safely imputed (in this case, by taking the average of the two adjacent values).

An alternative measure could have been to install a measuring device that could handle the specific type of wastewater without problems. The imputed data were given the tag 'imputed' in the meta-data. A similar issue occurred with a water level sensor: twice a day an enormous outlier occurred, this turned out to be caused by a defect in a lamppost located near to the monitoring location.

Francois Clemens-Meyer

9.1 INTRODUCTION

Data acquired from individual sensors and monitoring stations are prone to systematic and random errors. There are many causes varying from instrumental/device errors, human failure, software bugs, incorrect installations, discontinuities in data communication or power supply, electromagnetic interferences, etc.

This implies that raw data obtained from any monitoring system are not 100% flawless, making a 'blind' use of them potentially risky. Avoiding misleading decisions based on faulty, non-verified data is perhaps the most important reason why data should be carefully validated in any case (see the motivation anecdote). Other reasons to conduct data validation are e.g. avoiding system/catchment misunderstanding, and continuous maintenance and update of the monitoring system.

In addition, validating data on a regular and frequent basis, preferably in (almost) real-time modus, can reveal underlying causes of incorrect or missing data, and hence allow an early-on action to prevent undetected faulty recordings, and improve the maintenance protocols and tasks.

Furthermore, it can help to:

- Improve design and operation protocols.
- Detect failures of sensors and data communication.
- Identify errors which were man-made during installation and maintenance actions.
- In case of malfunctioning elements, preserve potential recourse claims involved.
- Detect and understand abnormal events that occurred at the monitoring location.

In the course of data validation, confidence grades are assigned to the subjected data, to ensure sufficient data quality as required for their purpose. In other words, data validation is a goal-driven process: required data quality changes according to the purpose of the subsequent data analysis. The level of quality strived for is different for, e.g. calculation of annual fluxes to comply with regulation obligations and real-time control of a complex system or process; data users may accept a lesser data quality for the first goal. While continuing with those two goals, the delay between records and validation is another key factor to take into account for the validation methods. If for annual fluxes data can be validated on a weekly or monthly basis, real-time control requires online data validation. The required methods depend on the purpose the data will be used for *and* the timeliness in which the validation can be accomplished after the data had been recorded. Passively measuring and collecting data without clear objectives and/or questions is not only inefficient but also makes the data validation difficult (Lindenmayer & Likens, 2018).

Prior to stepping into data validation and quality assessment procedures, some general conventions are introduced:

- First regarding data themselves:
  - A data point is a value recorded by a monitoring station from a given sensor.
  - This value can be raw (i.e. the raw data recorded by the system), 'processed' once the calibration correction has been performed (see Chapter 7), or have 'basic' and 'classical statistic' validation techniques applied once pre-validation and validation processes have been conducted.
- Then with respect to methods and procedures detailed in this chapter:
  - Data pre-validation is done by application of a sequence of basic procedures applied on corrected data aiming at automatically pinpointing data points which can be erroneous.

- Data validation is done by application of a sequence of more or less advanced procedures (including manual checks by experts) on pre-validated data.
- Data quality assessment is the output of those two steps: a 'validated' data point is then flagged with a colour (e.g. traffic light colour – green, orange or red) or a label (e.g. G for good, D for doubtful or U for Unsuitable, i.e. poor data quality – not fitting for the given use or purpose).

Data validation is about judging data quality in relation to the purpose the data are being meant to be used for. The quality of data points can be judged by a number of criteria:

- Plausibility: data points seem consistent with the expected conditions.
- Consistency: there are no internal inconsistencies in the data, e.g. no data beyond the physical defined interval of possible values.
- Accuracy: data points are too inaccurate and, therefore, meaningless.
- Auditability: this refers to the ability for users of the data set to obtain knowledge on the 'history' of the data, i.e. information on e.g. correction, interpolations, etc. being done on the data and the availability of meta-data on e.g. calibration and maintenance of sensors.
- Synchronicity: time stamps of measured data should be correct in relation to different global time systems, e.g. UTC (Coordinated Universal Time) and, again depending on the purpose the data is collected for, synchronized with associated sensor applications in the same network.

It is recommended to validate data as soon as possible after the measurements have been taken, for which an interval of one week has proven itself in practice, since many available meta-data such as the prevailing weather of the last seven days are mostly still mentally present.

In this sense, data validation is mostly done by computer software (see e.g. Mourad & Bertrand-Krajewski, 2002) largely since the amount of data gathered is normally too huge for manual validation. To date a 100% automatized data validation does not seem possible. What can be achieved, however, is a subdivision in data quality: 'fit for use', 'questionable quality' or 'unfit for use', i.e. in other words 'Good', 'Doubtful' and 'Unsuitable'. Since standardized and general applicable automated procedures are as yet unavailable, the assignment of those confidence levels to data points remains highly subjective with respect to the different methods discussed in this chapter, machine learning training data sets, annotating or labelling. The main challenge is to automate this subdivision in such a manner that false negative and false positive outcomes are minimal, while at the same time keeping the category 'questionable quality' as small as possible. The latter category represents data that may be of use when looked into in more detail, combining domain and process knowledge with familiarity with the system studied, the set-up applied and general engineering experience. Another very important source of information in this respect are the meta-data, such as logbooks (see also Chapters 5, 6, 7 and 10) in which information can be found on maintenance activities and calibration information for each measuring device. For this reason, it is of utmost importance during operation of the measuring systems that this information is logged by the operator with the greatest care and very promptly.

The main objective of this chapter is to make practitioners aware of the techniques that have been widely demonstrated (Figure 9.1) to be useful and work in practice for sensor signal quality within the urban hydrology context.

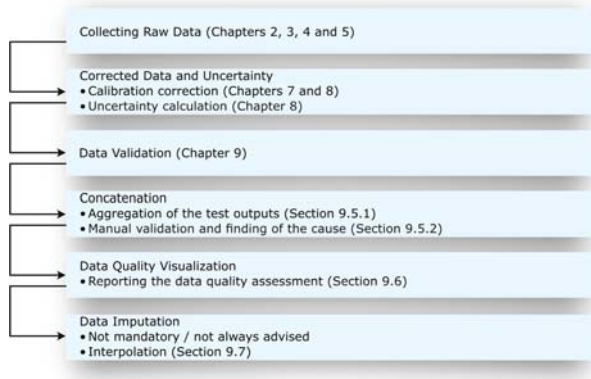


Figure 9.1 Flow chart of the data validation and quality assessment procedures. *Source*: Mathieu Lepot (TU Delft/Un poids une mesure).

This chapter reflects on data validation methods with a focus on urban drainage and stormwater management (UDSM) applications. It is by no means intended to be exhaustive on the subject, as the underlying methods find their roots in a vast and comprehensive research field in mathematics and are widely applied while generic, exhaustive texts are readily available (e.g. ESS, 2018). After defining the characteristics of 'good data', both basic and some 'classical' data validation routines with respect to their purposes are presented.

After a brief review of the different approaches (Section 9.2), this chapter is devoted to the description of the principles (Figure 9.1) of data validation of corrected data (i.e. after implementing calibration corrections on the raw data):

- Pre-validation tests with basic (Section 9.3) and advanced (Section 9.4) data analytical techniques.
- Once each data point has been flagged for each test (according to the results of each test, Section 9.5.1), those flags or labels have to be concatenated (Section 9.5.2) in order to label the data point (Section 9.5.4).
- The data quality representation (Section 9.6.1) and monitoring system analysis (Section 9.6.2) for communication purposes.
- Section 9.7 aims at introducing some methods for data imputation, i.e. replace unsuitable data to achieve some goals (e.g. calculated volumes or fluxes that require complete and equidistant time series). This step, mandatory for some applications, is not really recommended when not needed to avoid working with artificial (interpolated or imputed) data.
- Emerging techniques and methods are briefly introduced in Section 9.8.

This chapter does not aim at offering a complete guideline nor protocol for data validation: it is meant to be an introduction to data validation in itself, a review of the main existing methods (including their advantages and disadvantages) and a list of warnings regarding validation (a mandatory step in UDSM monitoring, but rather prone to bias).

Key messages on data validation

- KM 9.1: *Data validation is mandatory* – never use the data without a careful check.
- KM 9.2: Data validation based on the separation of concerns: *two steps* – (i) pre-validation (unified basic checks), (ii) goal-driven validation.
- KM 9.3: *Purpose dependency:* the results of the data validation depends on the anticipated use of the data.
- KM 9.4: *Subjectivity and reproducibility*: despite there being numerous methods and protocols, data validation remains a subjective process. Keep track of tasks performed.

9.2 CONCEPTS APPLIED IN DATA VALIDATION

9.2.1 What is data validation

Data validation is a process that determines if available data satisfy quality objectives (which have been *a priori* agreed upon) and requirements defined by the anticipated use of the data, here in the context of urban drainage and stormwater management. The process results in adding a quality indicator to each individual data point based on objective criteria as far as possible.

This quality indicator ideally reflects both the *correctness* and the *usefulness* of the data point. Whereas the correctness of a data point can be attributed to the physical meaning, the latter aspect indicates that there is no 'absolute' metric for the quality of a data point. To a certain extent, the evaluation of whether a data point is of high or poor quality depends on the purpose for which the data are to be used. Speaking in these terms, the process of data validation combines (i) an objective, physically-based assessment and (ii) a somewhat subjective perception of how confident the user can be that the measured data point reflects 'reality'.

Example: In real-time control (RTC) applications, there is very little time between obtaining data and using them, which implies that time for an extended validation of the data is limited at best. In such cases, a minimal (if any) validation is performed, e.g.:

- Is the data point there?
- Is the data point within the expected range?

If both questions are answered positively, the data can be used for feeding the RTC algorithm; if one question is negatively answered, the data point is omitted and a default action (in terms of RTC) is taken. In such a situation, it is good practice to store data and the outcome of the two tests mentioned as it allows for a posterior evaluation of the quality of the monitoring, and it may allow the future use of the data for other applications.

Therefore, prior to setting up a protocol for data validation in a given case, case-specific quality levels have to be agreed upon along with a method of organizing the meta-data (see Chapters 5 and 10 on this subject along with Section 9.2.5) that are produced by the validation protocol. This furthermore implies that, when starting a monitoring project, designing the data structure (see Chapter 5) essentially requires considering the envisioned process for the data validation.

9.2.2 How to quantify the quality of data

There are various ways of assigning confidence levels (i.e. quality flags) to individual data points as an indicator for data quality. Practically, the quality assessment of a data point may range from a very basic 0/1 flagging or a more distinct traffic-light labelling 'Good', 'Doubtful', 'Unsuitable' (Table 9.1) to a very refined system in which a wide range of specific qualifications can be added, e.g. attributed to a specific anomaly type (see Table 9.2).

Do's

- *'Only recordings that have a value can be assessed regarding their quality. Keep a record of the fact that there was a missed recording for as long as possible. Do not mix data quality assessment and data curation.'*
- '*The interpretation of data regarding its quality can substantially be qualified through meta-data information. Carefully document meta-data and associate them with data.'*
- *'Prior to assessing data quality, a thorough reflection is mandatory to ensure: (i) are the performed tests useful to reflect likely dubious behaviour of data? and (ii) can all available data be used to conduct individual tests?'*

Differentiating data into just two states, good and poor quality (dichotomous flagging) may be unambiguous and well-achievable for a machine, but insufficient for differentiation. For this reason, often three levels of confidence, e.g. good-doubtful-unsuitable, are assigned, allowing for a more distinguishing assessment. Still, the aspect when a data point is labelled *doubtful* can be somewhat subjective. One labeller may consider an obvious outlier as doubtful whereas the other labeller clearly labels it as *unsuitable*. Clear mind models or 'gold standards' need to be established to avoid subjectivity and allow cross-comparison within one data set. The term 'gold(en) standard' stands for an external criterion representing a kind of benchmark that is the best available under reasonable conditions.

Table 9.1 Example of a traffic-light-system for a gross quality assessment of a data point.

Primal label	Shortened as	Colour	Description
'Good'	'G'	Green	Data point passed all validation tests
'Doubtful'	'D'	Orange	Data point is physically valid but somewhat questionable when evaluated in a wider context
'Unsuitable'	'U'	Red	Data point is physically invalid or is definable erroneous so that it cannot be used
'Missing'	'M'	White or black	Missing data point

Table 9.2 Didactical example of advanced refinement of the quality assessment of a data point. Further examples are given in Leigh *et al.* (2019).

Minor label	Meaning
A	Sensor failure
C	Trend
D	Outlier
E	Constant offset
F	Time shift
G	Value $<$ lower bound of the valid range
H	Value $>$ upper bound of the valid range
I	Low variability, persistently constant value, freeze
J	Imputed by application method x or y
K	Wrong data format
M	Missing time stamp; no data point available

Defining a 'gold(en) standard' is a matter of consensus or opinion, not some kind of statistical property. Whereas dichotomous flagging can be accomplished by a machine, tripartite scoring mostly involves human assessment, i.e. expert knowledge. The general idea is to add relevant information to enhance the probability of finding the cause of a poor data quality.

Automatized flagging of individual signals (no additional information) results in a 0/1 assessment. Adding further information, i.e. extending it to a multi-signal analysis, allows tripartite scoring through a machine.

One can argue about whether or not to include missing data points ('M') in the data quality assessment. Strictly speaking, in a case where there is no measurement recorded, i.e. no data point available, the quality cannot be evaluated. On the other hand, the indication and qualification of gaps in time series at which a data point would have been expected, due to sensor failures, data communication outages, or erroneous data formats allows for characterizing time series regarding their consistency and completeness. The information on amount and distribution of periods at which no data is available may be decisive for the subsequent use of the data, but also for the data validation itself (Section 9.3.6).

Data points labelled as 'Unsuitable' or 'Doubtful' can further be qualified according to the (likely) cause of the less-than-ideal quality. A didactical example of such refined data quality labelling is given in Table 9.2. Note that qualification of quality labels can be supported through operational information, often referred to as meta-data. Meta-data, i.e. additional information on the sensor performance, operation of periphery devices, maintenance actions, and changes to the monitoring environment, are essential to interpret field data correctly (Section 9.2.5).

Authors suggest outputs of those tests: G, D or U. Those are only a suggestion and may or should change according to the monitoring purposes, legal regulations and the expected data quality. However, those suggestions are based on rather long experiences and we advise slight adaptations without completely changing the tests and their outputs.

Data available?

- CL 9.1: *Which?* – Which meta-data are available? Catchment, sewer, sensor, maintenance data.
- CL 9.2: *How?* – How can we make use of this information? Run-off model to correlate catchment, rain and discharge data.
- CL 9.3: *Missing data?* – Is there any data easily acquirable that could be used to conduct additional and relevant tests?

Performed tests

- CL 9.4: *Cover* – Do the applied tests cover any likely behaviour of my data?
- CL 9.5: *Complex situations* – Is (are) there any situation(s) that could bias the output of a few applied tests? Such as backflow effect, complex hydraulic geometry, etc.
- CL 9.6: *Full use* – Do the applied tests make full use of available data?

9.2.3 Subjectivity

The subjectivity in the process of data quality assessment is basically present in discriminating between data in categories 'Good', 'Doubtful' and 'Unsuitable' as defined in Table 9.1. Without going into the discussion of what 'truth' is and whether or not it can be known, 'Good' data is equivalent to 'passed all validation tests'.

This implies that the range of tests a data point is subjected to has a stringency convincing the data user that it is fit for its purpose in the case where the data point passes all these tests. But it does *not* automatically imply that it therefore necessarily reflects the 'truth'.

At the same time, one is striving for a data yield as high as possible, implying that the range of tests should produce a small portion of false positives and false negatives. In other words, the number of data labelled as 'Doubtful' should be minimized, as this fraction of data points requires further attention to investigate the cause of the imperfectness. This can be a very tedious job requiring domain knowledge, and in many cases also knowledge and understanding of the actual situation in the system at hand (e.g. documented as meta-data). Depending on the level of expertise and the solidity of the given information, different answers can be expected when asking a group of experts about the quality of a data point. A certain amount of subjectivity is introduced.

For instance, in the case where rehabilitation works are ongoing, this may result in abnormal sensor readings that may be classified as being 'Doubtful', while recorded values actually represent the (disturbed) process in reality. In the case where one is aware of such an event, the data may be useable after all; otherwise, the data may remain classified 'Doubtful'.

9.2.4 Automation of data validation

The example in the preceding paragraph nicely illustrates that it is likely that data validation cannot be 100% left to computerized algorithms. One way or the other there seems always to be a need for an expert judgement regarding the quality/useability of the data obtained. Having said that, it has to be added

immediately that for practical purposes the application of software, i.e. some degree of automation of data validation, is very favourable.

In principle, data validation can be done manually, which implies that trained individuals have to study the raw data obtained and judge whether or not the data obtained are fit for purpose. Manual data validation, however, has some serious drawbacks:

- It is very labour-intensive and therefore expensive.
- The criteria for accepting/rejecting data points are subjective and will result into a non-reproducible assessment.
- For some purposes, e.g. RTC applications, the processing time is simply too long to be practically applicable.

For these reasons, a certain degree of automated data validation is applied in practice. This may at least relieve the workload, although applied schemes seem to show a variation of success. For example, the validation scheme as proposed by Upton & Rahimi (2003) for validating data from tipping bucket rain gauges proved to be very efficient: up to 90% of the anomalies proved to be correctly identified after manually checking. However, when applying the same procedure to a grossly similar case, Schilperoort (2011) found a percentage of only 60% of correctly identified anomalies. This reduced yield in the latter case was caused by the huge amount of data missing due to data communication issues and the lack of meta-data, the latter underlining the importance of keeping track of such additional information.

9.2.5 Meta-data

Meta-data is essential to interpret field data correctly. When trying to identify causes of data being classified as 'Doubtful' or 'Unsuitable', the presence of additional information, i.e. meta-data, is vital. Ideally, this information on sensor operation and maintenance actions is in standardized logbooks. Meta-data should be collected systematically, i.e. formalized in individual categories, and continuously over time.

Meta-data information can comprise (non-exhaustive list):

- Sensor maintenance actions.
- Antecedent and last calibration results.
- Access to plans for and reports on construction works.
- Data from adjacent and/or related monitoring sites, e.g. rain gauges to discriminate between dry and wet weather or the reading from a sensor showing overlap in its readings – see also Section 6.2 on macro design.
- Weather reports, e.g. thunderstorms may cause loss of communication or damage caused by an electromagnetic pulse (EMP).
- Development over time of the sensor performance; sometimes a problem may repeat, so using the sensor history may hint at a (external) cause that induces the problem.
- Data on the performance of similar sensors (brand, production batch, etc.), this may reveal some inherent issues with the device(s) used. This information may be used to upgrade the system when replacing parts.

It is a managerial decision to what extent of detail one should go in gathering meta-data when operating a long-term observation campaign, as these administrative tasks tend to expand (certainly in large bureaucratic organizations). It is essential – not just against the background of an increasing degree of automation of data handling – to foresee gathering information on the performance of the monitoring system as a whole and in a systematic manner and for well-defined purposes only (Chapter 6). This implies man-hours spent on maintenance, data analysis and validation are monitored as well. Such systems allow fine-tuning of the

design and the operation of the monitoring system to ensure a certain level of data quality. The creation of standard operational procedures (SOPs) for the documentation of all meta-data to be recorded is recommended to ensure that the meta-data are documented as uniformly as possible and to reduce subjective elements as much as possible.

9.3 BASIC CHECKS

There are numerous data validation methods, from very simple to very advanced ones. This section provides an overview of the existing ones, their advantages, disadvantages, and limitations. The methods are divided in two categories, the basic checks (Section 9.3) and the advanced validation ones (Section 9.4). The choice of those tests is strongly dependent on what purpose(s) the data are collected for, the available skillset and knowledge of persons in charge, and the delay between measurement and validation. All the methods presented in this chapter are applied on data from calibrated sensors. Contrary to the methods discussed in Section 9.4, the basic methods can easily be automated.

Thresholds

Most of the tests presented hereafter are based on thresholds. The output of each test is directly dependent of the selected threshold(s). Careful attention must be paid to the threshold selection: the output can be too pessimistic or too optimistic.

This warning is valid for everyone: from data provider, data curator to the data user. Always keep in mind a famous quote from W.S. Churchill: 'I only believe in statistics that I doctored myself'.

9.3.1 Test on plausibility

Plausibility tests using numerical criteria or based on common sense usually do not require significant resources and are hence suitable for low-computation online validation.

9.3.1.1 Physical range

This is a first test, only based on physical boundaries of the measured phenomena: a water level at free surface flow mode cannot be negative or higher than the diameter of a circular pipe, the temperature of liquid water cannot be below 0 or above 100 degrees Celsius (at atmospheric pressure), rain intensity cannot be negative, etc. If the data values are outside the physical range, they should be labelled as 'Doubtful' or 'Unsuitable' for this test. A value $V_{1,t}$ recorded at the date t successfully passes this test if Equation (9.1) is verified.

$$V_{LL,PR} \leq V_{1,t} \leq V_{UL,PR} \tag{9.1}$$

where $V_{LL,PR}$ and $V_{UL,PR}$ are, respectively, the lower and upper limits of the physical range.

Example: In a circular pipe or 1000 mm of diameter, the water level values have the following threshold: $V_{LL,PR} = 0$ mm and $V_{UL,PR} = 1000$ mm. If a recorded value ($V_{1,t}$) is negative or higher than 1000 m, Equation (9.1) is not verified and, therefore, this value is flagged as 'Doubtful' or 'Unsuitable' with respect to this test on physical range (Figure 9.2).

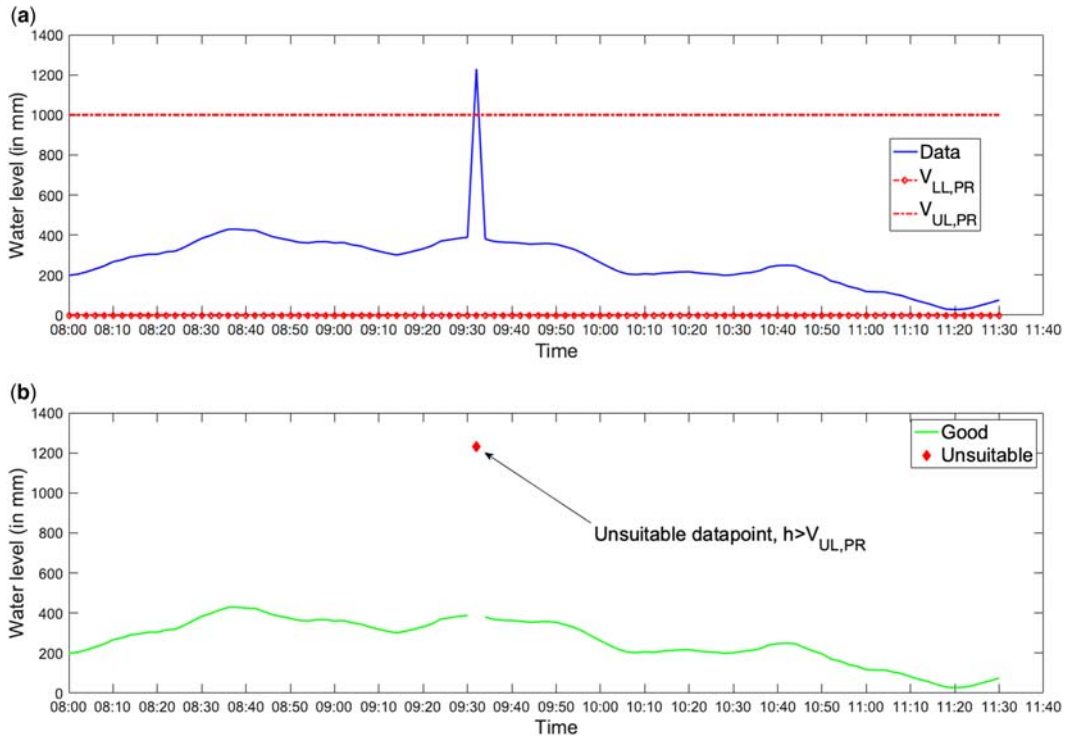


Figure 9.2 (a) water level data (blue) and physical range thresholds (red); (b) 'Good' (green) and 'Unsuitable' (red) data according to the physical range test. *Source*: Mathieu Lepot (TU Delft/Un poids une mesure).

Limitations

The statement made in Section 9.3.1.1 for a single data point is certainly correct. However, if pressurized flow may occur at the measurement location, piezometric water level sensors may give a water level greater than the pipe diameter (i.e. the flow pressure at the measurement section). Even if this test is rather easy to set up, it requires some expertise and knowledge about (un)likely conditions at the measurement point.

9.3.1.2 Measuring range

This test is rather similar to the previous one, but based on the measuring range of each sensor. Sensors are designed to measure and work over certain ranges of measurement or environmental conditions. If the recorded value is outside the measuring range or has been recorded in unusual conditions, it should be labelled as 'Doubtful' or 'Unsuitable' for this test (Equation (9.2)).

$$V_{LL,MR} \leq V_{1,t} \leq V_{UL,MR} \tag{9.2}$$

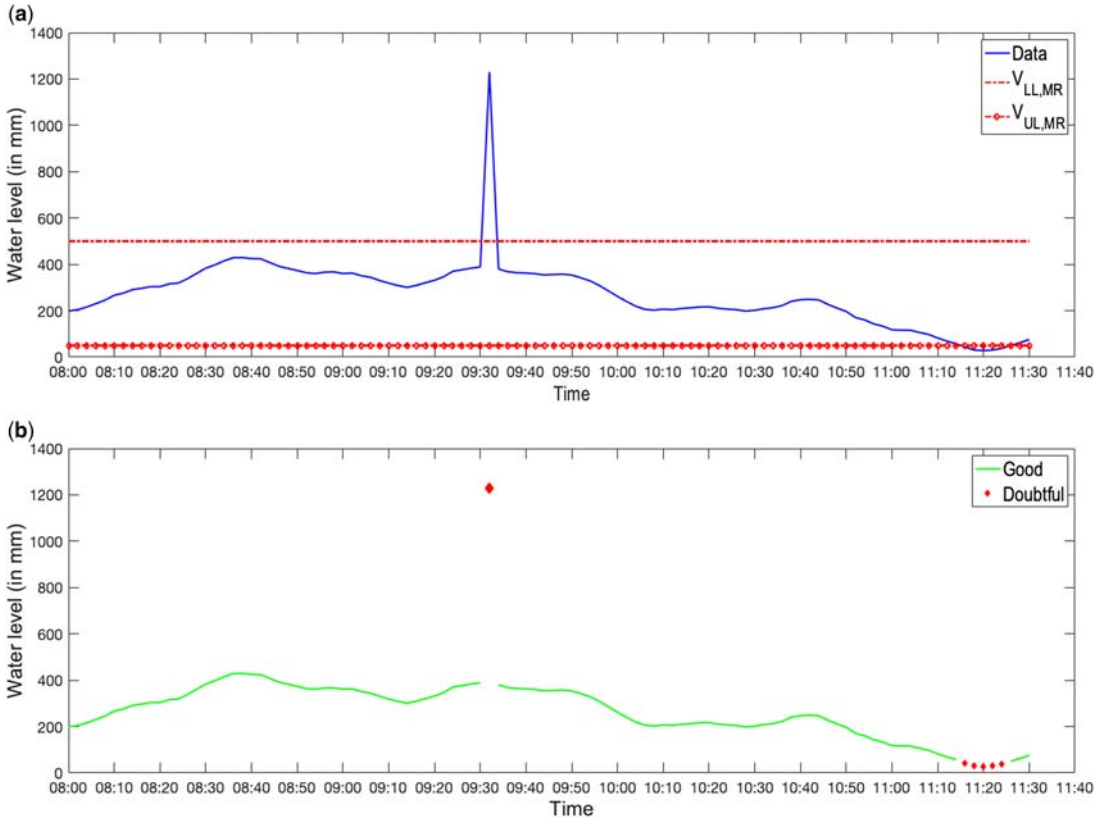


Figure 9.3 (a) water level data (blue) and measuring range thresholds (red); (b) 'Good' (green) and 'Doubtful' (red) data according to the measuring range test. *Source*: Mathieu Lepot (TU Delft/Un poids une mesure).

where $V_{LL,MR}$ and $V_{UL,MR}$ are, respectively, the lower and upper limits of the measuring range, i.e. they are sensor dependent. To avoid 'not good data' for this test, the measuring ranges of the different devices have to overlap.

Example: In the same pipe as in the previous example, the water level sensor has a measuring range between 50 mm and 500 mm (according to its specifications). The water level values recorded by this sensor have the following threshold: $V_{LL,MR} = 50$ mm and $V_{UL,MR} = 500$ mm. If a recorded value ($V_{1,t}$) is lower than 50 mm (e.g. 30 mm) or higher than 500 m, Equation (9.2) is not verified and, therefore, this value is flagged as 'Doubtful' with respect to this test on measuring range (Figure 9.3).

9.3.1.3 Calibration range

This test is quite similar to the previous ones. A sensor is calibrated over a given range, from the minimum to the maximum values of calibration standards. For this test also, if the value is outside the calibration range, it should be labelled as 'Doubtful' or 'Unsuitable' for this test (Equation (9.3)).

$$V_{LL,CR} \leq V_{1,t} \leq V_{UL,CR} \tag{9.3}$$

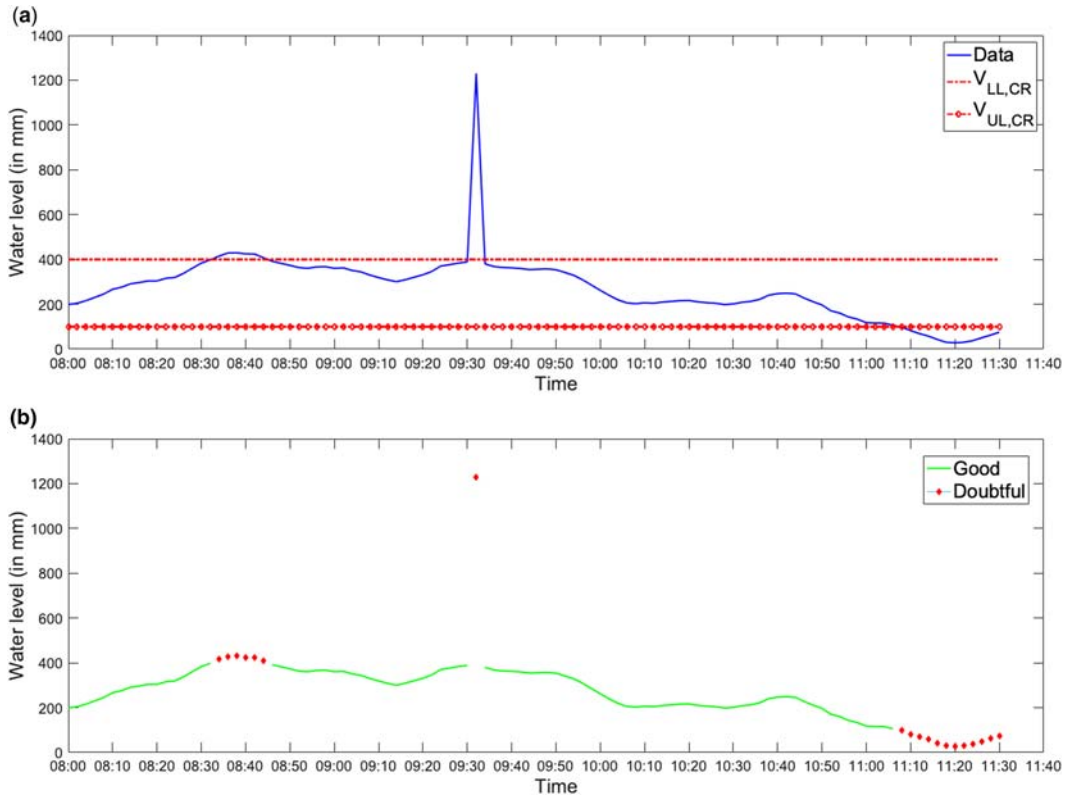


Figure 9.4 (a) water level data (blue) and calibration range thresholds (red); (b) 'Good' (green) and 'Doubtful' (red) data according to the calibration range test. *Source*: Mathieu Lepot (TU Delft/Un poids une mesure).

where $V_{LL,CR}$ and $V_{UL,CR}$ are, respectively, the lower and upper limits of the calibration range, i.e. calibration standards dependent. To avoid 'not good data' for this test, the calibration standard values should cover the full range of expected conditions.

Example: The water sensor used in the previous example has been calibrated from 100 ($V_{LL,CR}$) to 400 ($V_{UL,CR}$) mm. If a recorded value ($V_{1,t}$) is outside those boundaries, it should be flagged as 'Doubtful' with respect to this test on calibration range (Figure 9.4).

9.3.1.4 Expertise range

This range test requires domain knowledge and knowledge of the system under study. Despite all the previous checks, experts may judge that a value is doubtful if measured under certain conditions, generally narrower than the ones given in the design specifications of a sensor (Equation (9.4)). As an example, a Doppler probe can measure water levels in a range from 0 to 0.7 m, but experts may consider that data cannot be fully trusted outside 0.1 to 0.4 m due to the intrinsic limitation of the probe and acoustic attenuation of the signal.

$$V_{LL,ER} \leq V_{1,t} \leq V_{UL,ER} \tag{9.4}$$

where $V_{LL,ER}$ and $V_{UL,ER}$ are, respectively, the lower and upper limits of the expertise range.

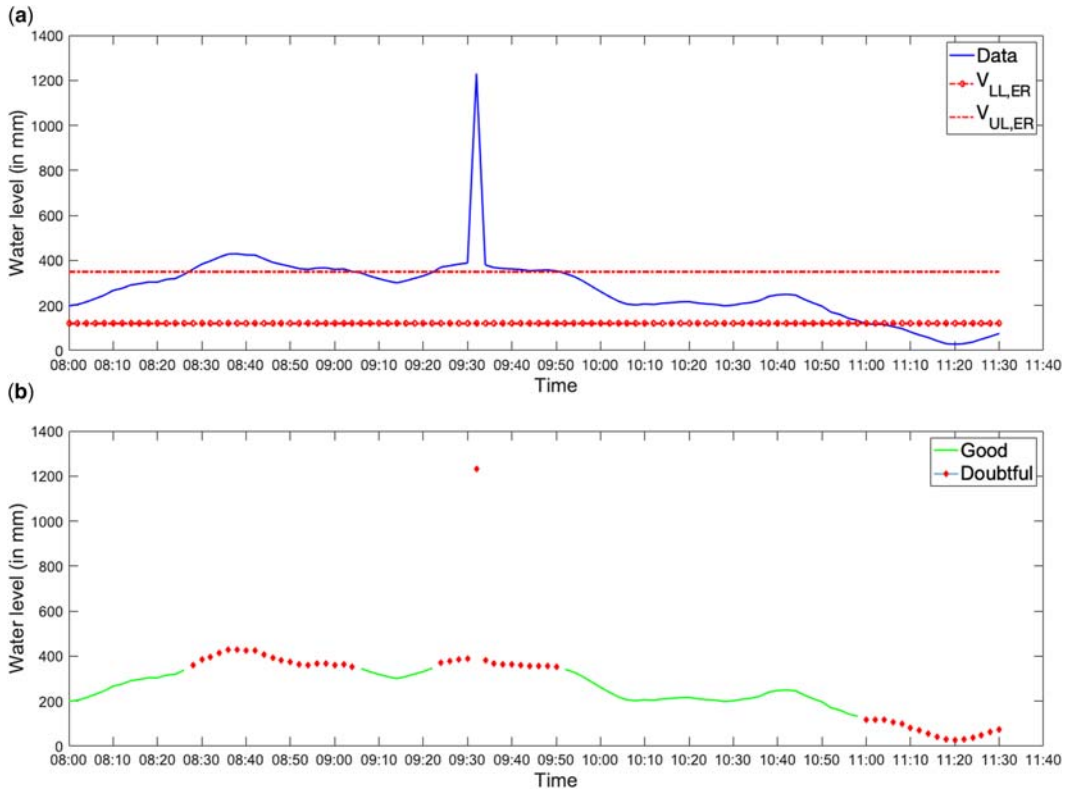


Figure 9.5 (a) water level data (blue) and expertise range thresholds (red); (b) 'Good' (green) and 'Doubtful' (red) data according to the expertise range test. *Source*: Mathieu Lepot (TU Delft/Un poids une mesure).

Example: Experience shows that the recorded values given by sensor are doubtful below 120 mm ($V_{LL,ER}$) and above 350 mm ($V_{UL,ER}$). If a recorded value ($V_{1,t}$) is outside those boundaries (e.g. 110 mm or 360 mm), it should be flagged as 'Doubtful' with respect to this test on expertise range (Figure 9.5).

9.3.1.5 Gradient range

Time series give information on phenomenon dynamics. With some expertise, the usual dynamics of the phenomena are known and can be used to validate or not the data. Time series showing no or too sudden dynamics can be considered as doubtful. Given a value V_1 recorded at two different dates (t and $t+\Delta t$), the value $V_{1,t+\Delta t}$ could be considered as doubtful if one of the Equation (9.5) is verified.

$$\begin{cases} G_{V1,t+\Delta t} = \dfrac{V_{1,t+\Delta t} - V_{1,t}}{\Delta t} > Gradient_{MAX} \\ V_{1,t+\Delta t} = V_{1,t} \\ G_{V1,t+\Delta t} = \dfrac{V_{1,t+\Delta t} - V_{1,t}}{\Delta t} < Gradient_{MIN} \end{cases} \tag{9.5}$$

where $Gradient_{MAX}$ and $Gradient_{MIN}$ are, respectively, the maximum and minimum likely gradients for the given phenomenon.

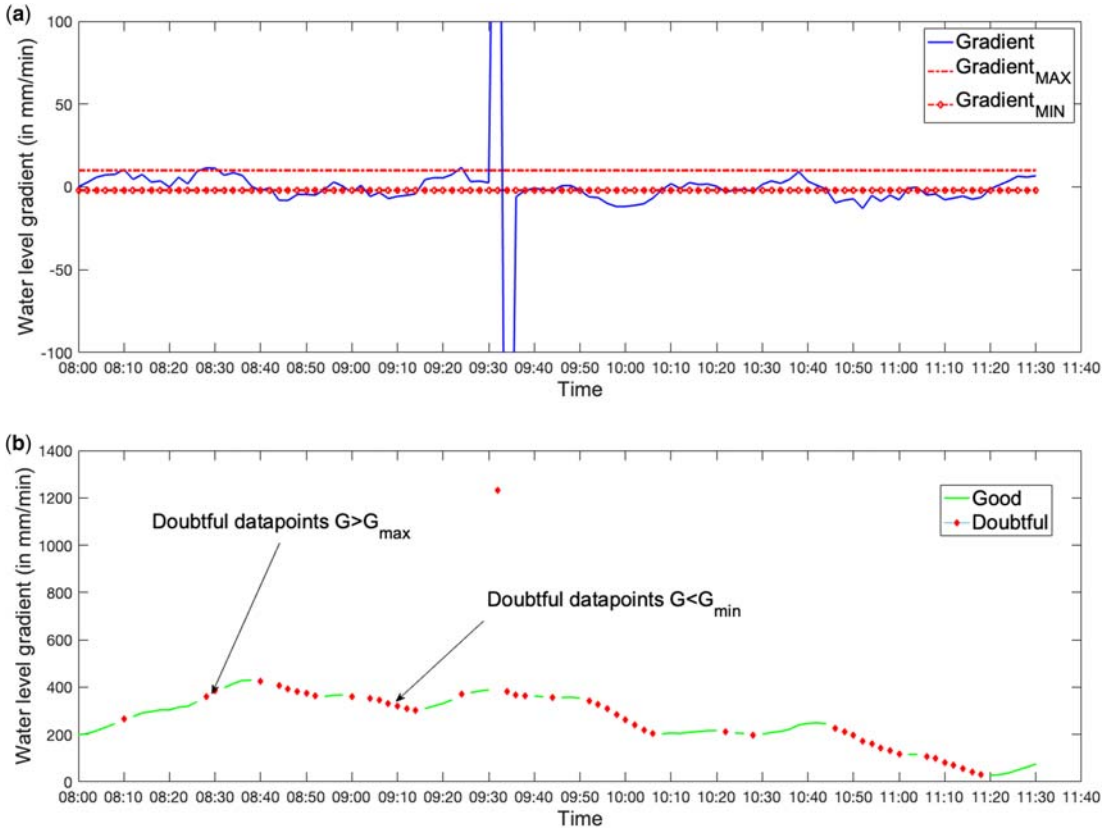


Figure 9.6 (a) water level data (blue) and gradient range thresholds (red); (b) 'Good' (green) and 'Doubtful' (red) data according to the gradient range test. *Source*: Mathieu Lepot (TU Delft/Un poids une mesure).

Example: Experience shows that the hydraulic dynamics of the catchment barely ever exceed 10 mm/min for the rising part of a storm event and are, in most cases, below 2 mm/min for the declining part (Figure 9.6).

Assuming a Δt equal to 2 min, the difference $(V_{1,t+\Delta t} - V_{1,t})$ should not be higher than 20 mm when the flow rises or be lower than 4 mm when the flow decreases. Otherwise, the value should be flagged as 'Doubtful'. As an example, the following couples $(V_{1,t+\Delta t}, V_{1,t})$ will flag $V_{1,t+\Delta t}$ as 'Doubtful': (160,190), (50,45) and respectively (70,70) – for these couples the gradients are, respectively, 15, −2.5 and 0 mm/min.

9.3.2 Test on consistency

Consistent data are logical and do not contradict themselves. Inconsistencies are usually caused by gross errors (DWA, 2011).

9.3.2.1 Comparison between redundant recordings (signal redundancy)

If, as advised, a monitoring station has redundant sensors to measure the same type of information (e.g., water level, velocity, etc.), each value can be compared to the other ones in order to identify if one or a

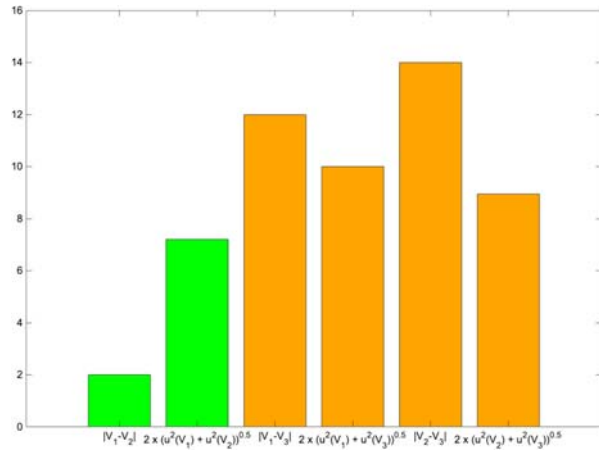


Figure 9.7 Comparison between absolute difference and their uncertainties (with 95.4% confidence level). *Source*: Mathieu Lepot (TU Delft/Un poids une mesure).

few of them are too different from the other ones. Given three measured values V_1, V_2 and V_3 and their associated standard uncertainties $u(V_1)$, $u(V_2)$ and $u(V_3)$, the value V_1 can be considered as doubtful if it is significantly different from the other ones, i.e. if V_1 satisfies the subsequent three Equation (9.6).

$$\begin{cases} |V_1 - V_2| < 2\sqrt{u^2(V_1) + u^2(V_2)} \\ |V_1 - V_3| \geq 2\sqrt{u^2(V_1) + u^2(V_3)} \\ |V_3 - V_2| \geq 2\sqrt{u^2(V_3) + u^2(V_2)} \end{cases} \tag{9.6}$$

This test requires at least three values. If there are only two, the test is just able to say that both values are significantly different, without pinpointing which one might be wrong. This test is applicable on recorded values or calculated values, such as water levels, velocities and discharges calculated from those two.

Example: At the same monitoring location, and while using the uncertainty calculation methods presented in Chapter 8, three water levels are recorded with known uncertainties: $V_1 = 50$ mm and $u(V_1) = 3$ mm, $V_2 = 48$ mm and $u(V_2) = 2$ mm and $V_3 = 62$ mm and $u(V_3) = 4$ mm. The three Equation (9.6) are verified: V_3 is flagged as 'Doubtful' while V_1 and V_2 pass the consistency test, i.e. are flagged as 'Good' (Figure 9.7).

9.3.2.2 Dynamic consistency

As an example, the dynamic behaviour of water level, velocity and discharge should be consistent: under standard condition, e.g. when no downstream effects occur, if the water level increases, the velocity increases and the discharge too. Consistencies between gradients could be checked to identify potentially doubtful data. While reusing the same notation as in Equation (9.5) for this example, i.e. $G_{WL,t+\Delta t}$, $G_{v,t+\Delta t}$ and $G_{Q,t+\Delta t}$ being the gradients for water level, velocity and discharge, the velocity can be considered as doubtful if Equation (9.7) are verified.

$$\begin{cases} (G_{WL,t+\Delta t} > 0 \text{ and } G_{Q,t+\Delta t} > 0) & \text{or} \quad (G_{WL,t+\Delta t} < 0 \text{ and } G_{Q,t+\Delta t} < 0) \\ (G_{WL,t+\Delta t} > 0 \text{ and } G_{v,t+\Delta t} < 0) & \text{or} \quad (G_{WL,t+\Delta t} < 0 \text{ and } G_{v,t+\Delta t} > 0) \\ (G_{v,t+\Delta t} < 0 \text{ and } G_{Q,t+\Delta t} > 0) & \text{or} \quad (G_{v,t+\Delta t} > 0 \text{ and } G_{Q,t+\Delta t} < 0) \end{cases} . \tag{9.7}$$

The potential combinations of such tests are endless and too site specific to draft an exhaustive list here.

Example: Figure 9.8 presents an example of such a test. The gradients for water level, velocity and discharge data are plotted at the top (Figure 9.8a). Based on sign analysis (Equation (9.7)), data are then labelled as 'Good' or 'Doubtful' according to this test (Figure 9.8b).

9.3.2.3 Time stamp consistency

Measurement data always have a time reference, as each individual measurement point has been observed and recorded at a specific time. If measurement data are recorded at a regular time interval (e.g. each minute), the distance between two consecutive time stamps is equal. However, depending on the quality of hard- and software installed, an expected equidistance might be interrupted, resulting in irregular time series causing loss of information. Irregular time series show unexpected time gaps or even different measurement data assigned to an identical time index.

Testing the time stamp consistency of measurement data requires knowledge of whether the signal is expected to be equidistant or have an irregular interval, and this must be communicated before the measurement is under operation. Estimating the correct periodicity after data has been collected would otherwise require statistical tests to be applied.

Nowadays, monitoring stations tend to measure and record at fixed and regular time intervals. However, irregularly-recording measurement stations are still maintained, e.g. to save battery life when remotely installed. Regular time changes due to daylight saving taking place twice a year can be a further cause of time stamp inconsistencies. If possible, these should be avoided by, for example, storing the measured data uniformly with a global time system, e.g. UTC (Coordinated Universal Time).

9.3.3 Test on accuracy

If a value is too inaccurate, i.e. if its standard uncertainty is higher than a given threshold adapted to its future use, it should be labelled as 'Doubtful' or 'Unsuitable' (Equation (9.8)).

$$u(V_{1,t}) \leq u_{MAX} \tag{9.8}$$

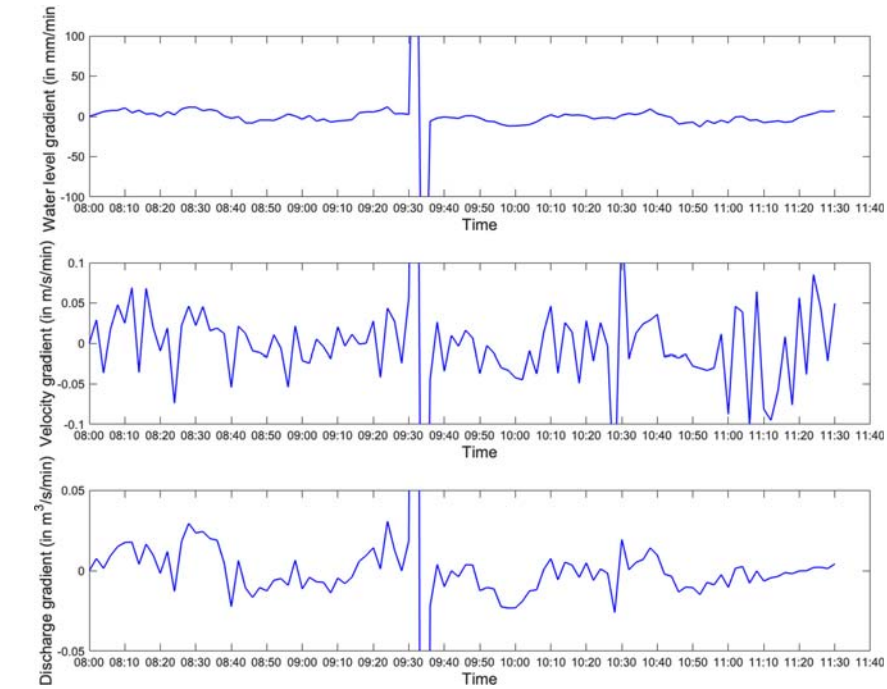
where $u(V_{1,t})$ is the standard uncertainty in the value $V_{1,t}$ and u_{MAX} is the threshold of the uncertainty. This test can be extended to two thresholds, one for 'Doubtful' and another one for 'Unsuitable'. This rather basic test is sensitive to the selected threshold, which is sensor specific and could either be absolute or relative.

Given certain standards, by law or for the final use, a value can be labelled as D or B if there is no uncertainty associated.

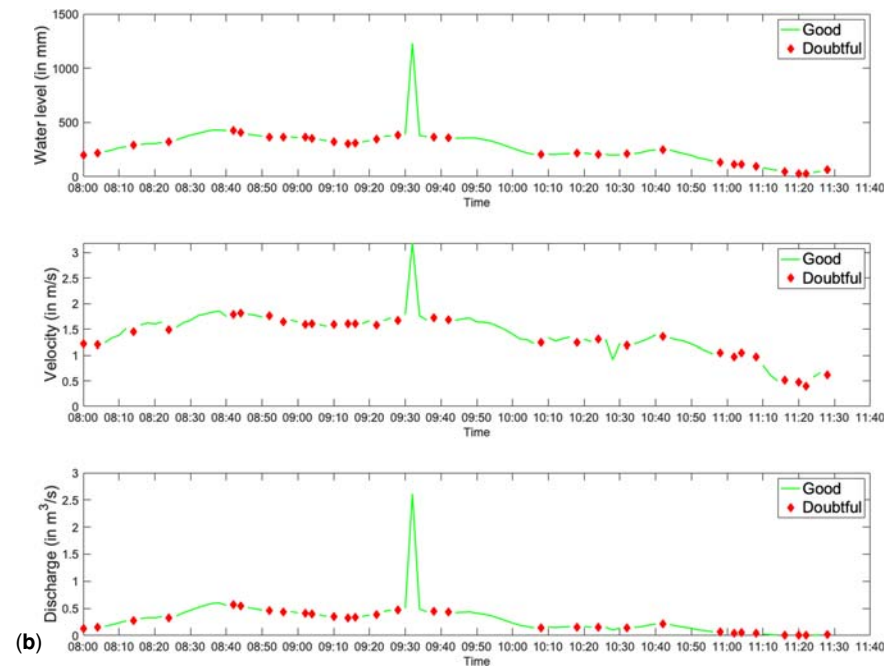
Example: At the same monitoring location, a value $V_{1,t} = 67$ mm is recorded by a water level sensor. Once the uncertainty calculations are done, $u(V_{1,t})$ is equal to 3 mm (see Chapter 8). Given a u_{MAX} of, e.g., 5 mm, $V_{1,t}$ is flagged as 'Good' (Figure 9.9).

9.3.4 Test on auditability

Although the term auditability is mostly used in accountancy, the principle of trackability of what 'happened' to a measured parameter value can be transferred to monitoring projects. In the end, the product that a monitoring project has to deliver is data of a known and well described quality. To be able to implement the underlying principle of quality control – 'collect data on the manner in which the procedures and protocols in an organization are applied and learn from evaluating them' – to monitoring projects, the following aspects are to be considered:



(a)



(b)

Figure 9.8 (a) gradient for water level, velocity and discharge time series; (b) output of the test based on dynamic consistencies. *Source*: Mathieu Lepot (TU Delft/Un poids une mesure).

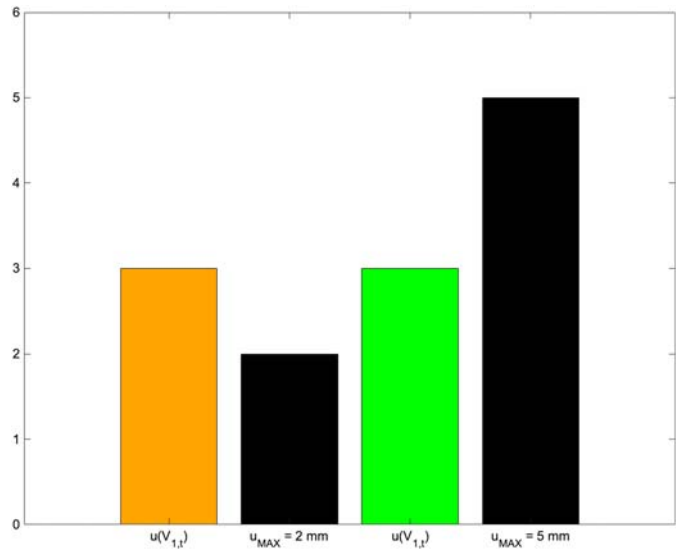


Figure 9.9 Comparison between absolute difference and their uncertainties (with 95.4% confidence level). *Source*: Mathieu Lepot (TU Delft/Un poids une mesure).

- Formulate clear procedures and protocols with respect to data storage, postprocessing and authorization levels of personnel working with or on the data.
- Make sure meta-data is kept accurate and up to date.
- Be sure to have back-ups of the 'raw' data at all times, this ensures being able to 'redo' postprocessing when in the course of time protocols or procedures applied earlier turn out to have flaws (e.g. some bug in a piece of computer code).

Testing on auditability may well be part of a regular/general quality systems check in an organization which not only tests the applicability and application of all protocols and procedures but also the motivation and awareness of people working with them.

9.3.4.1 Calibration

As introduced in Chapter 7 on calibration methods, sensors need to be calibrated regularly. If, for some reason, the data have not been corrected for calibration (no calibration has been done, the data of the calibration correction are not recorded or stored), i.e. the data are raw, those data points should be considered as 'Unsuitable'.

Example: If there is no calibration correction (Chapters 7 and 8), which can be identified by either the absence of corrected data, of calibration correction parameter or (but not always) no difference between the corrected and the raw values, this value has to be flagged as 'Unsuitable' according to the calibration test.

9.3.4.2 Latest calibration

The duration since last verification or calibration might be used for data validation. Given a sensor that requires a monthly calibration, the data recorded between 1 and 2 months after the latest calibration could be considered as 'Doubtful' for a longer delay. Given the maximum delay recommended between

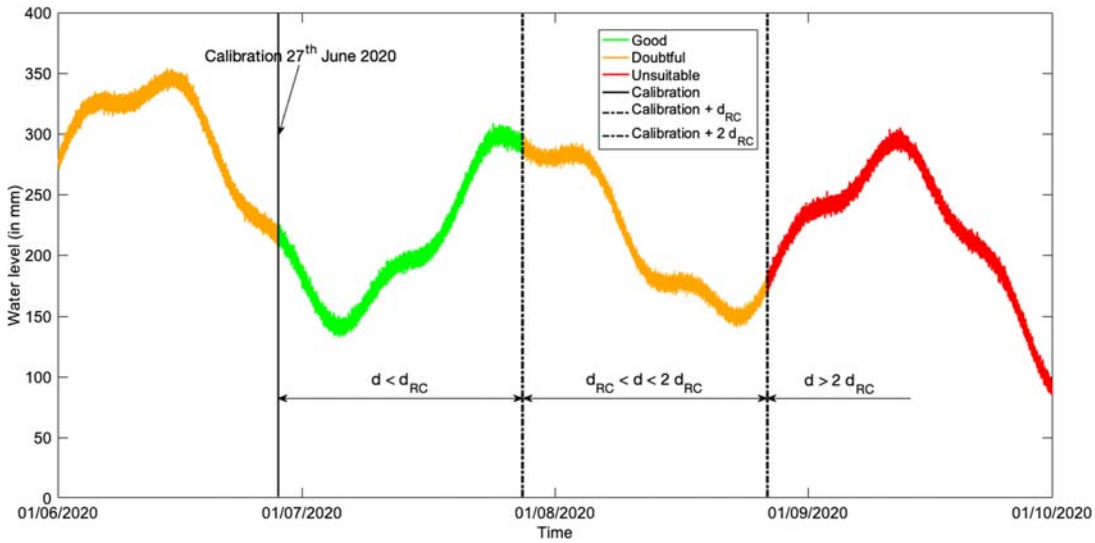


Figure 9.10 Test based on the duration since the last calibration. *Source*: Mathieu Lepot (TU Delft/Un poids une mesure).

two verifications or calibrations (d_{RC}), the duration since the last maintenance (d) should respect one of Equation (9.9):

$$\begin{aligned} &d \leq d_{RC} \\ &d_{RC} < d \leq 2 \times d_{RC} \\ &d > 2 \times d_{RC} \end{aligned} \tag{9.9}$$

The first test leads to data flagged as 'Good', the second one to 'Doubtful' and the third one to 'Unsuitable'. This refers to the ability for users of the data set to obtain knowledge on the 'history' of the data, i.e. information like corrections, interpolations, etc. being done on the data and the availability of meta-data on calibration and maintenance of sensors.

Example: Manufacturer or user expertise on a water level sensor advises a verification (and a potential re-calibration) every month. If the value has been recorded within a month since the last verification (or re-calibration), it is flagged as 'Good'. If this duration is longer than a month but shorter than 2 months, the value is flagged as 'Doubtful'. The flag is 'Unsuitable' if this duration d exceeds 3 months (Figure 9.10).

9.3.4.3 Maintenance

Sensors and data acquisition systems require maintenance. During maintenance or calibration, manually or automatically recorded data points should be considered as 'Unsuitable'. If there is no maintenance operation log in the system (automatic or logbook), this test cannot take place.

Example: A water level sensor has been cleaned between 2:00 and 2:30 pm. All the data recorded between those hours are flagged as 'Unsuitable' because the measurements have been disturbed by the cleaning actions. Outside this time slot, the values are considered as 'Good' according to this test.

9.3.4.4 Last maintenance

Maintenance has to be done on a periodic basis. According to manufacturer or expert recommendations, the delay between two maintenance activities should not exceed a certain duration d_{RM} for a given sensor. Therefore, if the data point has been recorded within this delay, it could be considered as 'Good', between this and twice this duration as 'Doubtful', and beyond twice this duration as 'Unsuitable' (Equation (9.10)).

$$\begin{aligned} &d \leq d_{RM} \\ &d_{RM} < d \leq 2 \times d_{RM} \\ &d > 2 \times d_{RM} \end{aligned} \tag{9.10}$$

Those durations and thresholds can be shortened based on expert judgements and site knowledge. It is not recommended to extend both values (durations and thresholds).

Example: A velocity sensor has to be cleaned every two months according to the manufacturer recommendations or your expertise. If the value has been recorded within 2 months since the last cleaning operation it is flagged a 'Good'. If this duration is longer than 2 months but shorter than 4 months, the value is flagged as 'Doubtful'. The flag is 'Unsuitable' if this duration d exceeds 4 months.

9.3.5 Test on synchronicity

Depending on their quality, quartz timers of measuring systems tend to drift (e.g. Leutnant *et al.*, 2015). In order to guarantee the accuracy of time stamps, measurement devices should be regularly synchronized, either automatically with an available time-server or manually in the case where an automated synchronization is not possible. As in urban drainage normally dynamic processes are studied involving an interest in the relation between e.g. rainfall and discharge, the mutual synchronicity between time series obtained from a monitoring network is of key importance. Depending on the goal of the analysis, a certain time shift between time series may be tolerable. However, preferably all individual series will share the same time basis and have an equal time interval between readings (temporal equidistance). The former is particularly important to avoid complications when analysing interrelations between different time series. An important, yet trivial, issue to address related to synchronicity is to make sure that, when daylight saving time is taken into account, all sensors in the network switch at the same moment in time, which in practice is not always easily achieved. Interpretation of time series can be significantly hampered by a disparate adjustment of the time stamps.

Deficiency analysis can be accomplished by a fragment-wise application of cross-correlation and/or the method of least squares. Sensor networks with wireless data transmission inherently ensure synchronicity between sensors as they synchronize regularly with an external time reference, such as the clock of the central computer used for data acquisition, a time-server via Network Time Protocol (NTP), through GPS signalling, or via DCF77. With DCF77, the legal time is transmitted from Frankfurt, Germany, across Europe according to the standards ISO 8601 or DIN EN 28601. DCF77 is registered on the international frequency list of the ITU (International Telecommunication Union) as 'Fixed Service' with the carrier frequency 77.5 kHz and the bandwidth 2.4 kHz.

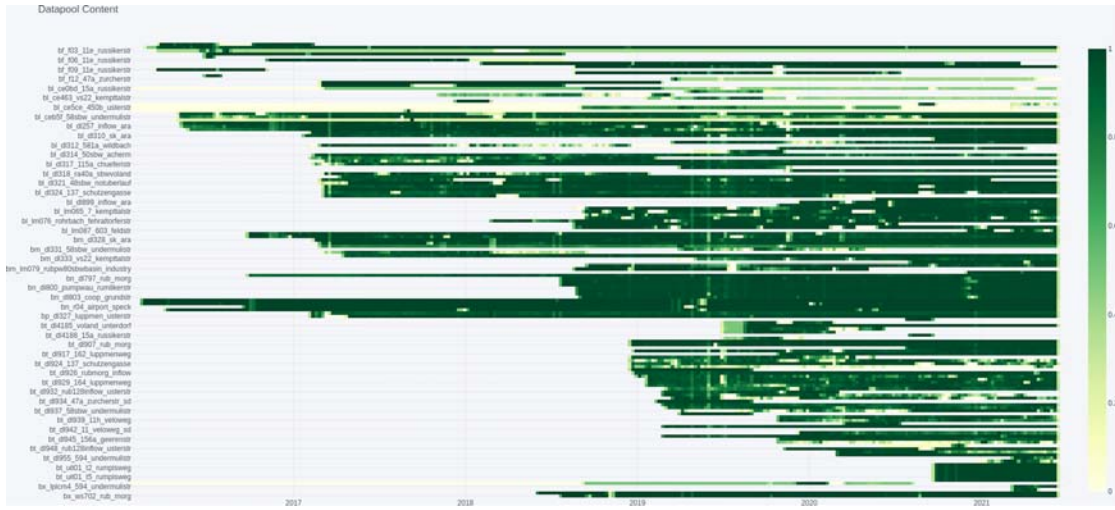


Figure 9.11 Heatmap that illustrates the consistency (here: degree of completeness) of various urban hydrology monitoring series. *Source*: Eawag. Web-based dashboard produced with the plotly app. Data: www.uwo-opendata.eawag.ch; Dashboard: Christian Förster (Eawag).

9.3.6 Test on completeness (degree of incompleteness)

To assess the usefulness of a set of measurements, it may be interesting to see how complete a time series is in terms of missing recordings (Figure 9.11). Depending on the sensors, the data communication and the operational environment at the monitoring location, the measuring system may be subject to outages at which an instance is not measured, not recorded or not transmitted. A missing recording, i.e. a gap in the time series, is the consequence. Considered over a longer period, the number of missing records may accumulate to a considerable amount. However, quite a few advanced data validation methods rely on gapless data series – they would not work with a single inconsistency in the time vector.

But depending on the distribution of missing recordings, e.g. small but frequently occurring gaps or few but large gaps, the missing data points may be curated or not (Section 9.7). Therefore, a *posterior* assessment of a data series with regard to degree of completeness and distribution of missing records is relevant. Even better, real-time data validation triggers an alarm in the case where recordings do not arrive at central servers, allowing *ad hoc* maintenance to be carried out. Assessment can be accomplished by analysing the plot of the gap distribution, i.e. number of days (% of total monitoring period) for which a certain accumulated gap length per day/hour is exceeded. Log(x) may be more favourable for a visual inspection. This test delivers two quality flags: 'Good' or 'Missing'. Note that the degree of completeness of a data set can change throughout the validation process, as some data that is identified as implausible may need to be excluded.

9.3.7 Summary of main basic tests available for data pre-validation

Table 9.3 proposes a non-exhaustive list of basic tests. Those tests are threshold dependent: the test outputs are sensitive to the chosen thresholds. The output can be 'Good', 'Doubtful', 'Unsuitable' or

'Missing' depending on the purposes, i.e. the required selectivity. Those tests can be combined to create new ones.

Selection of basic tests

- I 9.1: *Coverage* – Did the selected and applied tests cover all plausible 'Doubtful' behaviour of the corrected data?
- I 9.2: *Update* – Update, add but never withdraw tests from the list. 'Doubtful' behaviour can occur for numerous reasons. The list proposed (Table 9.3) is rather extended but is not exhaustive. Experiences will lead to the creation of new site- and sensor-specific tests.
- I 9.3: *Thresholds* – With the given warning at the beginning of this section, the authors advise a sensitivity analysis is conducted on the selected thresholds to ensure robustness in data quality assessment.
- I 9.4: *Redundancy* – Be aware that a few tests can identify data as 'Unsuitable' (U) or 'Doubtful' (D) several times but for the same reasons. Please check on this possibility and pay careful attention during the concatenation (see Section 9.5).

9.4 APPLIED CLASSICAL METHODS

This section presents some more advanced methods than Section 9.3. Those methods still aim at flagging a data point according to different tests. The main difference to the previous section comes from the complexity of those methods, either to perform the test or to interpret the result that leads to a certain flag. The methods presented hereafter remain strongly recommended, but they require some mathematical skills and, more importantly, a close evaluation of the results of those tests. At the same time, the examples given and the codes supplied are to be regarded as material to be used for illustration, and/or educational purposes as 'real world' applications are far more complicated and more advanced. As data validation is an activity in many fields of (scientific and industrial) application, the development of mathematical methods and their implementation into software is a field of science in itself. The interested reader is referred to the vast library of literature on the subject (e.g. Hamilton, 1994).

As with most classical statistically based methods, time series analysis implicitly assumes certain characteristics of the time series (stationarity, absence of autocorrelation and certain assumptions regarding the distribution of the data). These requirements are normally not (all) met in time series obtained in UDSM monitoring. In addition it is not always straightforward to manipulate the series in such a manner that the methods become applicable in a strict mathematical sense. Nevertheless, when taking into account these limitations, a naïve application of these methods can be effective when the results are used as a pre-filter for reducing the workload that comes with visual inspection by an expert. The latter aspect may be relaxed in the future by application of pattern recognition technology or other implementations of machine learning (ML).

Hereafter the detection of outliers, step and linear trends is described in a simple manner, just to illustrate the basics of these methods and their application. In practice much more complicated algorithms need to be applied but discussing them in detail is outside of the scope of this section.

Table 9.3 Possible basic tests for data validation. Output: G, Good, D, Doubtful, U, Unsuitable and M, Missing.

Category (Section)	Name (Section)	Output	Advantages	Disadvantages
Plausibility (9.3.1)	Physical range (9.3.1.1)	G U	Easy to set up Based on common sense	Basic
	Measuring range (9.3.1.2)	G D	Easy to set up Limits the sensor choice	Requires overlapping ranges
	Calibration range (9.3.1.3)	G D	Easy to set up	Depends on calibration standard ranges
	Expertise range (9.3.1.4)	G D	Easy to set up	Requires some expertise, site and sensor specific
	Gradient range (9.3.1.5)	G D	Easy to set up	Sensitive to noisy data and gaps
Consistency (9.3.2)	Redundancy (9.3.2.1)	G U	Easy to set up	Requires three measurements
	Dynamics (9.3.2.2)	G U	Easy to set up	Requires expertise, sensitive to special conditions
	Time step (9.3.2.3)	G U	Easy to set up	Requires time stamp
Accuracy (9.3.3)	Accuracy (9.3.3)	G U	Easy to set up once uncertainty is known	Sensitive to the threshold
Auditability (9.3.4)	Calibration (9.3.4.1)	G U	Easy to set up	Requires proper site book to record maintenance
	Last calibration (9.3.4.2)	G D U	Easy to set up	Requires proper site book and calibration data
	Maintenance (9.3.4.3)	G U	Powerful tool	Not easy to set up
Synchronicity (9.3.5)	Synchronicity (9.3.5)	G D	Easy to set up	Requires time stamp and a reference
Completeness (9.3.6)	Completeness (9.3.6)	G M	Easy to set up	'Good' data just means data point exists

A basic working sequence could be:

- Step 1: Perform basic validation methods.
- Step 2: Perform detection of outliers.
- Step 3: Detect step trends.
- Step 4: Detect linear trends.
- Step 5: Apply advanced methods.
- Step 6: Try to find the cause of data that do not pass the checks.
- Step 7: Decide what to do with discarded data points and how to proceed with data analysis.

Steps 2, 3 and 4 will be discussed in some detail while the more advanced methods are discussed in a more superficial manner, with reference to the emerging literature on e.g. ML techniques.

9.4.1 Detection of outliers

Outliers are data points that deviate significantly from the data points in their close vicinity (in time or space, the discussion here is limited to the time dimension). Outliers can occur due to e.g. (i) human error, (ii) some unforeseen process, e.g. clogging of a sensor, maintenance activities interfering with a sensor functioning, or (iii) erroneous sensor readings that are not filtered through on-board processing at the sensor and/or after applying basic validation (Section 9.3). Therefore, it should be kept in mind that data points identified as 'outliers' are not necessarily incorrect (e.g. when situation *ii* has occurred).

Outlier detection has become almost a science in itself. Many methods have been developed and applied in a wide range of application fields, and depending on system and signal characteristics, one or another may be preferential. However, identical methods utilized in different application fields are likely to be parameterized differently. Given our application field, UDSM, chosen methods and thresholds, confidence levels, etc. may have to be made adaptable between, e.g., storm and dry weather conditions.

In UDSM the main cause of outliers is likely to be found in malfunctioning equipment, so methods selected for outlier detection are logically chosen to 'catch' specific behaviour with this type of cause.

The interested reader is referred to literature, e.g. Barnett & Lewis (1996) provide a very comprehensive book on outliers in data sets and methods for detecting them, and Iglewicz & Hoaglin (1993) provide an exhaustive text on the fundamentals and application of a wide range of techniques for outlier detection. Here the discussion is limited to only a few of them that are found to be useful for time series in the UDSM context. As stated before, for most of these techniques some implicit assumptions are made:

- The data are equidistant in time (hence the importance of validating time stamps and synchronicity).
- There are no data gaps (no missing data).
- Time series are (piecewise) stationary.
- Uncertainties are assumed to be normally distributed.

A logical next step to take, once an outlier is detected, is to decide to either remove, correct or keep the data point. Simply removing the data point may hamper the application of analysing tools, correcting implies the need to 'make up' information while keeping it implies knowingly using wrong data. The following steps are distinguished:

- Detection of outliers (discussed hereafter).
- Deciding what to do with them (Section 9.7).
- Data curation (Section 9.7).

9.4.1.1 Z-test for outliers

A first, very basic and simple to implement test for the presence of outliers is the so-called Z-test, in which for each data point V_t a value is added according to Equation (9.11):

$$Z_t = \frac{V_t - \bar{V}}{s} \tag{9.11}$$

where $\bar{V}$ and s are, respectively, the mean value and the standard deviation of a shifting time window $w(V_t, \ldots, V_{t+n\Delta t})$.

This indicates the quotient between the absolute difference between the recorded value V_t and the average value in a time window w covering a shifting subset of the time series $(V_t, \ldots, V_{t+n\Delta t})$, and the corresponding standard deviation. The window size to be applied depends on the system characteristics and needs to be individually estimated. The test statistics of the Z-test are defined for the hypothesis H_0

(no outliers in the data set) against H_1 (at least one outlier is present). A comprehensive text on hypothesis testing is given in Wilcox (2016).

In the Z-test, a Gaussian probability distribution is implicitly assumed, which implies that when $Z_t > 2.5$ there is a $< 1\%$ probability that the corresponding reading V_t is not an outlier. Choosing an adequate threshold is a matter of preference. When setting the threshold very low (e.g. 0.5), this will result in many 'false alarms' leading to an increased effort ('manual labour') to decide whether to keep the data entry or not. On the other hand, when setting the threshold too high (e.g. >4), the risk of missing outliers increases, increasing the risk of obtaining incorrect information.

Another practical issue is that when using small time series (or a small shifting window), the Z value obtained can be misleading as the maximum value is limited to $(n-1)/n^{0.5}$, e.g. for $n = 10$ the maximum value is 2.84. When a threshold of 3 is applied, no outlier will be detected. Therefore, one is well-advised to carry out some test runs on available data sets and evaluate the effectiveness of a chosen threshold. In this process, information is obtained on the amount of time and means needed to manually process the indicated outliers against the improvement of the information which will be obtained.

9.4.1.2 Grubbs test

In the two-sided Grubbs test (Grubbs, 1969), the underlying hypothesis is the same as for the Z-test, the test value G is defined by Equation (9.12):

$$G = \frac{\max|V_t - \bar{V}|}{s} \tag{9.12}$$

Testing whether the minimum is an outlier is tested by Equation (9.13):

$$G = \frac{\bar{V} - V_{min}}{s} \tag{9.13}$$

And, correspondingly, for the maximum value (Equation (9.14)):

$$G = \frac{V_{max} - \bar{V}}{s} \tag{9.14}$$

At a significance level α the test statistics (the hypothesis H_0, i.e. no outlier) is rejected if:

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{\left[t_{\frac{\alpha}{2n}}, n-2\right]^2}{n-2+\left[t_{\frac{\alpha}{2n}}, n-2\right]^2}} \tag{9.15}$$

in which $t_{\alpha/2n,n-2}$ is the critical value of the Student t distribution with $n-2$ degrees of freedom. For the one-sided test (for maximum and minimum values) the significance level is α/n – functions for Student t values in Microsoft Excel® and Matlab® are given in Table 8.2).

9.4.1.3 Cook's distance

Another, often applied, metric to decide whether or not a specific measuring result is an outlier is known as the Cook's distance. Basically, the Cook's distance (Equation (9.16)) is a metric for the 'influence' an individual data point has on the 'fit' of a regression model for the time series of (monitoring) data. The

test is based on the linear regression (see also Section 9.4.2.1):

$$D_j = \frac{\sum_{i=1}^{i=n} (\hat{y}_i - \hat{y}_{i(j)})^2}{MSE} \tag{9.16}$$

The subscript $i(j)$ implies that when $i = j$, the element is omitted. MSE is defined by Equation (9.17):

$$MSE = \sum_{i=1}^{i=n} (\hat{y}_i - y_i)^2/(n-1) \tag{9.17}$$

A generally accepted criterion for detecting outliers is that an individual observation has a Cook's distance $D_i > 3\bar{D}$ with $\bar{D}$ the mean value for D_j $(j = 1{:}n)$, with n the number of observations in the time series.

9.4.1.4 Example of Cook's distance, Z and Grubbs tests

An example of a hydrograph is shown in Figure 9.12, where 10 apparent outliers are introduced. To illustrate the effect of choosing thresholds in the Z-test, the Grubbs test and the Cook's distance, this hydrograph is used as a test. Figure 9.13 shows the number of outliers detected using the Z-test as a function of the Z_{max} value.

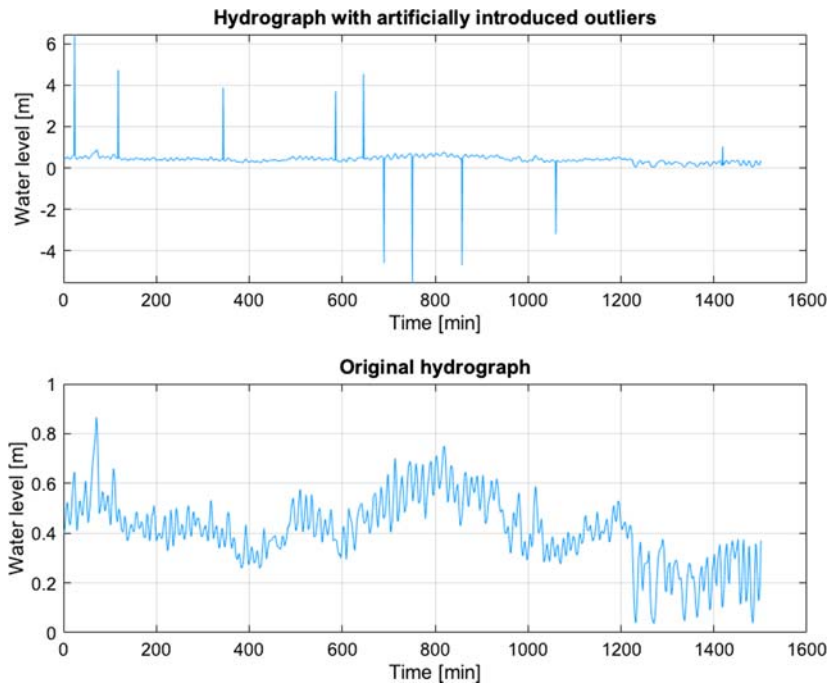


Figure 9.12 Example of a hydrograph with 10 artificial outliers. *Source*: Francois Clemens-Meyer (Deltares/TU Delft/NTNU).

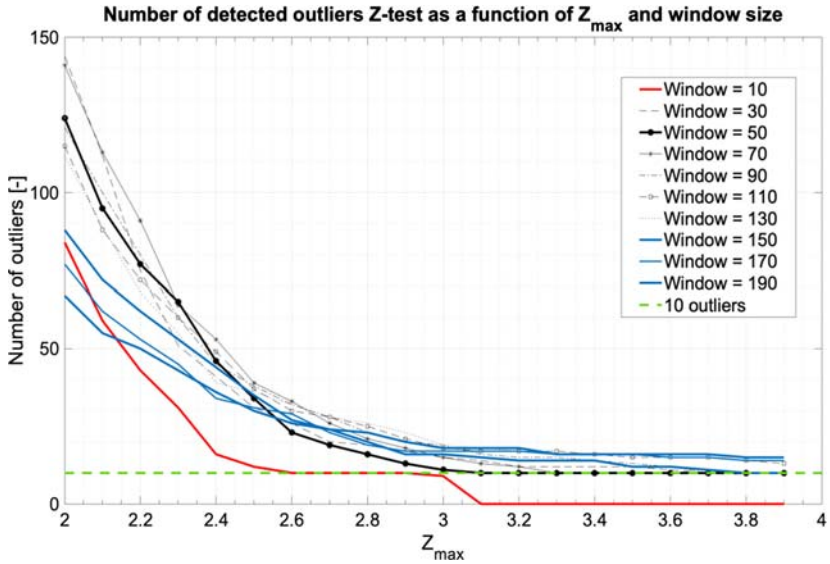


Figure 9.13 Number of outliers detected by the Z-test as a function of window size and Z_{max}. *Source*: Francois Clemens-Meyer (Deltares/TU Delft/NTNU).

As can be seen in Figure 9.13, the correct result (10 outliers) is only achieved for a limited number of combinations of window size and Z_{max}, a window size of 50 minutes seems to be a robust choice, as the outcome is constant for Z_{max} values >3.1. On the other hand, for window sizes >110 the number of detected outliers seems to be too high regardless of the Z_{max} value chosen.

Figures 9.14, 9.15 and 9.16 show some detailed results from some combinations of window size and Z_{max}.

It has to be emphasized that these graphs are only valid for the examples shown. The settings of the test parameters depend on the signal used, the defined rigidness in terms of false positives and false negatives,

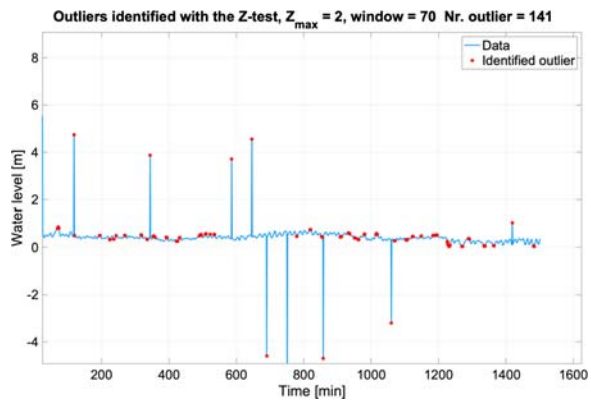


Figure 9.14 Detected outliers for window size = 70 and $Z_{max} = 2.0$. *Source*: Francois Clemens-Meyer (Deltares/TU Delft/NTNU).

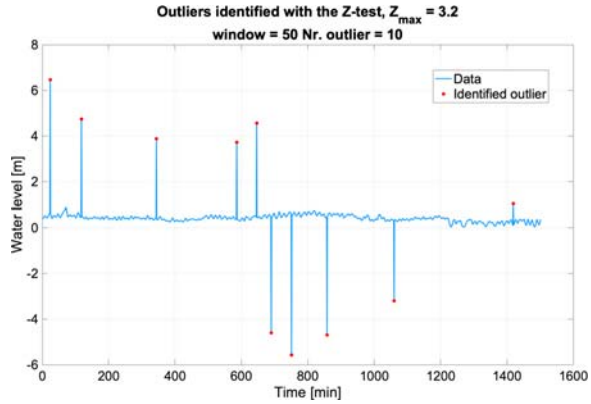


Figure 9.15 Detected outliers for window size = 50 and $Z_{max} = 3.2$. *Source*: Francois Clemens-Meyer (Deltares/TU Delft/NTNU).

and the subjective preferences of the user of the data. Nevertheless, as a rule of thumb, a Z_{max} value of 2.5 seems a good starting point, with respect to window size and the range of values between the relevant characteristic timescales of the processes (Section 6.3).

Figure 9.16 shows the performance of the Grubbs test on the hydrograph shown in Figure 9.12. The Grubbs test performs well for this example for a window size of 70 and for a range of significance levels ($\alpha > 0.015$). The test is sensitive to the window size, as for a window size of 90, the test overpredicts for any confidence level. Figure 9.17 shows some detailed results for the Grubbs test.

Finally, the hydrograph shown in Figure 9.12 was subjected to the Cook's distance test. Figure 9.18 shows the dependency of the result on the threshold. The Cook's distance test proves to produce reliable results for a wide range of threshold values. Threshold values between 2 to 7 times the mean values result in the correct identification of all outliers. Figure 9.19 shows an example of the detailed results of the Cook's distance test.

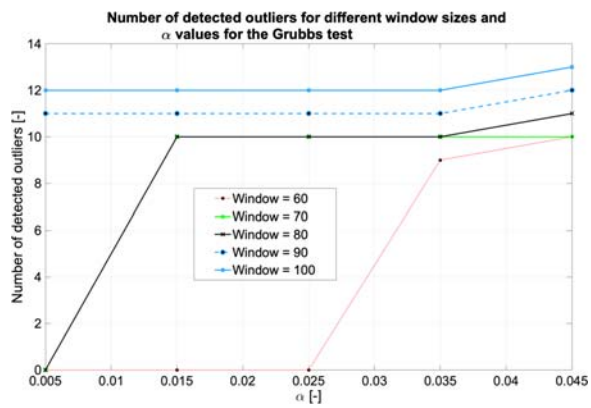


Figure 9.16 Performance of the Grubbs test as a function of window size and significance. *Source*: Francois Clemens-Meyer (Deltares/TU Delft/NTNU).

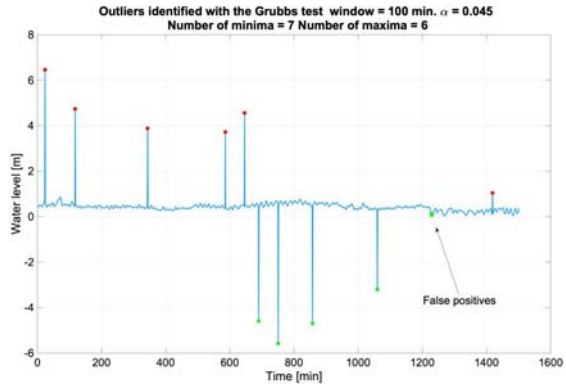


Figure 9.17 Example of the introduction of 3 local minima (very close together) falsely identified as outliers by the Grubbs test. *Source*: Francois Clemens-Meyer (Deltares/TU Delft/NTNU).

9.4.2 Detecting trends and sensor drifts

A trend in a time series is basically a variation of the process mean values in time and/or space.

A linear trend in a time series may point to either, (i) a change in the process under study, or (ii) zero drift of a sensor. Step trends (a sudden change in the mean value) may hint at (i) a change in reference level of a sensor (e.g. due to a wrong reinstallation after maintenance), or (ii) a change in the system studied (e.g. a sudden blockage of a conduit in a sewer system due to collapse or a closure during construction activities).

Methods to detect such trends are numerous (e.g. Gray, 2007). It is noted however that the detectability of (linear) trends depends on the variability of the process monitored, the uncertainty in the measuring system applied and the sampling frequency.

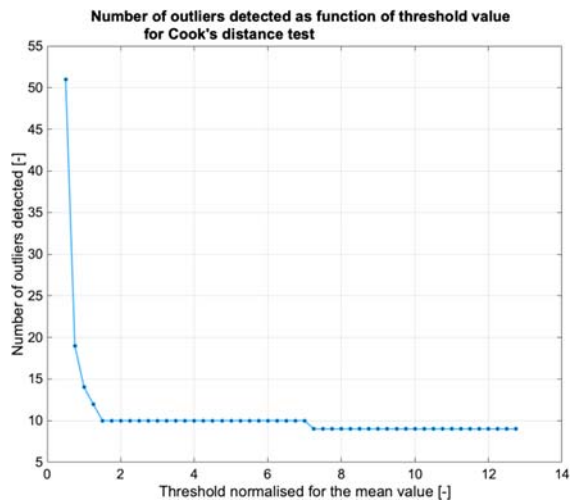


Figure 9.18 Performance of Cook's distance test. *Source*: Francois Clemens-Meyer (Deltares/TU Delft/NTNU).

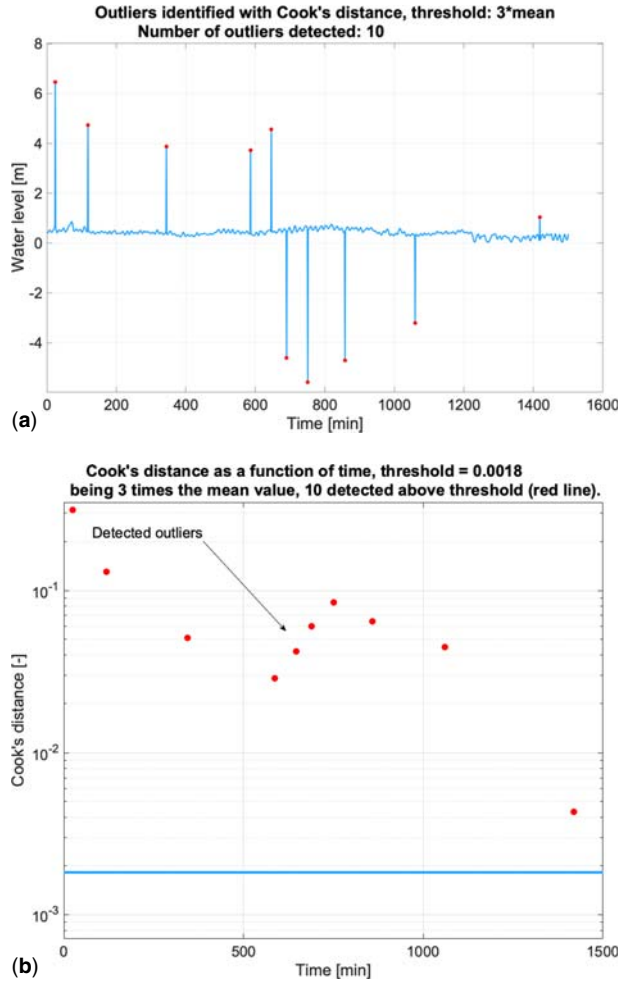


Figure 9.19 Results of Cook's distance test for the example hydrograph in Figure 9.12. (a) hydrograph with detected outliers; (b) values of Cook's distance for each individual data point in the series. *Source*: Francois Clemens-Meyer (Deltares/TU Delft/NTNU).

Lettenmaier (1976) approaches trend detectability in terms of statistical tests, with a null hypothesis H_0 stating 'no trend is present' against the hypothesis H_1 'a trend did occur'. Based on the data, either one of these hypotheses is rejected or accepted.

In statistical testing on accepting or rejecting a hypothesis, the relation between type I and type II errors and confidence and power are shown in Table 9.4. Emmert-Streib & Dehmer (2019) provide a review on hypothesis testing in general along with methods applied.

Obviously, the values for both α and β should be as small as feasible. As these values, apart from choices made like measuring frequency and uncertainty in the measured values, depend largely on the process studied, the settings chosen for acceptance limits of the test outcome (like a maximum value for α) can therefore never be regarded as generic.

Table 9.4 Accepting or rejecting a hypothesis.

	Test indication H_0	**Test indication H_1**
'Real' state H_0	Confidence $= (1-\alpha)$	Type I error $p = \alpha$
'Real' state H_1	Type II error $p = \beta$	Power $= (1-\beta)$

The discussion here is limited to two types of trends that occur frequently in UDSM, namely:

- The step trend (typically occurring after e.g. misplacing a water level sensor after maintenance).
- A linear trend, often attributable to sensor drift.

Most algorithms for trend detection are sensitive to the presence of outliers, therefore one is well advised to first analyse for outliers prior to analysing for the presence of trends.

9.4.2.1 Linear regression

A very simple method to detect a linear trend is by fitting a linear function through the measuring data using the relation: $y = ax + b$.

The values of a and b can be obtained using the method of maximum likelihood estimates, which, when assuming a Gaussian distribution for the residues, boils down to the classical ordinary least squares method.

The least squares estimators for a and b are:

$$\left.\begin{aligned} \hat{a} &= \frac{\sum_{i=1}^{i=n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{i=n}(x_i-\bar{x})^2} \\ \hat{b} &= \bar{y}-\hat{a}\bar{x} \end{aligned}\right\} \tag{9.18}$$

Giving the 'fitted' relation:

$$\hat{y} = \hat{a}x + \hat{b} \tag{9.19}$$

the residuals (difference between measured values and fitted results) are defined as:

$$r = y - \hat{y} \tag{9.20}$$

Assuming that the variance of the residuals is constant, their variance is estimated by:

$$\sigma_r^2 = \frac{\sum_{i=1}^{i=n} r_i^2}{n-2} \tag{9.21}$$

The standard deviations in the estimated parameter values follow from Equation (9.21) (neglecting covariance terms for the sake of simplicity – see Section 7.6.4.2 for more detail):

$$\left.\begin{aligned} \sigma_{\hat{a}} &= \sigma_r\sqrt{\frac{1}{\sum_{i=1}^{i=n}(x_i-\bar{x})^2}} \\ \sigma_{\hat{b}} &= \sigma_r\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{i=n}(x_i-\bar{x})^2}} \end{aligned}\right\} \tag{9.22}$$

When $|\hat{a}| > 0$, this may indicate there is a linear trend in the time interval studied. Or, more formally, the hypothesis tested is 'H_0: there is no linear trend present' against 'H_1: there is a linear trend present'. This essentially boils down to deciding whether or not the value of $|\hat{a}|$ deviates significantly from zero. For this, the Spearman's ρ test is used, as will be discussed later.

9.4.2.2 Relation between trend, window length, process characteristics, confidence and power of trend detection tests

Conover (1999) identifies the Mann-Whitney test and the Spearman's ρ test as best suited for the trends mentioned. Lettenmaier (1976) has shown that the detectability of a trend depends on the following parameters for:

- Detecting a step trend (Equation (9.23)):

$$\psi(\Delta t, T, T_r) = \frac{T_r}{2\sigma_x}\sqrt{N_{equi}(\Delta t, T)} \tag{9.23}$$

- Detecting a linear trend (Equation (9.24)):

$$\psi(\Delta t, T, T_r) = \frac{T_r}{N_{equi}(\Delta t, T)\sqrt{12}\sigma_x}\sqrt{N_{equi}(\Delta t, T)[N_{equi}+1][N_{equi}-1]} \tag{9.24}$$

The value N_{equi} is the equivalent number of samples in a series of N data points corrected for the mutual correlation between these N points (implying these hold redundant information).

N_{equi} depends on the autocorrelation function of the time series under study, as defined by Equation (9.25) (Bayley & Hammersley, 1946):

$$N_{equi}(\Delta t, T) = \frac{N(\Delta t, T)}{1 + 2\sum_{i=1}^{i=N(\Delta t,T)}\left[\left(1 - \frac{i}{N(\Delta t, T)}\right)\rho_x(i\Delta t)\right]} \tag{9.25}$$

in which $\rho_x(i\Delta t)$ is defined as:

$$\rho_x(i\Delta t) = \frac{\gamma_x(i\Delta t)}{\sigma_p^2 + \sigma_m^2} \tag{9.26}$$

and $\gamma_x(i\Delta t)$ is defined by:

$$\gamma_x(i\Delta t) = \begin{cases} \sigma_p^2 + \sigma_m^2 & i = 1 \\ \sigma_p^2\rho_p(i\Delta t) & i > 1 \end{cases} \tag{9.27}$$

When a time series is absolutely uncorrelated (i.e. a random sequence), then $N_{equi}(\Delta t, T)$ is equal to N. When all points are 100% correlated, then $N_{equi}(\Delta t, T)$ is equal to 1 (the first value in the series is the perfect predictor for all the rest). The power and confidence of the tests are related as Equation (9.28):

$$1 - \beta = \phi\left[\psi(\Delta t, T, T_r) - \xi\left(\frac{\alpha}{2}\right)\right] \tag{9.28}$$

This implies that, given a value for α (type I error, or significance), a time series and the measuring uncertainty, the limits of detecting a step and/or linear trend are defined when choosing a certain window size $N(\Delta t, T)$ in terms of confidence and power of the test.

Note that in the obtained values for power and confidence of the tests, the effects of (auto) correlation and measuring uncertainty are accounted for. This allows for determining the characteristics of trends that are detectable given the measuring frequency, measuring uncertainty and window length.

Figure 9.20 shows some results of the application of the relations between the variables in Equations (9.23–9.28) (Matlab instructions can be found in `lin_step_power.m`). When striving for equal probability for type I and type II errors, one would like to have equal levels for confidence and power of a test. As can be seen in Figure 9.20 for the linear trend in the example time series, this is not a feasible option within the range of window sizes and T_r/σ_p values. For the step trend, however, it is feasible for all window sizes provided $T_r/\sigma_p > 2$. The 'optimum' window size for both step and linear trend is approx. 250 minutes, as it produces the highest values for the power over the whole range of T_r/σ_p. So, in this case, one would start with analysing the series with a window size of 250 min. Further it has to be emphasized that the detectability of trends is largely decided upon in the macro design of the monitoring network (Section 6.2).

9.4.2.3 Mann-Whitney test for step trend detection

The Mann-Whitney U test (or Wilcoxon rank-sum test) basically tests whether there is a difference in level (median) of two partitions y and z ($y = (x_1, x_2, \ldots, x_m)$, $z = (x_{m+1}, \ldots, x_n)$) of a vector (i.e. a time series) $X = (x_1, x_2, \ldots, x_n)$. The hypothesis H_0 is that $p(y < z) = 0.5$ (no step trend) against H_1 $p(y < z) < 0.5$ (a step trend is present and y has a lower overall value than z) and H_2 $p(y < z) > 0.5$ (a step trend is present

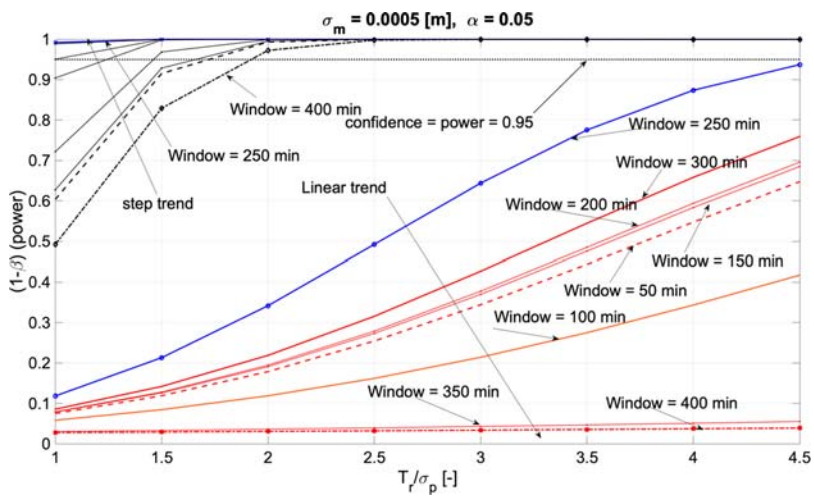


Figure 9.20 Results for the relation between confidence and power for the example hydrograph using different sizes of shifting windows and trend ratios. *Source*: Francois Clemens-Meyer (Deltares/TU Delft/NTNU).

and y has a higher overall value than z). The test value is defined in Equation (9.29):

$$\left.\begin{aligned} z &= \frac{\sum_{i=1}^{i=n} R(y_i) - \dfrac{m(n+1)}{2} + K}{s_w} \\ \text{in which} \\ K &= \begin{cases} \sum_{i=1}^{i=n} R(y_i) - \dfrac{m(n+1)}{2} > 0.5 : K = 0.5 \\ \sum_{i=1}^{i=n} R(y_i) - \dfrac{m(n+1)}{2} = 0.5 : K = 0 \\ \sum_{i=1}^{i=n} R(y_i) - \dfrac{m(n+1)}{2} < 0.5 : : K = -0.5 \end{cases} \end{aligned}\right\} \tag{9.29}$$

where $R(y_i)$ is the rank of y_i in the vector X, and s_w is defined by Equation (9.30):

$$s_w = \sqrt{\frac{m(n-m)(n+1)}{12} - \frac{m(n-m)\sum_{i=1}^{k}(t_i^3 - t_i)}{12n(n-1)}} \tag{9.30}$$

where t_i is the number of subjects having the rank i, and k is the number of (distinct) ranks in the data. H_0 is rejected in favour of H_1 or H_2 (step trend is present) when $z > z_{1-\alpha/2}$ or $z < z_{\alpha/2}$, respectively, in which z_q is the quantile for $\alpha/2$ (for $\alpha = 0.05$, $z_{\alpha/2} = -1.96$ and $z_{1-\alpha/2} = 1.96$). Common statistical software suites provide a standard function for this test. A simple implementation is provided in the Matlab® code `step_trend.m` (available for download at https://doi.org/10.2166/9781789060102) by defining a shifting window and applying it to the time series at a chosen value for α.

9.4.2.4 Spearman's ρ test for linear trend detection

Spearman's ρ test is used to decide whether or not a detected linear trend, as described in Section 9.4.1.1, is to be regarded as significant or not. The test parameter is defined as:

$$\rho = \frac{\dfrac{1}{n}\sum_{i=1}^{i=n}\left[i - \dfrac{n+1}{2}\right][R(x_i) - \overline{R(x)}]}{\sqrt{\sum_{i=1}^{i=n}\left[i - \dfrac{n+1}{2}\right]^2 \sum_{i=1}^{i=n}[R(x_i) - \overline{R(x)}]^2}} \tag{9.31}$$

If the value of ρ is negative, the trend is descending; if ρ is positive, the trend is ascending. At the same time, the value of ρ is used as a statistical test variable, assuming a normal distribution. This implies that when choosing a p-value for the test, α, the hypothesis that no linear trend is present is rejected when either $\rho n^{0.5} < z_{\alpha/2}$ or $\rho n^{0.5} > z_{1-\alpha/2}$ ($p < \alpha/2$ or $p > 1-\alpha/2$). A simple implementation in Matlab® is `[r,p] = corr(x,y, 'Type', 'Spearman')` here 'r' is the value for ρ and p is the p-value (to be related to α in the preceding text).

9.4.2.5 Examples of trend detection

Hereafter a rather naïve working sequence will be demonstrated which is not very sophisticated but will show the limitations of the methods applied and the need for some human supervision (although self-learning software may be expected to, at least partially, take over this task, see Section 9.8).

An important fact to consider it that the shifting time window over which the trend analysis is applied is crucial in recognizing any trend. A first logical step to take is to identify time intervals that behave in a more or less similar manner, at times where changes occur. An approach in relation to the former is to find a piecewise linear fit to the original signal. A simple implementation is shown in the Matlab® codes `piece_lin_fit.m`, `step_trend.m` and `linear_trend.m` (available for download at https://doi.org/10.2166/9781789060102).

When studying the lower graph in Figure 9.21, it can be seen that the critical z values are only surpassed below the z_{min} value (with $\alpha = 0.05$, this is -1.96). The Mann-Whiney test is applied in such a manner that the z values for both hypotheses are tested (z_1 tests hypothesis H_1): this implies that when $z_1 > 1.96$, the hypothesis that no trend is present is rejected in favour of a trend in which a sudden increase occurs. Regarding z_2, the situation is likewise: when $z_2 < -1.96$, the H_0 hypothesis is rejected in favour of the presence of a sudden decrease. In the graph, there are three time windows (indicated as 'A', 'B' and 'D') and one cluster of short windows (indicated 'C') that, given the chosen α, are marked to reject the hypothesis that no trend is present. As can be seen, the example used is composed of a range of (linear and step) trends. As we are 'hunting' for step trends using the Mann-Whitney test, we focus on the time windows indicated. In window A, it is clear from visual inspection that this would qualify a linear trend, indeed the test outcomes are not very convincing, so possibly choosing a somewhat smaller value for α would have eliminated this candidate. Window B shows that here the discrimination between step trend and linear trend seems to function well, which cannot be said for window D. Indeed, from visual inspection, it is clear that the signal is more or less constant, which is reflected in a somewhat ambiguous test result. In Figure 9.13 the maximum relative trend size as a function of window size is shown. As can be seen, the maximum relative trend that occurs in the example hydrograph is less than 2, and for most window sizes approx. 1.6–1.7. From Figure 9.20, it is concluded that for this range of trends, the power of the tests in not very high (not for step trends and certainly not for linear trends), which hints at ambiguous test results (Figure 9.22).

With respect to the cluster of windows C (Figure 9.21) overall, apart from the sub window at the far right, there is clear evidence that a sudden decrease in the water level occurs. Again, here a smaller value for α

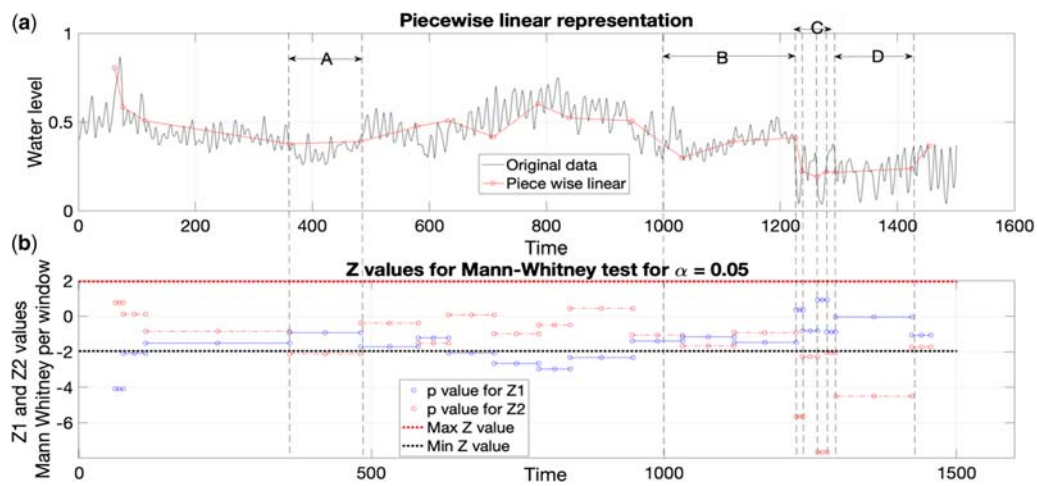


Figure 9.21 A piecewise linear representation of the original signal. (a) outliers removed; (b) results of the Mann-Whitney test. *Source*: Francois Clemens-Meyer (Deltares/TU Delft/NTNU).

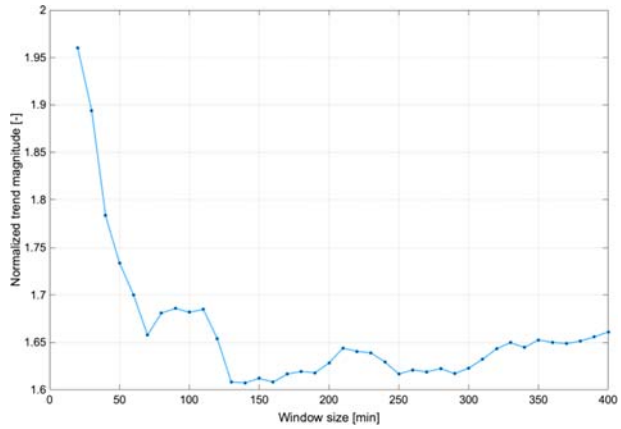


Figure 9.22 Maximum relative trend as a function of window size. *Source*: Francois Clemens-Meyer (Deltares/TU Delft/NTNU).

would have eliminated this candidate. After eliminating A, D and the far-right sub window C, a step trend in the time window between $t = 1226$ and $t = 1294$ minutes is recognized. In this respect one has to realize that when windows become small (i.e. <20 observations), implicit assumptions (most important normality) underlying many of the statistical methods applied are no longer valid. Filtering on window size prior to conducting the analysis is thus strongly advised.

After identifying the step trend, linear trends can be detected using the Spearman's ρ test applied on the piecewise linearization. In Figure 9.23, the results of the Spearman's ρ test are shown. It can be seen that at a significance level of 0.05, six time windows (noted A-F) are identified to contain a significant linear trend, the sign of the ρ values indeed corresponds with the visually observed trends (either ascending or descending). With respect to window F, it contains less than 20 data points and is therefore to be treated as an artefact of the test result. Notice that the time window B in Figure 9.23 corresponds with the time window B in Figure 9.21, which implies that both the Mann-Whitney test and the Spearman's ρ test detect, respectively, a step and a linear trend that has statistical significance. After visual inspection, one has to conclude that a linear trend is more obvious than a step trend in this case. In spite of the strictly taken non-compliance of the underlying data with the pre-assumptions set for the statistical test presented, these tests prove to be of value when validating time series, even though some manual inspection is needed as demonstrated in the examples.

9.4.3 Detecting abnormal processes

9.4.3.1 Using spline function

Villez & Habermacher (2016) propose a method to detect anomalies in processes. The methodology is based on *shaped-constrained splines* and is applicable for any univariate or multivariate time series. Without entering into all the details of this method, which is rather more complex than the previous ones, the overall idea is to identify abnormal trends or behaviours in time series, while fitting spline functions into different parts of time series.

As a basic example, let us look into the evolution of water levels, velocities and discharges in a sewer pipe once a rainfall peak is passed. Those three values are supposed to decrease while following a convex shape.

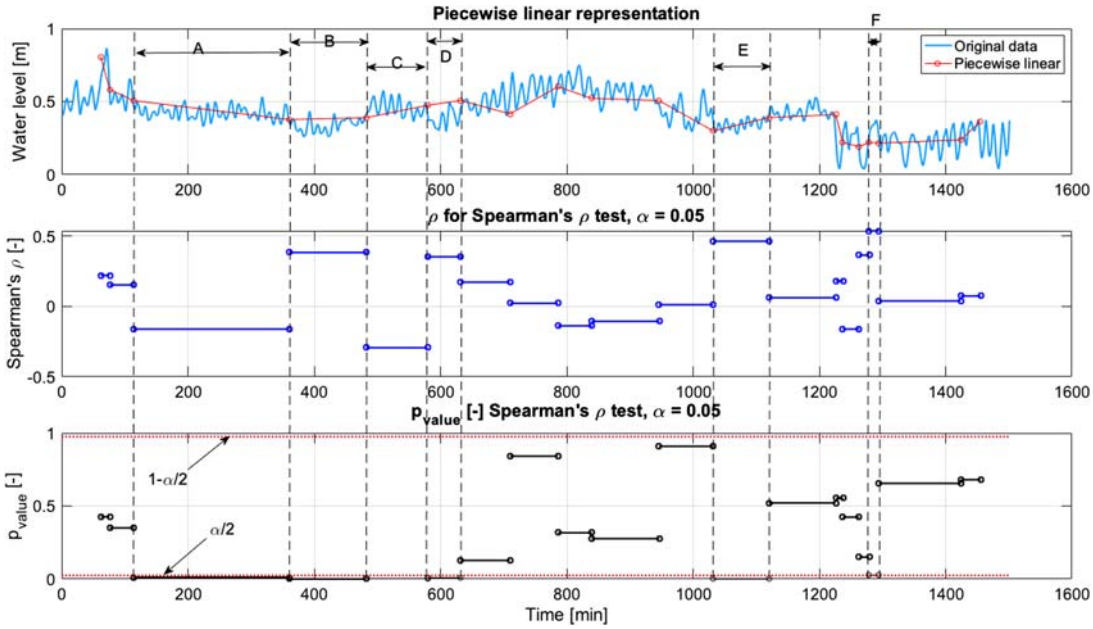


Figure 9.23 Results for Spearman's ρ test. *Source*: Francois Clemens-Meyer (Deltares/TU Delft/NTNU).

If, for a given period, one time series show a concave shape, the values obtained over this period should be considered as 'Doubtful' or 'Unsuitable'.

9.4.3.2 Detecting abnormal episodes based on conditional dependencies

Conditional dependencies identify if inconsistencies in data emerge as violations of these dependencies. That is, for instance, including topological information on a flow path network in the form of rules and using partially redundant information from up- or downstream located sensors. This allows detection of abnormal measurements (Figure 9.24).

In Figure 9.25, sensors F03 and F06 are installed at the same location. Sensor F04 is located 1 km downstream of F03 and F06. F04 should always show higher flows than F03 and F06. This can be questionable to a certain extent for dry weather periods. The two labelled anomalies in F04 appear questionable. A cross-comparison with correctly aligned adjacent measurements leads to the conclusion that the early anomaly obviously occurs due to hydraulic disturbances, while the later anomaly is obviously a non-natural artefact.

9.4.3.3 Detecting abnormal episodes based on the hydraulic gradient h_c

The consistency of flow observations can be verified by means of the hydraulic gradient h_c based on the Manning-Strickler relation. This method can be applied for flow ranges in which no disturbance due to minimal water levels and/or backwater effects is expected. The hydraulic gradient h_c is calculated according to Equations (9.32) and (9.33).

$$h_c = k_{st} * \sqrt{I} = \frac{Q}{(A * {R_{hyd}}^{2/3})} \tag{9.32}$$

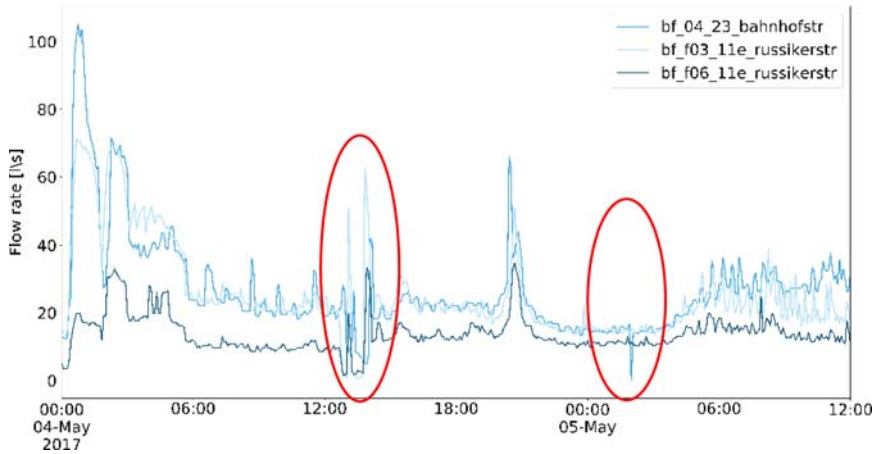


Figure 9.24 Example of the cross-comparison of hydrographs from three different sensors in the same drainage network. *Source*: Frank Blumensaat and Andy Disch (Eawag).

with

$$R_{hyd} = \frac{A}{l_u} \tag{9.33}$$

h_c is calculated for each recording, i.e. each time step. Subsequently a monthly/weekly average is determined. Inconsistencies can be detected by deviations from the mean value over the course of time.

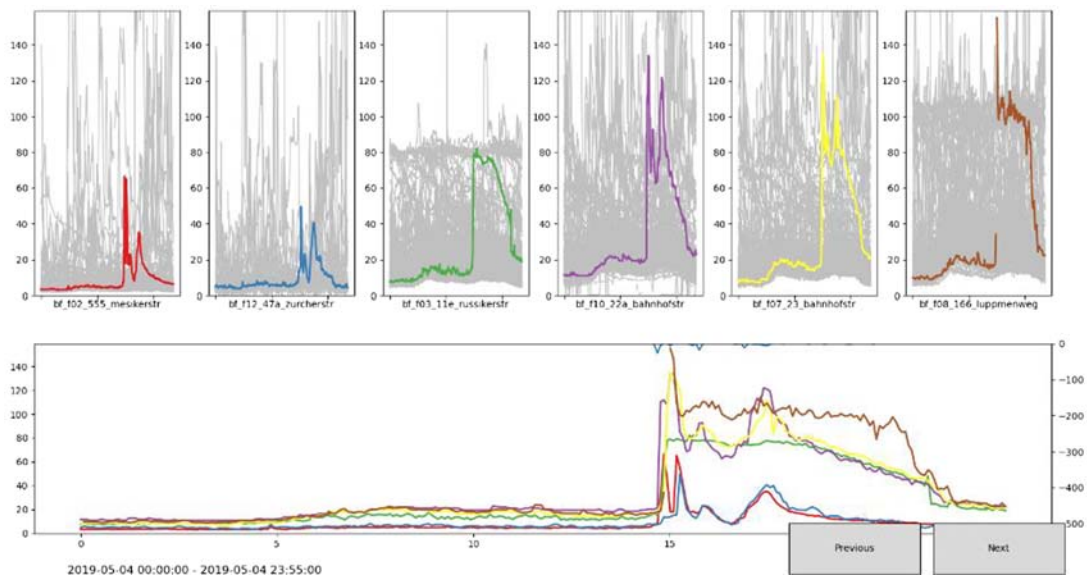


Figure 9.25 Example of multi-signal cross-comparison (here on a daily basis). Rules apply depending on the topological relation on the flow path network. *Source*: Frank Blumensaat and Andy Disch (Eawag).

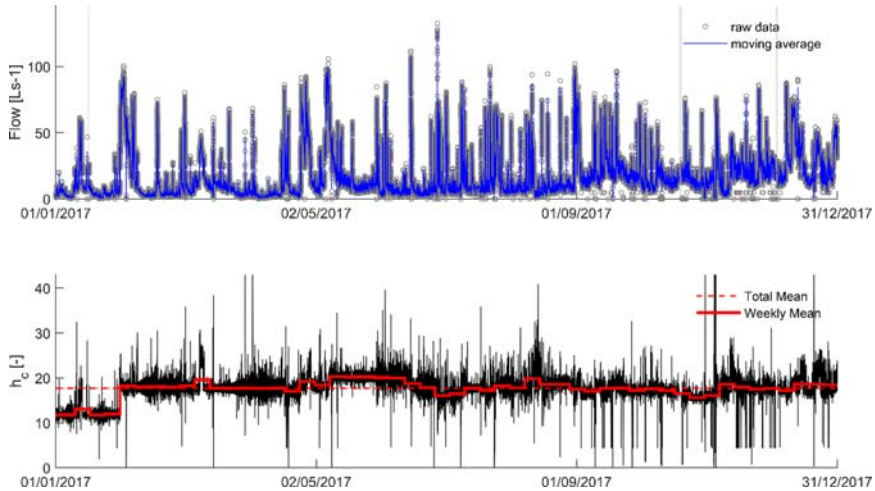


Figure 9.26 Hydrograph obtained through a wedge sensor that measures flow velocity (US backscatter and cross-correlation) and water level (pressure gauge). *Source*: Frank Blumensaat (Eawag).

Weeks/months for which h_c values considerably deviate from the mean are to be questioned; periods should be excluded from further usage, e.g. for calibration, or should be further analysed. The application of the hydraulic gradient test is exemplified in Figure 9.26. Here, in February 2017, the sensor was cleaned and re-configured without showing an impact on the resulting flow signal. The jump in the h_c value, however, reveals the hidden anomaly due to sensor maintenance (Figure 9.26).

9.4.3.4 Detecting abnormal flow conditions based on the Q(h) relation

Flow observations should be checked for plausibility by comparing measured data with the theoretical $Q(h)$ relationship, i.e. with the part-full circular pipe flow curve (Figure 9.27). For this purpose, the theoretical $Q(h)$ relationship according to Manning-Strickler (Equations (9.34) to (9.36)) is calculated for a given pipe diameter D.

$$Q = k_{st} \times A \times R_{hyd}^{\frac{2}{3}} \times \sqrt{I} \tag{9.34}$$

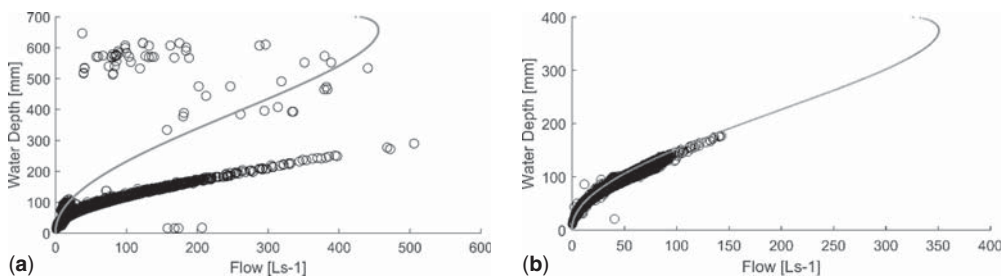


Figure 9.27 $Q(h)$ relationships plotted against the theoretical Manning-Strickler relation (parable test). Examples for a poor (a) diameter 700 mm and good (b) diameter 400 mm matches. *Source*: Frank Blumensaat (Eawag).

with

$$A = D^2 \frac{4 \times \arcsin\left(\sqrt{\frac{h}{D}}\right) - \sin\left(4 \times \arcsin\left(\sqrt{\frac{h}{D}}\right)\right)}{8} \tag{9.35}$$

$$l_u = 2 \times D \times \arcsin\left(\sqrt{\frac{h}{D}}\right) \tag{9.36}$$

9.4.4 Validation between correlated monitoring points (time series, ARMA models)

In many cases, time series obtained from two or more measuring devices in an urban drainage system show a mutual correlation structure. For example, water level sensors in a wastewater system will reflect more or less the same daily pattern in the recorded water levels, or rain measurements with discharge time series. In many cases an auto regressive moving average (ARMA) model can be used to obtain a description of a time series in the form of a polynomial function, which, to a certain extent, can provide a basic tool for forecasting or (re)constructing a missing value in the time series (see also Section 9.7 on data curation). The theory of ARMA models and applications is comprehensively discussed in e.g. Choi (1992).

The general equation for an ARMA(p,q) model is given in Equation (9.37):

$$V_t = c + \varepsilon_t + \sum_{i=1}^{i=p} \gamma_i V_{t-i} + \sum_{j=1}^{j=q} \theta_i \varepsilon_{t-j} \tag{9.37}$$

The first summation over p represents the auto regressive (AR) part while the second summation over q represents the moving average (MA) part of the model. c is a constant, ε_i represents white noise in step i and V_i is the dependent variable at step i, γ_i and θ_i are the polynomial coefficients for, respectively, the AR and MA parts of the model. The coefficients of the model, for given values of p and q, can be found by e.g. using the maximum likelihood estimate method. In most software packages like Matlab®, Python®, R®, etc. fast routines for ARMA fitting are available as standard.

There is no general manner or protocol for determining the values for p and q. A first requisite is to achieve (piecewise) stationarity of the time series. Using transformation techniques, stationarity of the series can be achieved. Apart from removal of trends and/or periodic signals in the series, differentiating is a popular and effective manner to achieve stationarity. Transforming the time series into a series of incremental differences between successive parameter values often achieves the sought after stationarity. The goodness of fit between the ARMA model and the original data can be expressed in a range of metrics, of which the Akaike information criterium (AIC) is the most commonly applied in this context as it not only takes into account the 'goodness of fit' of the model, but also penalizes for overfitting.

AIC is defined as:

$$\mathrm{AIC} = \ln(\sigma_r^2) + \frac{2(p+q)}{n} \tag{9.38}$$

where n is number of elements in the time series and σ_r is the standard deviation of the residues.

A simple stepwise approach is as follows:

(1) Plot the autocorrelation function (ACF), and the autocorrelation function of the differentiated series (DACF) (Matlab® commands: `autocorr(x1)` and `autocorr(diff(x1))` respectively, where the differentiated time series of x1 is defined as $x_{d,i}, = x_{i+1} - x_i$).
(2) A first estimate for p is indicated by the DACF, where the DACF becomes (almost) zero, defines this first estimate.
(3) A first estimate for q is obtained from the ACF where it starts to tail off to zero.
(4) Estimate the model parameters: in most standard available applications, this is normally done by the application of e.g. the maximum likelihood estimates method.

In Figures 9.28 and 9.29 an example is shown. The Matlab® script `arma_space.m` can be downloaded here https://doi.org/10.2166/9781789060102. Based on the DACF, the value for p is expected to be approx. 5 while the q value is less easy to deduce, as the ACF tails off to zero at moderate time lags. A value of $q = 1$ or 2 would be a first guess.

Using the AIC metric, an ARMA ($p = 5$, $q = 2$) model using the differentiated time series was found to produce the best fit to the data over the first 60 minutes. An ARMA model on a differentiated time series is

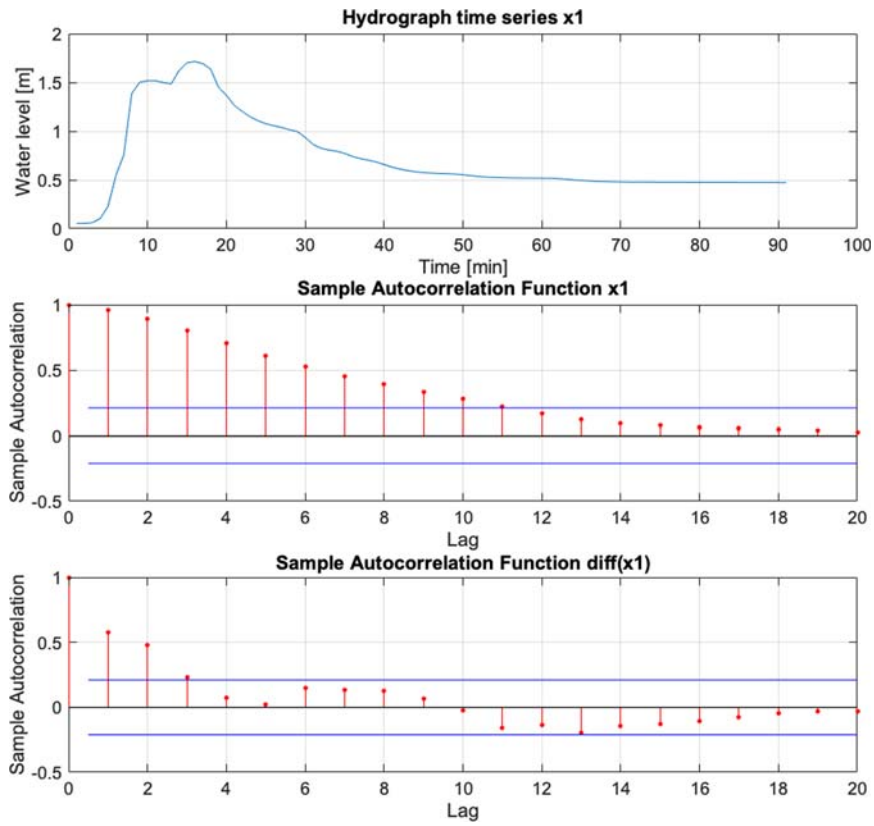


Figure 9.28 Time series and the corresponding ACF and DACF. *Source*: Francois Clemens-Meyer (Deltares/TU Delft/NTNU).

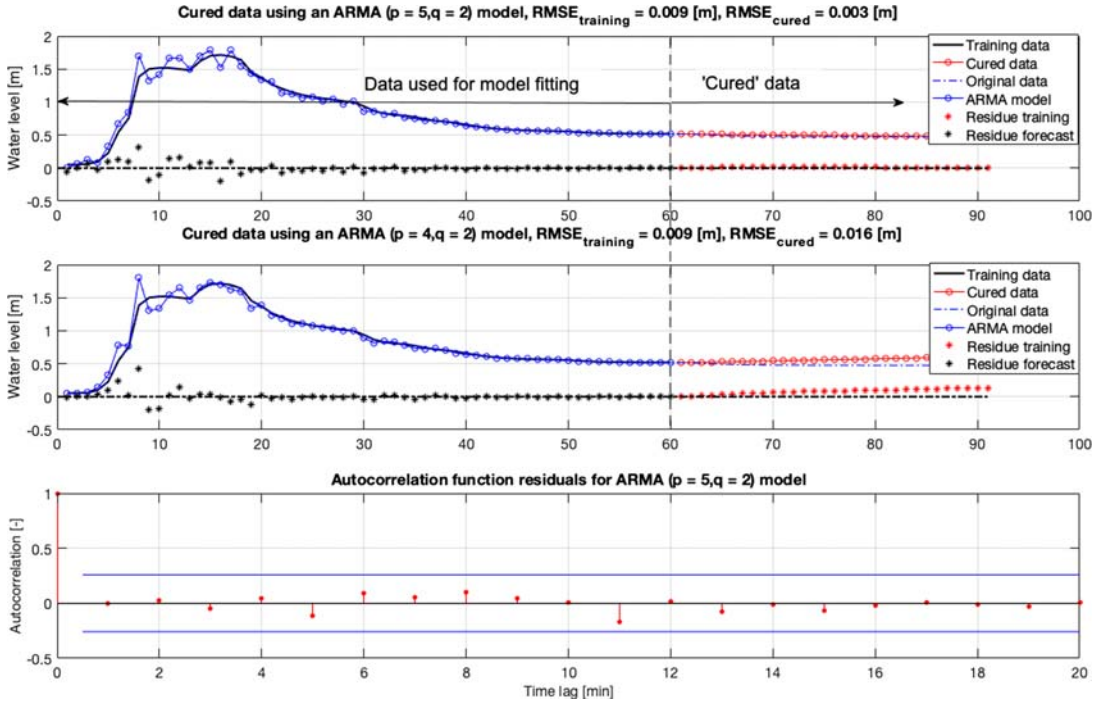


Figure 9.29 Example of the application of an ARMA model to forecast data. (ARIMA(p,d,q), in which d stands for differentiating (in this case $d = 1$: differentiating once). *Source*: Francois Clemens-Meyer (Deltares/TU Delft/NTNU).

also known as an ARIMA model (most software packages like Matlab®, R®, etc. contain standard functions for this). This model was used to forecast 30 minutes of additional data. Actually these data were measured as well, allowing for a comparison between forecasted and observed results. In Figure 9.29 the upper graph shows the results of the ARMA model that showed the best fit to the training data, which results in forecasted data with an RMSE (root mean squared error, chosen here since it offers a more 'intuitive' understanding compared to a value for the AIC, the model selection, however, has been done based on AIC) of 3×10^{-3} m. The lower graph shows the results for a slightly different ARMA model ($p = 4$, $q = 2$) that results in a significantly different forecast result (RMSE = 0.015 m), although the quality of the model-fit to the data was only incrementally different from the ARMA ($p = 5$, $q = 2$) model. Of course, these forecasts can be refined with confidence intervals, but the message from the example is clear: forecasting is a possibility but one is advised to apply it for short time windows only and test a range of model configurations (i.e. p and q values), as a 'good' fit to the training data does not guarantee the 'best model forecast'. The latter statement holds for any other type of model (be it process based, or a statistical model), the validity beyond the calibration domain cannot be taken for granted.

An alternative application is to use similarity between time series. Figure 9.30 shows an example: two hydrographs from locations 1 and 2 are shown together with their difference (top graph in Figure 9.30). The middle and bottom graphs show, respectively, the autocorrelation function for the difference between the time series (i.e. $x_3 = x_1 - x_2$) and the differentiated difference (i.e. $\texttt{diff}(x_3)$).

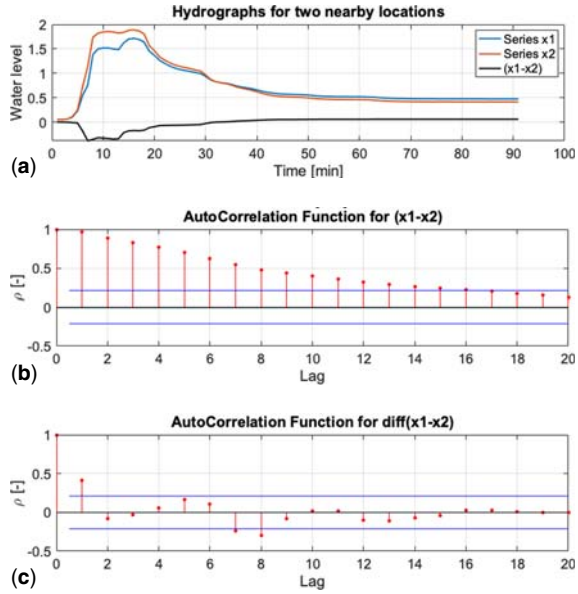


Figure 9.30 (a) two hydrographs and their difference; (b) ACF; (c) DACF. *Source*: Francois Clemens-Meyer (Deltares/TU Delft/NTNU).

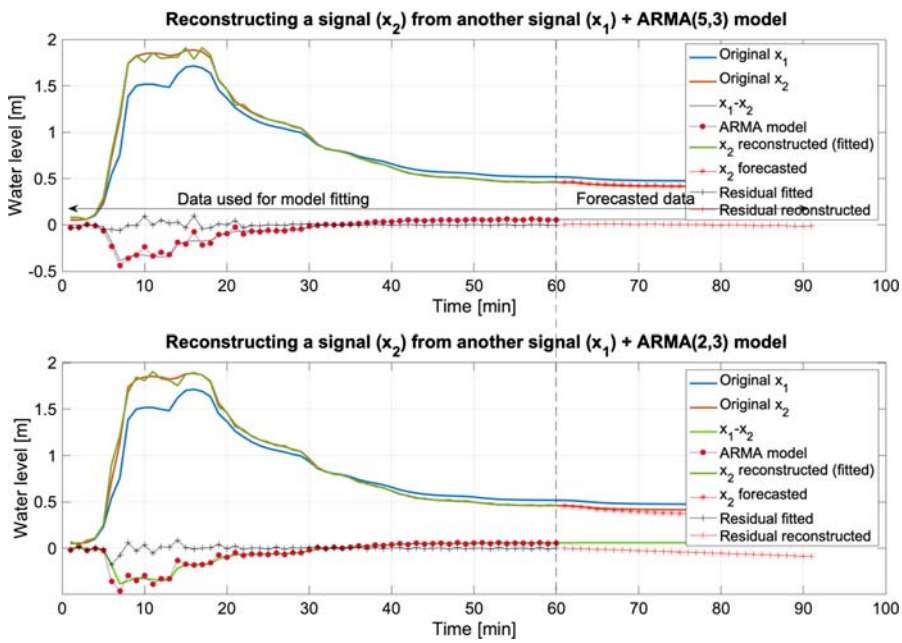


Figure 9.31 Results of forecasting of one signal based on an ARMA model for the signals differences combined with one signal available. *Source*: Francois Clemens-Meyer (Deltares/TU Delft/NTNU).

In this case a first estimate for p would be 2, while for the value of q the same reasoning is followed as before, as the ACF start tailing off to zero in the first few lags. Thus a first estimate for q would be 1 or 2. Again the best model is decided upon based on the AIC values.

Notice that the p value is found to be 5, in contrast to the indication from the DACF (see Figure 9.30). This illustrates that the indication obtained from the DACF does not necessarily correspond with the 'best' model using AIC as a metric. It is therefore, again, suggested to test a range of values for p and q.

Figure 9.31 shows some results: the upper graph for the best model configuration found, and the lower graph for a slightly different model configuration. Again, it is seen that a slightly different model results in significantly different results again. The advantage of the latter approach based on differences between two observation locations is that only one model needs to be maintained enabling the possible curation of two sensors' outputs. A regular update of the model configuration is suggested, as the recorded processes may change over time in terms of level of stationarity. In that respect, when it is found that the best fitting model changes over time in terms of AIC result or even in variation for the order (p, q) of the model, it can be used as an indication for changes in the system observed.

9.5 MAKING QUALITY FLAGS OPERATIONABLE

9.5.1 Concatenation of quality flags

Each individual recording can be checked and assessed through different tests. In the previous sections, these tests have been described, producing a more or less differentiated output for each individual test and for each recording. For each value V_t, i.e. each sensor and each time stamp, several outputs are available to further specify data quality according to the N_T tests presented above (see Table 9.5).

In order to assess a complete data set, the quality labels for individual recordings need to be concatenated: manually validated by a trained staff member to split the values labelled with a D into the G or the U categories.

There are several methods to achieve such concatenations, i.e. to perform a dichotomous flagging or label data points as 'Good' or 'Unsuitable', while further differentiating the quality of recordings labelled 'Doubtful':

- Method 1 (worst case): assign the final quality as the common minimum, i.e. the lowest quality.
- Method 2 (arithmetic mean): calculate an average, while assigning a numerical value to qualitative flags, e.g. 1 for 'Good', 0.5 for 'Doubtful' and 0 for 'Unsuitable' and assigning thresholds.
- Method 3 (median): based on the same principle as Method 2, but it calculates the median of the cardinal qualities.

Each of those methods has pros and cons: the first one being rather pessimistic, the second one being sensitive the grade attribution and potential weights in the average calculation, and the last one being

Table 9.5 Possible basic tests for data validation. Output: G, Good, D, Doubtful and U, Unsuitable.

Value	Test 1	Test 2	Test 3	Test 4	…	Test N_T
V_t	G	G	D	U	…	D

sensitive to a series of failed tests. Based on the output in the didactical example given in Table 9.5, the output of the concatenation will be:

- Method 1: 'Unsuitable', due to Test 4 or Test N_T (5 in this case).
- Method 2: 'Doubtful' while assigning the same weight for each test (average score of 0.6 $(1 + 1 + 0.5 + 0 + 0.5)/5$).
- Method 3: 'Doubtful', the median is equal to 0.5.

The *advantage* of the concatenation is its flexibility: the tests are taken into account, the weights assigned to each output can be changed according to the different purposes the data have been recorded for.

The *disadvantages* may be that (i) one reduces the overall amount of useful data when applying a stringent method (Method 1), or (ii) one introduces a bias when transforming an ordinal scale (G > D > U) into a cardinal, i.e. ratio scale (Method 2).

Those concatenations can be done automatically but are prone to subjectivity regarding the selected tests, thresholds, weights and the retained method to concatenate the outputs. However, this step is mandatory to simplify a subsequent manual validation. At the end of the automatic concatenations, a value can have one or several labels about its quality, one for each purpose.

A suggested concatenation method

- I 9.5: *Test* – Test the three proposed methods and compare the result to select the most appropriate for your needs and uses.
- I 9.6: *Sensitivity* – Try different values and weights if you use methods based on mean and median.
- I 9.7: *Update* – Update and design new concatenation methods if you are not satisfied with the proposed methods.
- I 9.8: *Report* – Always report the methods used in the meta-data, the values and the weights used for the concatenation.

9.5.2 Finding causes of unreliable data being rejected

Validating measurements generally aims to distinguish between dubious and plausible data. Unreliable data may lead to wrong findings and consequently have to be excluded before further processing. This, however, often requires manual intervention to explain *why* some data need to be labelled 'Unsuitable' or 'Doubtful'.

Labelling data as discussed in Section 9.2 supports this process, but it needs to be backed with domain knowledge of UDSM systems behaviour, monitoring techniques, signal processing and data transformation. Losses in data quality may occur during the entire data collection process due to: (i) inappropriate selection of sensor and/or location; (ii) inadequate sensor installation; (iii) sensor specific issues, e.g. failing barometric compensation or electrolytes in the wastewater; (iv) data-logger related issues, e.g. synchronicity, interval, smoothing algorithms, power supply, data transfer; (v) data storing techniques (database management); and (vi) the validation process itself (software, data import routines). Persons in charge of investigating doubtful data are required to be qualified accordingly or seek professional (and mental) support.

Depending on (i) the complexity of the monitoring set-up, (ii) the environmental conditions when the value has been recorded and (iii) the existence of a site-book, investigations may require several experts:

- Persons in charge of the design, construction, maintenance of the monitoring set-up.
- Experts in metrology.
- Experts in IT, electronics and signal processing.
- Local experts with in-depth knowledge about system and locations, e.g. hydraulic conditions in a specific pipe.
- Sensor manufacturers and data acquisition suppliers.

The process of *manual validation* can be done in several steps. The main way suggested for processing is summarized as follows: Be very strict in assigning the labels 'Good' or 'Unsuitable' and any manual modification (from 'Doubtful' to 'Good', 'Unsuitable' – and sometimes 'Doubtful', if the manual validation did not result in assigning another label) must be commented and recorded in order to keep track of the conducted investigations, e.g.:

- The values of water level and velocities, flagged as 'Doubtful' because of inconsistencies between them, have been finally flagged as 'Good'. Reason: recession phase during a storm event, and strong hysteresis (see Chapter 3 or the example in Section 9.4.3.4).
- Discharge values that are flagged as 'Doubtful' were linked to a change of position of the measuring device and finally found to be 'Unsuitable'. Reasons: error in the probe positioning, and the observed bias was confirmed by tracer experiments, as outlined in Chapter 3.

The final flagging and the reasoning that may have led to manual modification should be recorded as meta-data in the database (see Chapters 5 and 10).

Manual validation requires time, expertise and common sense for complicated cases: meta-data on maintenance, sensors, storm events etc., collected in a site book, are essential to conduct such investigations properly. There is no 'silver bullet' to investigate the reasons for 'Doubtful' flags. However, a few basic questions could help to draft hypotheses regarding doubtful data points:

- Are data points flagged as 'Doubtful' assigned for a single or several time step(s)?
- Does this phenomenon occur for a single, several or every time series?
- If several time series are flagged as 'Doubtful', is there any correlation between them? Do they occur through the same computer, the same data acquisition set-up, a specific software? Do data stem from the same measurement location?
- Can a change in data quality be associated with a specific event, such as an intense storm, a maintenance operation, the installation of a nearby sensor?

The list of possible causes can be long. Unfortunately, there is no known generic method or protocol that, when applied, can ensure that in all cases the underlaying cause for poor data quality, or missing data, can be determined. Common sense, domain expertise, site and maintenance logs, and a complete documentation of the monitoring station are necessary to increase chances of problem identification.

9.6 COMMUNICATING DATA QUALITY

9.6.1 Presenting validated data

Once the data set has been validated, one of the key methods to get a user-friendly overview on its overall quality consists of plotting the data quality. Such plots will help the data providers and data users to easily assess the quality of the monitoring set-up, the recorded data set and implicitly all the procedures along the

monitoring processes. Here it is of importance that the quality level of the data is indicated in a manner that is unambiguous and fit for purpose. In this respect a difference has to be made between 'managerial' information and 'operational' information. For example, from a managerial point of view, indicators related to the overall performance of the monitoring system as discussed in Section 9.6.2 are of interest while for the operator of the same network detailed information on the level of individual sensors is sought after. Especially in long term monitoring projects, graphs visualizing the data quality provide an easy access to data quality on different levels.

9.6.1.1 Types of quality and availability plots

There are numerous ways to produce representations of data quality, mainly based on colour scale:

- 'Shades of grey' style (e.g. in Figure 9.32).
- Heatmap dashboard illustrating data consistency, i.e. availability, completeness and interpretability (e.g. in Figure 9.33).

From Figures 9.32 and 9.33, the availability and quality of data are easily recognized. Using this kind of charts allows for a quick identification of whether or not enough data are available for a given purpose. In Figure 9.32, for each sensor (Lev1-Lev27 and R1, R2) for each day (24 h), different shades of grey indicate the quality of the available data. The white columns in Figure 9.33 highlight likely failure of the whole system, since no data have been recorded by any sensor.

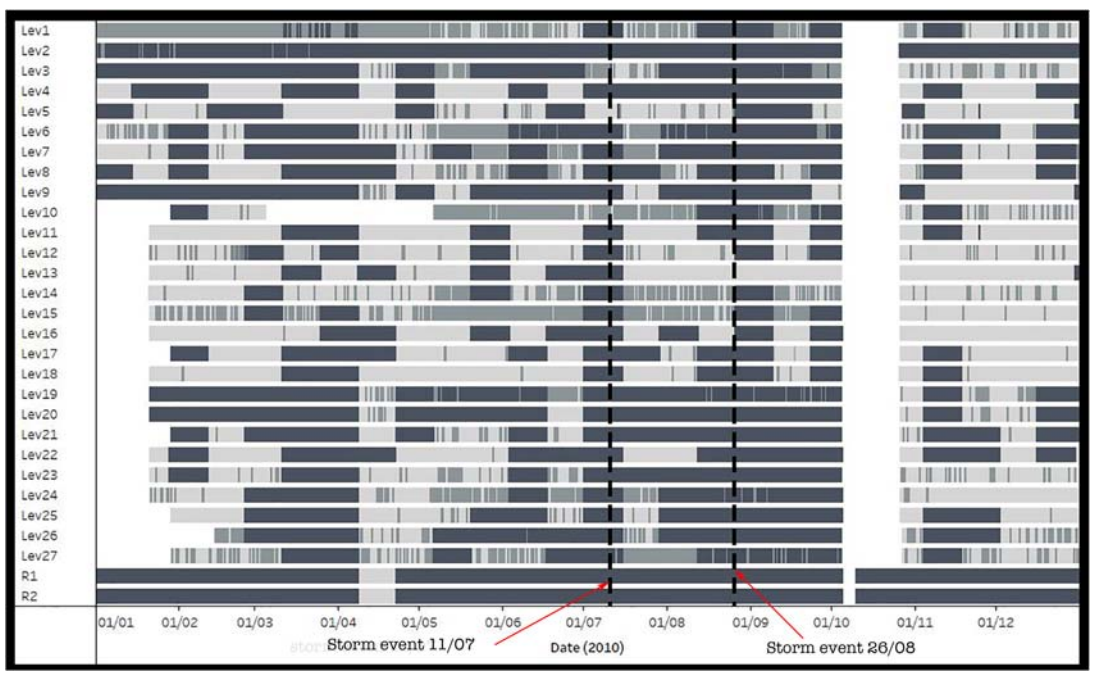


Figure 9.32 Example of a chart granting a 'quick' impression of the quality of a data set, discriminating between a range of labels indicating a range of possible 'issues' with data. *Source*: van Bijnen (2018).

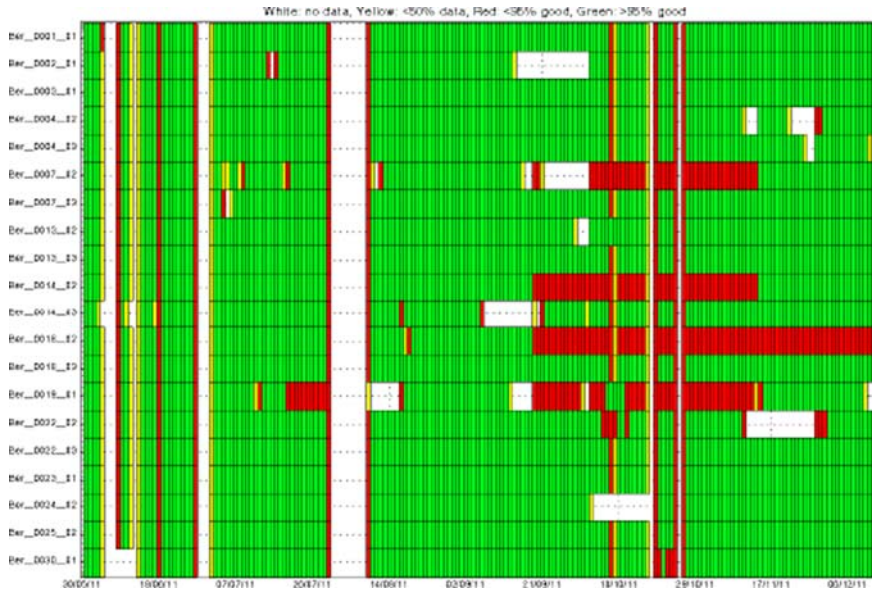


Figure 9.33 Example of a chart showing data quality in 'traffic-light' coding: green is >95% Good data, red is <95% Good data, orange indicates that <50% of data is available and needs a manual check prior to application. A blank cell indicates no data are available at all. Each row represents one sensor, each column represents a 24 h window. *Source*: van Bijnen (2018).

The choice of map style requires some reflection. On the one hand, if basic traffic light colours (or similar plots – Figure 9.33) are relatively easy to interpret, they lack information on important data or facts. On the other hand, adding too many layers on the colour map (e.g. in Figure 9.32) increases the information given by the graphic while decreasing its readability. Both plots can be produced:

- A basic one to give an overview to the management team and draft the main outlines on the data quality.
- A more complex one for technical meetings and discussions aiming at understanding and improving the current performances of the monitoring system.

Once the type of plots and the legend have been selected, they should not change over time to allow a quick assessment of the evolution of data quality, without having to learn a new way of reading for the updated version of those graphics.

9.6.1.2 Data to plot

Plots as discussed in Section 9.6.1.1 can be created with several types of data (results of the tests, automatic and final quality grades).

Creating those graphics with the results of individual tests (like consistency, gradient, etc.) can help to identify tests that all fail or succeed and, later on, can help to adjust the tuning of those tests or identify permanent errors between sensors.

Comparing graphics on global data quality grade, before and after manual validation, can help to highlight changes in the manual validation, between different data validators or over time.

The possibilities are endless and existing tools to generate plots do not really limit the number of graphics that can be produced. However, multiplying the number of plots requires more time to perform such analysis and may lead to unseen problems. The tendency to produce more and more graphics is not always suitable: it might be useful at the beginning, but only the relevant plots should be drawn once the experience is sufficient to identify the meaningless ones.

9.6.1.3 Use of those graphics

Graphs as shown in Figures 9.32 and 9.33 highlight 'Good' and 'Unsuitable' data: column(s) with 'Missing' or 'Unsuitable' data will most likely indicate an error in the data acquisition system and row(s) presenting the same characteristics will indicate a problem with one or more sensors (with various plausible causes). Adding on the timeline every single event that took place on site (storm event, maintenance, calibration, probe change, etc.) will give key information to understand what might have been the cause leading to either an 'Unsuitable' or a 'Doubtful' flag.

The main utilities of such plots are to:

- Understand what occurred on the system.
- Improve the quality of the data while proposing and testing solutions of the causes.
- Communicate the overall quality of the data to the final users (modellers, managers) or financers, while, in most cases, regularly delivering some statistics on data quality.

9.6.2 Using statistics as indicator of the overall monitoring system quality

Even if graphics are rather user-friendly tools to communicate data quality, basic statistics offer some additional highlights, especially for reporting and conducting good asset management of the monitoring networks.

9.6.2.1 Additional information given by statistics

Statistics involved in this part are really basic: they mainly consist of calculating the percentage of each flag (typically, 'Good', 'Doubtful' or 'Unsuitable') for the recorded data set. Those percentages can be and should be calculated for different subsets of the data sets: per sensor, per group of sensors (e.g. inside or outside of the sewer, sensors connected to the same hardware, same type – such as rain gauges, water level probes), per monitoring station, per catchment, etc.

Those statistics will help the data user to highlight and quantify the first impression derived from the previous graphics. Weekly, monthly or yearly values will highlight the data quality trends over longer durations.

9.6.2.2 Suggested indicators

Various indicators could be calculated with the validated data set:

- Percentage of available data (Equation (9.39a) or Equation (9.39b))

$$100 \times \frac{N_G + N_D + N_U}{N_E} \tag{9.39a}$$

$$100 \times \frac{N_E - N_M}{N_E} \tag{9.39b}$$

- Percentage of 'Good' data (Equation (9.40))

$$100 \times \frac{N_G}{N_A} \tag{9.40}$$

- Percentage of 'Doubtful' data (Equation (9.41))

$$100 \times \frac{N_D}{N_A} \tag{9.41}$$

- Percentage of 'Unsuitable' data (Equation (9.42))

$$100 \times \frac{N_U}{N_A} \tag{9.42}$$

- Percentage of 'Doubtful' data finally considered as 'Good' after the manual validation (Equation (9.43))

$$100 \times \frac{N_{D \to G}}{N_D} \tag{9.43}$$

- Percentage of 'Doubtful' data finally considered as 'Unsuitable' after the manual validation (Equation (9.44))

$$100 \times \frac{N_{D \to U}}{N_D} \tag{9.44}$$

- Percentage of data that remain 'Doubtful' after the manual validation (Equation (9.45))

$$100 \times \frac{N_{D \to D}}{N_D} \tag{9.45}$$

Those indicators should be calculated for several subsets of the entire data set:

- For individual sensors.
- For groups of sensors, e.g. inside/outside pipe, water level/velocity/discharge probes, by manufacturer/sensor connected to the same data acquisition hardware.
- For each monitoring station.
- For each catchment.

This type of indicator can be used when judging the performance of a monitoring network. Especially for long-term monitoring activities, keeping track of the monitoring system performance can provide crucial information for, amongst others, improving the data yield.

9.6.2.3 Indicators and asset management of the monitoring system

Data quality indicators and their evolution over time offer multiple opportunities to conduct asset management of a monitoring system, especially while using the meta-data associated with the validated data set. Any positive or negative change in the data quality indicators can lead to confirmation of good

decisions/practices or to new decisions to improve the monitoring system. Since an exhaustive list of cases/situations/conclusions is nearly impossible to draft, a few examples are listed below:

- The percentage of data labelled as 'Doubtful' increases a few weeks after a cleaning procedure (meta-data) of a water sensor. This behaviour occurs several times. The delay between two cleaning procedures should be shortened.
- The percentage of data labelled as 'Good' is higher for a pressure sensor than for a US water level sensor at the same location. Pressure sensors seem to be more suitable at this location.
- The percentage of available data for a group of sensors (connected to the same hardware) dropped after its replacement. The new hardware and its installation need to be checked.
- The percentage of data labelled as 'Good' increases after the refurbishment of a monitoring station. The new design and set-up are better than the previous ones: the future replacement of existing monitoring stations should be done the same way.
- The quality of data decreased after some changes in the maintenance/calibration protocols. Those protocols and their realizations need to be carefully checked and compared with the previous ones in order to identify potential issues in the new protocols.

The list of examples is virtually endless. The general rule consists of having a deep look into data quality indicators, keeping track of any change and correlating those behaviours with other data and meta-data, i.e. rain event, maintenance events, hardware or software upgrades, etc.

9.7 DATA CURATION

When deciding on the methods, protocols and their thresholds and further settings, a decision has to be made on what to do with missing or discarded data (i.e. data in category 'Doubtful', 'Unsuitable' and 'Missing'). In many cases missing a few data does not influence the decisions ultimately taken based on them. For instance, when it comes to the evaluation of the environmental performance of a combined sewer system, three or four missing data points over a period of a year is not an issue. However, in the case where six months of data are missing in the same situation, this may make the data set useless.

At the other end of the spectrum, when e.g. model calibration is the main objective of the monitoring project, a time shift of just 1 minute can be enough to make data completely worthless. In the former case no action is needed, other than reporting that out of the 20,000 recorded data points 4 were missing. In the latter case a decision has to be made; either discard the data set altogether and wait for better times, or 'repair' the data obtained and 'make them useable'. From a strictly scientific perspective, the latter is considered a death sin, as information is added that was not actually measured and as such cannot be accepted as an objective basis to work upon.

However, when relaxing this point of view a little bit, it can be argued that when there is convincing evidence that an observed time shift is due to a cause that has been identified, (e.g. a documented application of an incorrect reference level), a correction of the data can be accepted and imposed under the condition that this is clearly stated, hence the label 'Doubtful' for cured data.

In any case, it is strongly suggested to always keep and maintain two data sets: (i) the original time series, with gaps and (ii) the 'cured' time series, with interpolated values. This solution offers some advantages: (i) original data are not overwritten by interpolated data, (ii) the two data sets may serve different purposes and needs, and (iii) alternative interpolation methods can be tested and performed afterwards.

Can we trust or rely on interpolated values? Some points have to be clearly stated: (i) an interpolated value is, as repeatedly stated, a virtual value. Unusual, but real, phenomena may have occurred during the gaps, (ii) interpolation methods are likely not to be able to reconstruct such a phenomenon.

How can the uncertainties of interpolated values be quantified? In line with one of the main messages of this book, the standard uncertainty associated with each value should be estimated. The standard uncertainty of an interpolated value has to take into account two sources: the uncertainty of the measurement (like this value has been normally recorded) and an additional uncertainty due to the interpolation process itself.

Imputation of data in time series is a subject not restricted to the field of UDSM. The study of time series and all their aspects are comprehensively discussed and explained in textbooks (e.g. Hamilton, 1994). Wongoutong (2020) provides a state-of-the-art review on methods applied for imputation of missing values in time series.

9.7.1 What to do with outliers, trends or data gaps in general?

Once gaps, outlier(s) or trends are detected in a data set, the question arises over what to do with them. A first omni-important action to take is to try to find out what caused the outlier, trend or missing data. In many cases malfunctioning, or wrongly installed equipment proves to be the cause. On the other hand, in many cases the cause remains unknown, and could therefore represent a 'real' value and as such hold information on the system studied that might be important. In such cases it is worth trying to find out whether there is some temporal pattern in the occurrence of outliers at the given location for the specific sensor. The example presented in the beginning of the chapter (disturbing lamppost) illustrates the importance of meta-data. In this case the outliers were first marked as 'outlier', which is as such a warning for the use for further analysis. After unambiguously determining the cause and interpolation, the meta-data was changed into 'imputed'. Of course, this type of protocol has to be designed and applied to the specific demands of a given project.

Basically, applying the following sequence with respect to applying data imputation is suggested:

(1) Try to use the data in a piecewise manner, that is, use those time windows in which 'Good' data is present without gaps.
(2) If 1/does not apply, look for the time windows with as much 'Good' data as possible.
(3) If 2/does not result in enough data for analysis, a first data imputation can be made.
(4) A first step is to impute data for single point outliers (see Section 9.7.2).
(5) If data gaps are present (more than a few time steps), data reconstruction may be considered, e.g. from known correlations with other measuring locations or from a model running in parallel.

The latter situation is the most difficult one as it is not simple to set a limit on the allowable amount of missing data. Essentially this boils down to carrying out a risk assessment in terms of making a wrong decision. For pure scientific purposes, the situation is relatively simple: when no data are available, no analysis can be made, so the challenge is to obtain more data and improve the quality until the data are usable. For practical applications, it is more complicated: depending on the purpose for which the data are being collected, more or less missing data can be acceptable, or imputation is seen as 'normal and accepted practice'. For example, if a system has been monitored on dry weather flow patterns for a few years, missing a few days of data per year does not pose a serious impact on the uncertainty on e.g. the total volume discharged per year (one could impute the expected behaviour from historical data). On the other hand, when data gaps occur in a real time controlled system, this has a direct impact on the effectivity of the system. Therefore, again, how to handle missing data is largely a matter of subjective decision making. Regardless of the circumstances, however, the fact that a data value stems from imputation has to be explicitly clear from the meta-data.

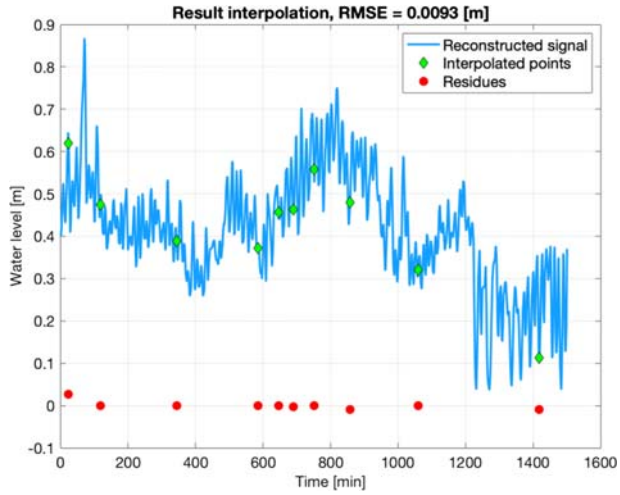


Figure 9.34 Example of interpolated data points. *Source*: Francois Clemens-Meyer (Deltares/TU Delft/NTNU).

9.7.2 Imputation of small data gaps

When 'curing' a single outlier, linear midpoint interpolation (Equation (9.46)) is the easiest and most straightforward method (Figure 9.34):

$$V_t = 0.5(V_{t-\Delta t} + V_{t+\Delta t}) \tag{9.46}$$

Figure 9.34 shows the results of the midpoint interpolation of artificially removed data points. The RMSE of the residues is approx. 9 mm, which is in the same order of magnitude of the confidence range as normally expected for water level sensors. This implies that the curation through interpolation in this case has no noticeable effect on the confidence interval of the data point (see Lepot *et al.*, 2017).

Al Janabi (2005) suggests the following interpolation values for up to two successive missing values, based on the previous and following value:

One point missing (Equation (9.47)):

$$V_t = \sqrt{V_{t-\Delta t} \times V_{t+\Delta t}} \tag{9.47}$$

For two successive points missing (Equation (9.48)):

$$V_t = \sqrt[3]{V_{t-\Delta t}^2 \times V_{t+2\times\Delta t}} \tag{9.48a}$$

$$V_{t+\Delta t} = \sqrt[3]{V_{t-\Delta t} \times V_{t+2\times\Delta t}^2} \tag{9.48b}$$

When filling larger data gaps, a more advanced method is to use a (calibrated) model to fill in gaps, e.g. either a deterministic model, or a conceptual model like an ARMA model, spline fitting method or ML applications.

Gaps may, and will, occur in data sets. There are numerous methods to fill the gaps (if needed) and, in any case, those interpolated values should be labelled as 'D' or 'U' (see Table 9.1), depending on the goals the data have been recorded for. Gaps vary in: (i) size, a single value missing or more; (ii) continuity, a single

gap or a series of several gaps; and (iii) the number of impacted time series, typically if a gap is due to a sensor failure or if gaps are due to a data acquisition system failure (a few time series present gaps at the same time). Lepot *et al.* (2017) wrapped up the state of the art regarding interpolation in time series. The next section presents a brief summary of this review, while being restrained to the main method, adding a few examples and advice for practitioners.

9.7.3 Imputation of larger data gaps

Applied methods for filling larger data gaps (i.e. >2 consecutive missing records) are divided into two categories: the deterministic and the stochastic ones. This distinction, based on the existence or not of residuals (differences between prediction at known location and observations) in the interpolation function, deals also with uncertainty.

9.7.3.1 Deterministic methods

The easiest method, not highly recommended for large gaps in dynamic time series, is the nearest-neighbour interpolation: the interpolated values are equal to the closest recorded ones. Other straight forward to implement methods are LOCF (last observation carried forward) or NOCB (next observation carried backward). Such methods are fast and simple, and find their main application in RTC systems.

Example: There is a gap in water level data with missing values between 10:51 am and 10:59 am. At 10:50 and 11:00, the recorded water levels are respectively 10 cm and 12 cm. With this method, interpolated water levels are 10 cm until 10:55 and 12 cm afterwards (Figure 9.35).

In order to avoid this discontinuous behaviour, smoother functions can be used to interpolate data e.g. with linear or polynomial interpolation methods.

Example: While re-using the same example with a linear interpolation, the interpolated values will slowly vary from 10 to 12 cm, while reaching 11 cm at 10:55 am (Figure 9.35).

Phenomena in urban hydrology are rarely linear: polynomial interpolation or application of ARMA models (see Section 9.4.4) will better mimic the expected shape of the time series. Methods based on

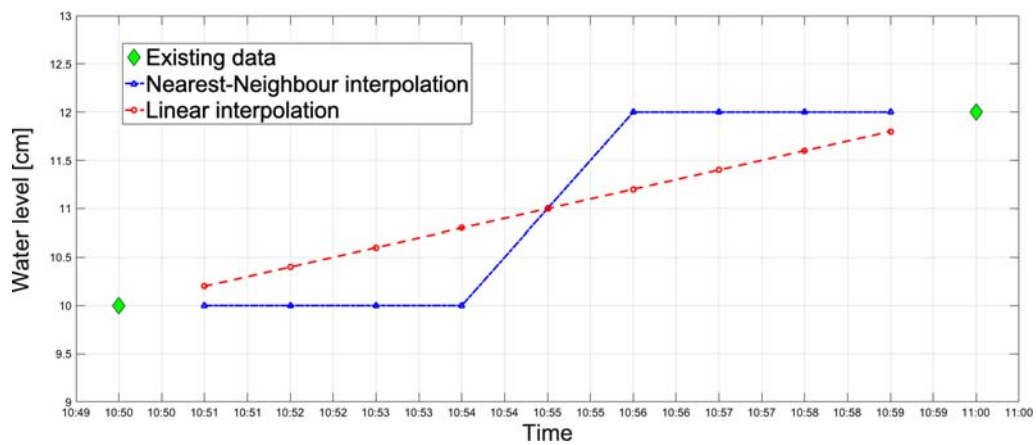


Figure 9.35 Existing and interpolated data. *Source*: Mathieu Lepot (TU Delft/Un poids une mesure).

distance-weighting could be also used: an average, weighted by the distance between the interpolated and recorded values.

Other functions can be used to achieve such interpolation, especially equations that reflect the phenomenon: dispersion of pollutant, correlation between water level, velocity and discharge (like the Manning-Strickler equation), run-off models, etc. Those methods are less based on mathematics and more on the physical processes that occur at the measuring location. A few of those approaches are nearly impossible to perform if the entire data acquisition system collapses.

9.7.3.2 Stochastic methods

Machine learning approaches such as neural networks, kernel methods and kriging are also available for data interpolation. Methods based on data dynamics seem to be more appropriate in urban hydrology and deserve more detail.

The k-Nearest Neighbours (k-NN) takes into account the cyclic variations of a time series (e.g. discharge during a dry business day). Assuming or knowing the dynamics are similar from day to day, the gap can be fulfilled with data from another day. There are several metrics to identify which part of the recorded values is most suitable for the interpolation e.g. city block, Euclidean or Chebychev. Box-Jenkins models are suitable for polycyclic data, including seasonality and daily or weekly patterns. Pratama *et al.* (2016) provide a review on handling missing data in time series.

9.7.3.3 Uncertainty assessment

The uncertainty of interpolated data has two components: (i) the uncertainty of measurement if this value has been normally recorded and (ii) the uncertainty from the interpolation process itself. If the first component has been detailed in Chapter 8, the second one requires a few tips given in this section.

Given two known values (V_t and $V_{t+\Delta t}$), the standard uncertainty $u(V_{I,i})$ of an interpolated value $V_{I,i}$ is calculated according to Equation (9.49) (Lepot *et al.*, 2017):

$$u(V_{I,i}) = \sqrt{\frac{1}{2}\sigma_M^2\left(3 + |\rho(V_t, V_{t+\Delta t})| - 2|\rho(V_t, V_{I,i})| - 2|\rho(V_{t+\Delta t}, V_{I,i})| - \frac{(\rho(V_{t+\Delta t}, V_{I,i}) - \rho(V_{t+\Delta t}, V_{I,i}))^2}{1 - \rho(V_t, V_{t+\Delta t})}\right) + \sigma_P^2} \tag{9.49}$$

where σ_P is the process variance, ρ is the autocorrelation function and σ_M is the measurement error.

Figure 9.36 shows several methods to assess uncertainties of interpolated data: the Law of Propagation of Uncertainty (LPU, top left), Monte Carlo simulations (MC, top right), a method proposed by Schlegel *et al.* (2012) (bottom left) and Equation (9.49) (bottom right). Only Equation (9.49) proposes uncertainties for interpolated values, which present a correct trend: (i) being higher than the measurement ones, and (ii) a continuity at the edge of the interpolation area.

9.8 DATA-DRIVEN METHODS

The discussion of methods related to machine learning (ML) – hereinafter referred to as *data-driven methods* – has been intentionally discarded here for two main reasons. Firstly, those methods have only very recently been applied, in various fields with differing success. Their application in the UDSM field has not yet reached the level of maturity that would allow an objective judgement of their usefulness. Secondly, data-driven methods are often based on a black-box approach: the method parameters often lack physical interpretation, data may be rejected (e.g. as false positive) without explicitly providing a

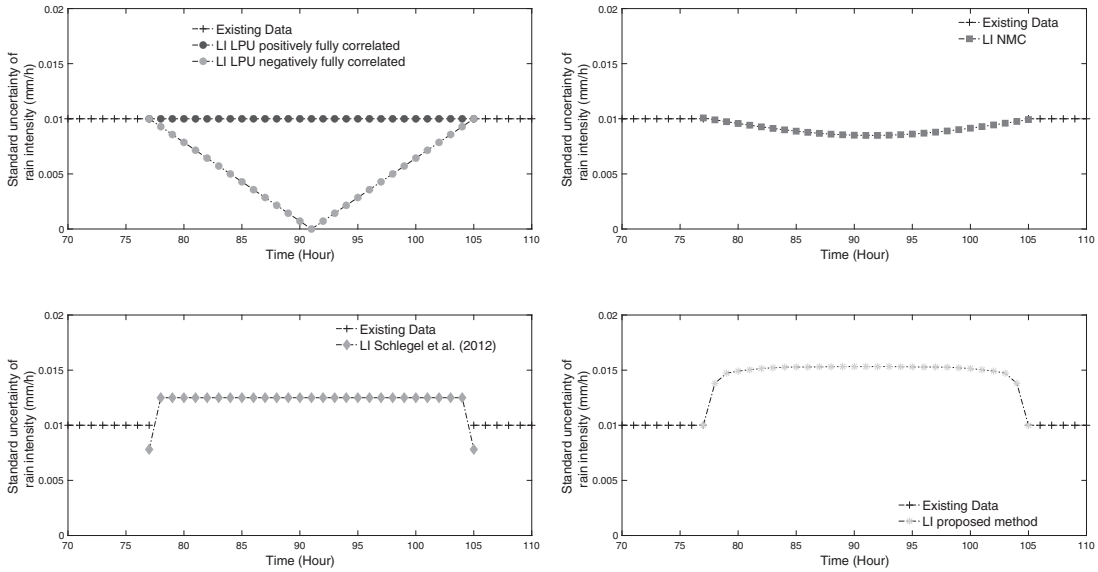


Figure 9.36 Uncertainties of interpolated values *vs.* uncertainties of existing data (black +) for a linear interpolation. *Source*: adapted from Lepot *et al.* (2017).

reason, i.e. the input to understand this rejection and better understand the monitored system. This approach is contrary to the one we want to propose in this book. However, data-driven approaches are rapidly developing.

It is acknowledged that the suite of data-driven concepts may gain further popularity once domain knowledge has sufficiently been integrated into purely machine/data-driven approaches. In the following section, a brief, but clearly limited review of why such concepts can be useful and what challenges are associated with their application is presented.

9.8.1 Motivation

Existing guidelines (Bertrand-Krajewski & Muste, 2007; DWA, 2011) and automated data validation pipelines (e.g. Alferes & Vanrolleghem, 2016; Branisavljević *et al.*, 2010) specifically developed for UDSM applications provide very useful solutions. Generally, these approaches suggest a consecutive application of standard rule-based methods for basic data validation. Rule-based methods – some of which are discussed in Sections 9.3 and 9.4 – imply the use of parameters defined based on expert knowledge about sensors and system behaviour (min/max ranges, acceptable changes, pre-defined correlations). Integrating this expert knowledge in the form of manual intervention and subjective judgement is rather expensive, and it does not necessarily lead to reproducible results. Other limitations become obvious in the case of real-time applications requiring minimum latency, e.g. if it is necessary to simultaneously check several signals of different types in real time to trigger control, and/or if computationally expensive analysis methods are used.

Conducting *automated* data validation for identifying abnormal behaviour of the deployed sensors is therefore becoming increasingly critical. Data-driven methods promise easy help here. They suggest a timely, coherent, complete and, with increasing data volume, more efficient assessment of data quality

(Aggarwal, 2017). The algorithms ability to 'learn' suggests higher efficiency with minimal human intervention while increasing flexibility.

While data-driven methods are experiencing a boom in the field of image processing, natural language processing, speech recognition, stock portfolio management, and other fields, so far only very few applications are known in the field of urban hydrology. Troutman *et al.* (2017) combine Gaussian processes (dry-weather flows) and dynamical System Identification (wet weather discharge) aiming to simulate rainfall-run-off dynamics in a combined sewer network purely based on sensor data. Although the detection of novelty in monitored data had not been the primary objective, this approach could be applied to do so. Russo *et al.* (2019) present an anomaly detection method based on a convolutional neural network (CNN), i.e. a deep autoencoder, for validating urban drainage monitoring data. However, the suggested methods come along with some deficiencies: immense data pre-processing is required, and a high false positive rate is still present, although the latter aspect is partially justified with a somewhat high complexity of the data, i.e. in-sewer flows, used in the study. Rodriguez-Perez *et al.* (2020) applied a range of artificial neural networks on water-quality data (turbidity and conductivity) with high temporal resolution to evaluate the 'best' performing model depending on the variable-, environment-, and anomaly type. Common anomaly types present in online water quality data (with characteristics very similar to variables monitored in urban drainage) were previously categorized by Leigh *et al.* (2019). Rodriguez-Perez *et al.* (2020) in turn found that semi-supervised classification was better able to detect instantaneous faults (e.g. spikes), whereas supervised classification had higher accuracy for predicting long-term anomalies, such as drifts.

Despite the fact that results of these studies look rather promising, further systematic evaluation on *different* real-life data sets applying *different* data-driven approaches is required to show the usefulness and likewise the limitations of such approaches.

9.8.2 Challenges and constraints

Direct application of purely data-driven methods for the validation of urban hydrological data is challenging for several reasons: (i) system-determining rainfall events with random occurrence do not result in easily recognizable patterns; (ii) the range of values of measured state variables is sometimes limited on one side (e.g. in the case of flow restrictions caused by a throttle) resulting in unilaterally constrained data; (iii) processes are inherently non-linear – the boundary conditions are difficult to define; and (iv) the complexity of some urban drainage signals, i.e. the decomposition of overlapping fluxes of different dynamics, can be challenging.

Current research in the field of ML mostly focuses on new and signal-specific methods, or method comparisons with limited representativeness. Comparisons and proofs of application are often carried out with synthetic, i.e. modelled data sets or data from other fields. Description or necessary steps for the preceding data preparation is often omitted, or their effect on the final result is often unclear. Against this background, it is obvious that more studies are required, applying various methods to different data sets (benchmarking of algorithms applied on open-access data). The repetition of experiments, i.e. reproduction of results using new methods, would increase the trustworthiness of data-driven methods.

Existing and future studies in the UDSM field should be carefully evaluated. The specific urban drainage context (constrained/unconstrained, dry/wet weather) must be taken into account. If the community manages to integrate sufficient domain knowledge into machine learning concepts (basically shifting from black box to grey box models), data validation for urban drainage applications may experience a

significant push. However, it has to be kept in mind that ML may be a convincing hammer, but not every problem is a nail!

9.9 SUMMARY AND TRANSITION

Validation of field measurements is crucial to ensure data consistency and to allow for optimal interpretation of the data. High-quality data increases the trustworthiness of the derived information enabling informed decisions, but obtaining high-quality data is not an easy task.

This chapter attempts to remedy this situation. It discusses individual aspects of checking the plausibility of data points, assessing their quality, and it strives for options to curate data as an inherent part of the validation process, and/or as a subsequent step. A brief excurse on the use of machine learning techniques is given due to its increasing popularity, also in the field of data validation.

The process of data validation should be understood as a stepwise approach, which can be split into a basic check of consistency and plausibility, and a more subjective assessment. If applied, the latter should clearly be dependent on the purpose the data is used for. Meta-data is considered decisive for correct data interpretation. This additional, often non-numeric information should be collected systematically and archived with a distinct relation to the corresponding data point(s).

Various tests for data quality assessment are introduced: from the simpler to the more complex ones. Once the outputs of individual tests are calculated, results may be concatenated to obtain a single metric per data point. Despite the fact that methods are mathematically founded, there can be substantial subjectivity associated with their application. Some of the statistical techniques described require – in a strict mathematical sense – data properties that are fulfilled. Still, these techniques are successfully applied and widely accepted in practice. In such cases, this has been indicated, but scholastic correctness has been considered as subordinate in favour of practicality. End users must be aware of such subjectivities. In any case, pedantic documentation and transparent communication on parameters, weights, and methods applied is highly recommended.

With the rise of new sensors and data communication technologies, and as we are adopting the Internet of Things (IoT), collecting data has become less cumbersome, even in such challenging environments as UDSM systems. This will inevitably lead to a substantially increased amount of data. But it is anticipated that the quality of that data will not necessarily increase in the course of this trend. This, in turn, raises the importance of a *quantitative* data quality assessment, as it is for the automation of this process.

In a positive sense, this trend will stimulate the development of new methods, and their integration into automatized data validation pipelines. But it will also come with new challenges: higher complexity, higher diversity, an increased risk of confusion, and a lack of transparency in the process. The following strategies will help to efficiently tackle present and future challenges:

- Introducing a basic level of harmonization, effectively resulting in somewhat standardized approaches, including commonly agreed interfaces in the data assessment pipeline. What does it bring? It allows the comparison of the quality of different data sets, for instance in the form of benchmarking between different data providers and different systems. It enables performance assessment of different evaluation approaches. Defined interfaces enable straightforward integration of new methods.
- Development of data literacy across different qualification levels, that is from a technician that does sewer maintenance to the CEO that manages a wastewater utility.

- Establishing a culture of open-data, i.e. data sharing internally and between organizations. The evaluation of anonymized data sets in the form of a benchmarking process leads to objective performance assessment and continuous improvement.
- Increasing the degree of automation in the data validation process.

Along these lines, it can be expected that, in the (near) future, subjectivities in the data validation process can be minimized, and the efficiency increased by using advanced data analysis tools. From the perspective of scientific importance as well as for practical applications, all this would be beneficial to further facilitate the use of validated data for evidence-based decision making in the field of urban drainage and stormwater management.

REFERENCES

Aggarwal C. C. (2017). *Outlier Analysis*, *2nd edn*. Springer, New York (USA), 488 p. ISBN 978-3319475776.

Al Janabi M. A. M. (2005). *Financial Risk Management: Application to the Moroccan Stock Market*. Al Akhawayn University Press, Ifrane (Morocco), 137 p. ISBN 9789954413470.

Alferes J. & Vanrolleghem P. A. (2016). Efficient automated quality assessment: dealing with faulty on-line water quality sensors. *AI Communications*, **29**(6), 701–709. doi: 10.3233/AIC-160713.

Barnett V. & Lewis T. (1996). *Outliers in Statistical Data*, *3rd edn*. John Wiley & Sons, Chichester (UK), 604 p. ISBN 978-0471930945.

Bayley G. V. & Hammersley J. M. (1946). The 'effective' number of independent observations in an autocorrelated time series. *Supplement to the Journal of the Royal Statistical Society*, **8**(2), 184–197. doi: 10.2307/2983560.

Bertrand-Krajewski J.-L. & Muste M. (2007). Data validation: principles and implementation. In *Data Requirements for Integrated Urban Water Management*, T. Fletcher & A. Deletic (eds.), Taylor & Francis, London (UK), Urban Water series – UNESCO IHP, 103–126. ISBN 9780415453455.

Branisavljević N., Prodanović D. & Pavlović D. (2010). Automatic, semi-automatic and manual validation of urban drainage data. *Water Science and Technology*, **62**(5), 1013–1021. doi: 10.2166/wst.2010.350.

Choi B. (1992). *ARMA Model Identification*. Springer, New York (USA), 212 p. ISBN 978-1461397472.

Conover W. J. (1999). *Practical Non-Parametric Statistics*, *3rd edn*. John Wiley & Sons, New York (USA), 584 p. ISBN 978-0-471-16068-7.

DWA (2011). *Merkblatt DWA-M 181 Messung von Wasserstand und Durchfluss in Entwässerungssystemen [Measurement of water level and flow in urban drainage systems]*. DWA – Deutsche Vereinigung für Wasserwirtschaft, Abwasser und Abfall, Hennef (Germany), 10 p. ISBN 978-3-941897-94-6. Available at https://webshop.dwa.de/de//dwa/download/?link=TV8xODFfMDlfMjAxMS5wZGY= (accessed 21 Dec. 2020).

Emmert-Streib F. & Dehmer M. (2019). Large-scale simultaneous inference with hypothesis testing: multiple testing procedures in practice. *Machine Learning & Knowledge Extraction*, **1**(2), 653–683. doi: 10.3390/make1020039.

ESS (2018). *ESS Handbook – Methodology for data validation 2.0 – Revised edition 2018*. European Commission, Eurostat, European Statistical System, Brussels (Belgium), 85 p. Available at https://ec.europa.eu/eurostat/cros/system/files/ess_handbook_-_methodology_for_data_validation_v2.0_-_rev2018_0.pdf (accessed 21 Dec. 2020).

Gray K. L. (2007). *Comparison of Trend Detection Methods*. PhD thesis, University of Montana, Missoula (MT), USA, 97 p. Available at https://scholarworks.umt.edu/cgi/viewcontent.cgi?article=1247&context=etd (accessed 21 Dec. 2020).

Grubbs F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, **11**(1), 1–21. doi: 10.1080/00401706.1969.10490657.

Hamilton J. D. (1994). *Time Series Analysis*. Princeton University Press, Princeton, NJ (USA), 820 p. ISBN 978-0691042893.

Iglewicz B. & Hoaglin D. C. (1993). *How to Detect and Handle Outliers*. American Society for Quality Control, Milwaukee (USA), 87 p. ISBN 9780873892476.

Leigh C., Alsibai O., Hyndman R. J., Kandanaarachchi S., King O. C., McGree J. M., Neelamraju C., Strauss J., Talagala P. D., Turner R. D. R., Mengersen K. & Peterson E. E. (2019). A framework for automated anomaly detection in high frequency water-quality data from in situ sensors. *Science of the Total Environment*, **664**, 885–898. doi: 10.1016/j.scitotenv.2019.02.085.

Lepot M., Aubin J.-B. & Clemens F. H. L. R. (2017). Interpolation in timeseries: an introductive overview of existing methods, their performance criteria and uncertainty assessment. *Water*, **9**(10), 796, 20 p. doi: 10.3390/w9100796.

Lettenmaier D. P. (1976). Detection of trends in water quality data from records with depended observations. *Water Resources Research*, **12**(5), 1037–1046. doi: 10.1029/WR012i005p01037.

Leutnant D., Hofer T., Henrichs M., Muschalla D. & Uhl M. (2015). Model-based time drift correction of asynchronous measurement data in urban drainage. *Proceedings of the 10th International Conference on Urban Drainage Modelling*, 20–23 September, Mont Sainte Anne, Quebec, Canada, 131–134.

Lindenmayer D. F. & Likens G. E. (2018). Maintaining the culture of ecology. *Frontiers in Ecology and the Environment*, **16**(4), 195. doi: 10.1002/fee.1801.

Mourad M. & Bertrand-Krajewski J.-L. (2002). A method for automatic validation of long time series of data in urban hydrology. *Water Science and Technology*, **45**(4–5), 263–270. doi: 10.2166/wst.2002.0601.

Pratama I., Erna Permansari A., Ardiyanto I. & Indrayani R. (2016). A review of missing values handling methods on time-series data. *Proceedings of the 2016 International Conference on Information Technology Systems and Innovation (ICITSI)*, 24–27 Oct., Bandung, Bali (Indonesia), 1–6. doi: 10.1109/ICITSI.2016.7858189.

Rodriguez-Perez J., Leigh C., Liquet B., Kermorvant C., Peterson E., Sous D. & Mengersen K. (2020). Detecting technical anomalies in high-frequency water-quality data using artificial neural networks. *Environmental Science and Technology*, **54**(21), 13719–13730. doi: 10.1021/acs.est.0c04069.

Russo S., Disch A., Blumensaat F. & Villez K. (2019). Anomaly detection using deep autoencoders for in-situ wastewater systems monitoring data. *Proceedings of the 10th IWA Symposium on Systems Analysis and Integrated Assessment (Watermatex 2019)*, September 1–4, Copenhagen, Denmark, 7 p. Available at https://arxiv.org/abs/2002.03843 (accessed 21 Dec. 2020).

Schilperoort R. P. S. (2011). *Monitoring as a Tool for the Assessment of Wastewater Quality Dynamics*. PhD thesis, Delft University of Technology, Delft, The Netherlands, 330 p. ISBN 978-90-8957-021-5.

Schlegel S., Korn N. & Scheuermann G. (2012). On the interpolation of data with normally distributed uncertainty for visualization. *IEEE Transactions on Visualization and Computer Graphics*, **18**(12), 2305–2314. doi: 10.1109/TVCG.2012.249.

Troutman S. C., Schambach N., Love N. G. & Kerkez B. (2017). An automated toolchain for the data-driven and dynamical modeling of combined sewer systems. *Water Research*, **126**, 88–100. doi: 10.1016/j.watres.2017.08.065.

Upton G. J. G. & Rahimi A. R. (2003). On-line detection of errors in tipping-bucket raingauges. *Journal of Hydrology*, **278** (1–4), 197–212. doi: 10.1016/S0022-1694(03)00142-2.

van Bijnen J. A. C. (2018). *The Impact of Sewer Condition on the Performance of Sewer Systems*. PhD thesis, Delft University of Technology, Delft, The Netherlands, 202 p. ISBN 978-94-6233-987-3.

Villez K. & Habermacher J. (2016). Shape anomaly detection for process monitoring of a sequencing batch reactor. *Computer and Chemical Engineering*, **91**(4), 365–379. doi: 10.1016/j.compchemeng.2016.04.012.

Wilcox R. R. (2016). *Introduction to Robust Estimation and Hypothesis Testing*, *4th edn*. Elsevier Academic Press, London (UK), 810 p. ISBN 978-0128047330.

Wongoutong C. (2020). Imputation for consecutive missing values in non-stationary time series data. *Advances and Applications in Statistics*, **64**(1), 87–102. doi: 10.17654/AS064010087.