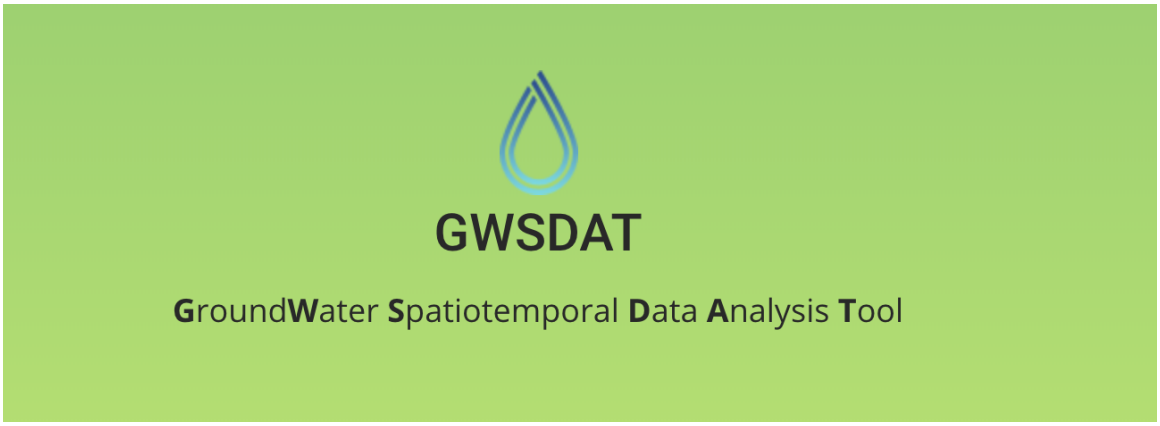


Version 4.0 User Manual

Wayne R. Jones (Wayne.W.Jones@shell.com)
Principal Data Scientist, Shell, London

Luc Rock (Luc.Rock@shell.com)
Soil and Groundwater Scientist, Shell, Amsterdam

Claire Miller (Claire.Miller@glasgow.ac.uk)
Marnie Low (Marnie.Low@glasgow.ac.uk)
Craig Alexander (Craig.Alexander.2@glasgow.ac.uk)
Adrian Bowman (Adrian.Bowman@glasgow.ac.uk)
School of Mathematics & Statistics, The University of Glasgow

16 January, 2021

Contents

Acknowledgements

The authors gratefully acknowledge the many different people who have willingly contributed their knowledge and their time to the development of GWSDAT.

The authors wish to express their gratitude to Ludger Evers and Daniel Molinari from the Department of Statistics, University of Glasgow, for their invaluable contributions to the statistical aspects of GWSDAT. Thanks also to Ewan Crawford for his assistance in the development of the original GWSDAT user interface.

We acknowledge and thank the R project for Statistical Computing and all its contributors without which this project would not have been possible.

A big thank you to Shell's worldwide environmental consultants for assistance in evaluating and testing GWSDAT. Thanks also to the Shell Year in Industry students Tess Brina, Rosemary Archard, Emma Toms, Stephanie Marrs and Rachel Stroud who spent a great deal of time using GWSDAT and making suggestions for improvements.

We thank our colleagues Matthew Lahvis, George Devaull, Matthijs Bonte, Hayley Thomas, Karina Cady from Shell Projects & Technology; HSE Technology: Soil & Groundwater and Philip Jonathan, Shell Chief Statistician, for their support, vision and advocacy of GWSDAT.

The original idea of GWSDAT was inspired by Marco Giannitrapani.

Introduction

The GroundWater Spatiotemporal Data Analysis Tool (GWSDAT) has been developed by Shell Global Solutions and the University of Glasgow to help visualise trends in groundwater monitoring data. It is designed to work with simple time-series data for solute concentration and ground water elevation, but can also plot non-aqueous phase liquid (NAPL) thickness if required. Spatial data is input in the form of well coordinates, and wells can be grouped to separate data from different aquifer units. The software also allows the import of a site basemap in GIS shapefile format. Trend and contour plots generated using GWSDAT can be exported directly to Microsoft PowerPoint and Word to expedite reporting.

GWSDAT version 3.0 can be operated on-line or downloaded to run locally. Access to the on-line version, and instructions for downloading the local version, are available at `gwsdat.net`.[1] The underlying statistical calculations and graphical output are generated using the open source statistical program R (R Development Core Team (2008)). More details on the statistical methods can be found in Section .

Potential applications where GWSDAT can add value (cost savings and reduction in environmental liabilities) through improved risk-based decision making and response include:

- Early identification of increasing trends or off-site migration.
- Evaluation of groundwater monitoring trends over time and space (i.e., holistic plume evaluation).
- Nonparametric statistical and uncertainty analyses to assess highly variable groundwater data.
- Reduction in the number of sites in long-term monitoring or active remediation through simple, visual demonstrations of groundwater data and trends.
- More efficient evaluation and reporting of groundwater monitoring trends via simple, standardised plots and tables created at the 'click of a mouse'.

Disclaimer: There is no warranty for the Program (GWSDAT), to the extent permitted by applicable law. SHELL, Affiliates of SHELL, the copyright holders and/or any other party provide the Program 'as is' without warranty of any kind, either expressed or implied, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The entire risk as to the quality and performance of the Program is with the LICENSEE. Should the Program prove defective, the LICENSEE assumes the cost of all necessary servicing, repair or correction.

[1] Version 2.2 of GWSDAT uses Microsoft Excel as the primary user interface and this version remains available from the *API* `https://www.api.org/oil-and-natural-gas/environment/clean-water/ground-water/gwsdat` and *Cl:aire* (`https://www.claire.co.uk/projects-and-initiatives/gwsdat`) websites.

Accessing GWSDAT

The simplest way to use GWSDAT is through the on-line version available at (gwsdat.net). This web site also has general information about the the software tool, help files and videos, and case studies of its use.

However, it will sometimes be helpful to be able to use GWSDAT without internet access. Also, datasets may sometimes need to be retained and analysed locally. In these circumstances, GWSDAT can be downloaded onto any computer. The instructions below describe how to do that.

GWSDAT uses the widely available, open source, statistical computing environment `R` (R Development Core Team 2008). This should be downloaded from `www.r-project.org` where versions for all major computing platforms are available. Installation is a very simple process. You may also find it convenient to install the *RStudio* 'front end' for `R`, freely available from `www.rstudio.com`. This manages some aspects of the `R` environment in a helpful way.

When `R` or `RStudio` is launched, one of the visible windows is a 'console'. GWSDAT is available as a package in `R` and this can be installed by typing the instruction

```
install.packages("GWSDAT")
```

in the console window. Note the capital letters, as `R` is case-sensitive. The package is retrieved from the `R` archive (`CRAN`), so an internet connection is required for this step. The package will then be installed locally. GWSDAT uses several other `R` packages and these will be installed at the same time. There may be a warning message about a mismatch between the version of `R` used to build the package and the version of `R` installed on your computer if this is not the most recent one, but this is unlikely to cause any difficulty. The installation step is required only once. GWSDAT will now be available on your computer at any time, with or without an internet connection.

To launch GWSDAT, issue the following two instructions in the console window:

```
library(GWSDAT)
launchApp()
```

The first instruction loads the package so that it can be used in the current session of `R`. The second instruction launches GWSDAT.

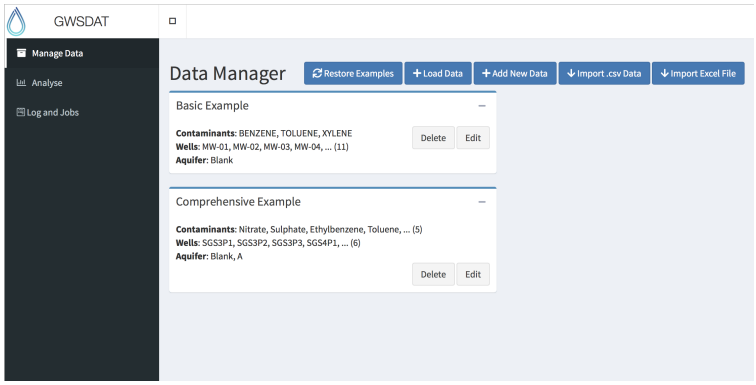
```
getwd()
```

```
## [1] "C:/Users/Wayne.W.Jones/GitHub/GWSDAT_User_ManualTemp2"
```

Excel installation instructions to be added

Using your own data

GWSDAT comes supplied with examples of groundwater monitoring data and these can be used to experiment with the tools for analysis and visualisation. If you would like to do this you can move immediately to Section~X of this manual. However, the aim of GWSDAT is of course to provide means of analysing user-supplied data and various options are supplied to allow this. These options are all available from the *Manage Data* page which is the point of entry when GWSDAT is launched but can also be accessed at any time from the *Manage Data* item on the menu bar on the left hand side of the window.



A *Basic Example* and a *Comprehensive Example* are supplied but the buttons at the top of the page allow other data to be entered, in different forms.

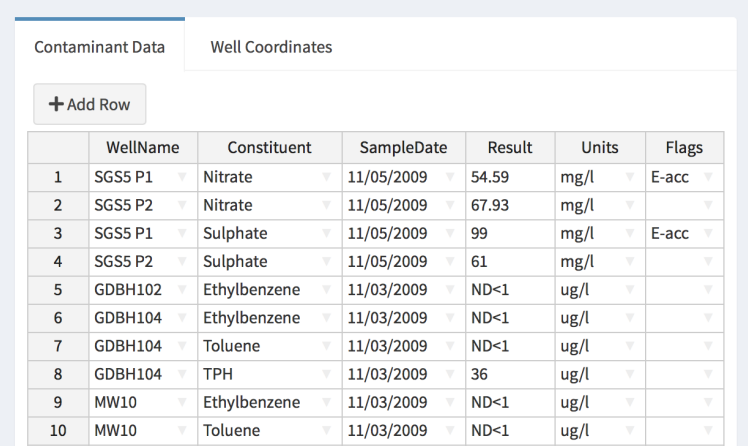
Restore Examples	This allows the two in-built examples to be restored at any point, should this be needed.
Load Data	Use the *Browse* button to loaded from a previous GWSDAT session.
Add New Data	This allows data to be entered manually into a spreadsheet (and saved for later use).
Import .csv Data	Use the *Browse* buttons to load separate 'Contaminant Data' and 'Well Coordinates' files, each in '.csv' form. One or more shapefiles may also be loaded.
Import Excel File	Use the *Browse* button to load an Excel spreadsheet containing both 'Contaminant Data' and 'Well Coordinates'. One or more shapefiles may also be loaded.

The details of data entry are described below but it is helpful first to describe the format of the data which is required. The essential information is in are two spreadsheets - one which contains the contaminant data, and a second one which gives the co-ordinates of the wells. There is also an option to use shape files to superimpose map information on spatial plots.

An example of the first spreadsheet is shown for the *Basic Example* in the screenshot below. The full spreadsheet can be viewed by clicking the *Edit* button for the *Basic Example* in the *Manage Data* page of GWSDAT. (When you wish to return to the *Manage Data* page, click the 'back-arrow' button in the top left hand corner of the *Edit Data* page.) The columns of the spreadsheet are explained in detail below.

Wayne advises against the use of the term 'contaminant' but this is the terminology used in the software.

Well Data



	WellName	Constituent	SampleDate	Result	Units	Flags
1	SGS5 P1	Nitrate	11/05/2009	54.59	mg/l	E-acc
2	SGS5 P2	Nitrate	11/05/2009	67.93	mg/l	
3	SGS5 P1	Sulphate	11/05/2009	99	mg/l	E-acc
4	SGS5 P2	Sulphate	11/05/2009	61	mg/l	
5	GDBH102	Ethylbenzene	11/03/2009	ND<1	ug/l	
6	GDBH104	Ethylbenzene	11/03/2009	ND<1	ug/l	
7	GDBH104	Toluene	11/03/2009	ND<1	ug/l	
8	GDBH104	TPH	11/03/2009	36	ug/l	
9	MW10	Ethylbenzene	11/03/2009	ND<1	ug/l	
10	MW10	Toluene	11/03/2009	ND<1	ug/l	

Each row of this table gives the details of a particular measurement. There should be no empty rows in the table.

- **WellName**: the name or identifier of the well (or soil boring) from which the sample was collected. Well names must be consistent and unique. For example, `MW-1' and`MW1' will be treated as different wells.
- **Constituent**: the name of the solute, for example `Benzene` or `Toluene`. The name of a solute must be consistent and unique. The name `GW` is reserved for measurements of groundwater level and `NAPL` is reserved for NAPL thickness data. There are further details on this below.
- **SampleDate**: the date at which the well was sampled (not the date the results were returned from laboratory analysis). Please use a calendar date format, the preferred format is `dd/mm/yyyy`. Do not include a time of day.
- **Result**: the value of the measurement made. This will be a solute concentration, a groundwater level or a NAPL thickness, as specified in the *Constituent* column.

For solute concentrations, non-detect values should be entered as either `<X` or `ND<X`, where `X` is the detection limit specified by the laboratory. For example, if the detection limit is $100\mu g/l$ then either `<100` or `ND<100` is acceptable. The non-detect threshold value must be specified so `ND` on its own is not permissible. In the absence of known detection limits, a sensible value must be substituted. This could be the lowest measured value for the solute in the dataset.

Groundwater level data is entered as an elevation above a common datum, such as metres or feet above sea level or some other common reference height. All groundwater measurement entries should have the same units, such as metres or feet, and the `Constituent~ field should be set to`GW'. In the presence of NAPL, please ensure that the groundwater level has been corrected for NAPL density.

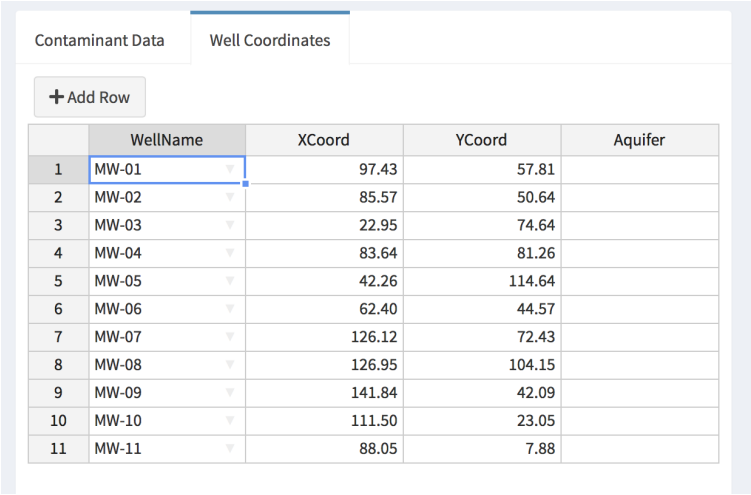
For NAPL thickness data, all entries should have the same units, such as feet or metres, and the `Constituent' field should be set to`NAPL'. If no NAPL is present, do not add a NAPL entry with zero thickness – simply do not add an entry.

Where NAPL is recorded in soil borings that do not reach the water table, the NAPL thickness should be entered as zero.
Well location markers for soil borings or wells where NAPL has been recorded are highlighted in red.

- **Units**: the units of the *Result*.This must be specified for each entry, with consistent units for each type of *Constituent*. Solute concentration data can either be `mg/l` or μ`g/l`. Groundwater level and NAPL thickness data should be set to one of `mm`, `cm`, `metres`, `inches`, `feet` or `level`.
- **Flags**: these are available to modify the way in which certain types of data are handled by the software. The `E-Acc` (Electron Acceptor), `NotInNAPL` and `Redox` flags are used to identify results which are to be

omitted in the event that the user activates the NAPL substitution function; see the Section on NAPL handling. Note that it is only necessary to flag one data row in this way for all rows containing that constituent to be excluded from NAPL. The `Omit` flag can be used to exclude any individual data row from the analysis.

Well co-ordinates



	WellName	XCoord	YCoord	Aquifer
1	MW-01	97.43	57.81	
2	MW-02	85.57	50.64	
3	MW-03	22.95	74.64	
4	MW-04	83.64	81.26	
5	MW-05	42.26	114.64	
6	MW-06	62.40	44.57	
7	MW-07	126.12	72.43	
8	MW-08	126.95	104.15	
9	MW-09	141.84	42.09	
10	MW-10	111.50	23.05	
11	MW-11	88.05	7.88	

An example of a *WellCoordinates* table is shown above. This can be viewed by clicking on the *Well Coordinates* tab in the *Edit Data* page for the *Basic Example*. This table is used to store the co-ordinates of the groundwater monitoring wells or soil borings. This information is essential for spatial and spatiotemporal analyses. For most of the purposes of GWSDAT modelling, it is only the relative distances between wells which are important. This means any arbitrary cartesian co-ordinate system can be used as long as well co-ordinate values have an aspect ratio very close to 1 – in other words, a unit in the x-coordinate corresponds to the same distance as a unit in the y-coordinate. Hence, well co-ordinates can be measured directly from a map, or given in easting and northing, etc.

The columns of the spreadsheet are described below. There should be no empty rows in the table.

- **WellName**: the name or identifier of the well or soil boring. Well names must be identical to those specified in the *Contaminant Data* table. On a point of detail, it is better to name wells using the convention of `MW-01` rather than `MW1` so that plots in GWSDAT are correctly ordered.
- **XCoord**: the x-coordinate of the well.
- **YCoord**: the y-coordinate of the well.
- **Aquifer**: an optional column which allows wells or soil borings to be associated with particular subsurface features such as aquifers or sub-strata, in case these data need to be modelled separately. The name (maximum of 8 characters) of the aquifer or sub-stratum can be entered or a letter (A–G) can be used. The aquifer field can also be used to partition the dataset from a large site, in the event that multiple unrelated plumes are present or wells are grouped into well separated clusters. When analysis begins, the user is asked to select an aquifer (subsurface feature) to analyse. Plots generated using data associated with particular subsurface features have the feature name appended to the title.

Notice the *Coordinate Unit* button in the box to the left of the spreadsheet. This enables the spatial units to be specified as metres or feet, through a drop-down menu.

Shape Files

A further possible component of the data is a site map in the form of a *shape file*. This is optional but can be very useful in superimposing map information on spatial plots. Examples will be given later in the manual.

The paragraphs below need to be updated.

A 'shapefile' is actually a collection of several files, typically created using ARC-GIS.[2] The *Shape Files* table consists of a single column where each entry is the name of a file (including its path) containing site plans. These must be in shapefile format. Filenames can be entered manually or, in interactive mode, the `Browse ...' can be used. Only the location of the main shapefile (file ending with a`.shp' extension) needs to be specified in this table - the associated data files (such as `.dbf`, `.sbn`, `.sbx`, `.shx`) will be picked up automatically, but they must be in the same folder. An example is given in the Figure below.

It is possible to overlay multiple shapefiles up to a maximum of seven. There should be no empty rows in the table.

Entering your own data

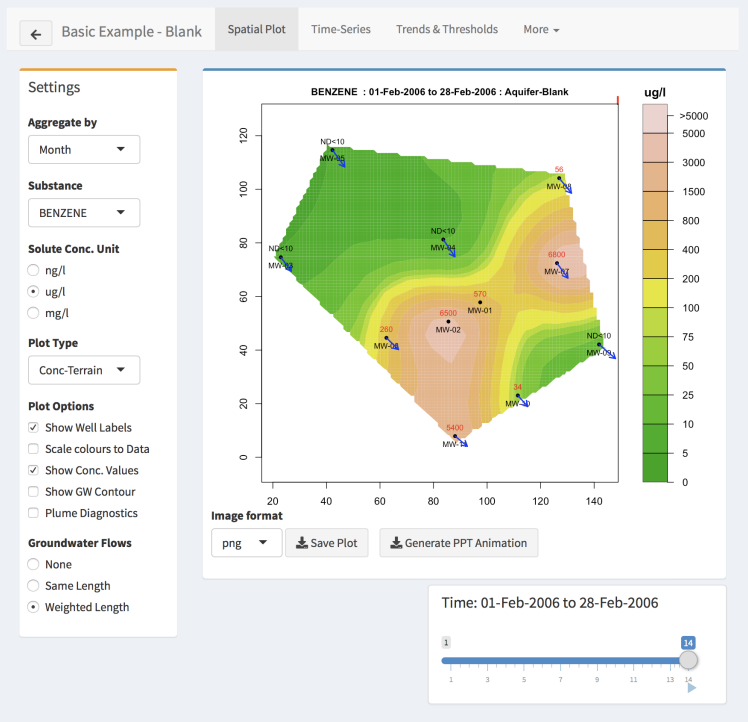
The various ways in which data can be entered are described below. The easiest option is likely to be to create the spreadsheets in a format you are familiar with, such as *Excel*, and then load the files into GWSDAT.

Import Excel File	The file to be loaded should contain two spreadsheets labelled 'Contaminant Data' and 'Well Coordinates' and have the structure described above. An example spreadsheet is available at *web address at gwsdat.net*.
Import .csv Data	The data can be exported from *Excel() into '.csv' format. In this case there should be separate '.csv' files for the *Contaminant Data* and *Well Coordinates*. Click on the *Browse* button to select these files. For convenience, GWSDAT gives options for the *Column separator* and the *Quote for Character Strings* at the foot of this page. Click on the *Import|* button to read the data.
Add New Data	This creates blank spreadsheets with the required structure. Entries can by typed in directly or copied and pasted from another source. There is an *+Add Row* button and right clicking gives further options for adding and removing rows.
Load Data	This allows data which has previously been saved from GWSDAT to be loaded again.

[2] See `http://en.wikipedia.org/wiki/Shapefile` for more information on this format.

Analysing the data

GWSDAT is designed to produce informative visualisations of groundwater monitoring over space and time. To illustrate this, we will use the *Basic Example*. If you are still in the detailed pages of the *Manage Data* section, click on the 'back-arrow' in the top left hand corner to return to the main *manage Data* page and then click on *Analyse* on the left had sidebar. You will then be asked to specify which dataset you would like to analyse. In addition to the standard examples, any other datasets you have created will be listed here too. For the moment, click on the *Select* button for the *Basic Example'*. You should now see a screen similar to the image below.



The tabs at the top of this page give access to several different forms of analysis. These are described in detail below.

Spatial plot

A key feature of GWSDAT is the ability to produce estimates of contaminant concentrations over space and time simultaneously. This gives a more effective method of analysis than the examination of concentration maps at isolated time points, or of time trends at isolated locations. The simultaneous use of information over space and time allows estimates at particular locations and times to 'borrow strength' from neighbouring data. Use the slider at the foot of the page to explore how the estimates of Benzene concentration change across the month of October. Note that the slider box at the foot of the page can be moved to any convenient position by clicking and dragging with the mouse. The 'Play' symbol (forward-arrow) in the bottom right hand corner of the slider activates a 'movie', which can be paused by pressing the button again.

The *Settings* box gives control over many aspects of the display.

- **Aggregate by** provides a drop-down menu which allows the temporal resolution to be altered (Day, Month, Quarter, Year). This gives different perspectives on the trends present in the data. Of course, the highest resolution (here, day) will be constrained by the frequency of data collection.
- **Substance** allows different contaminants to be inspected, if these have been recorded in the dataset. Use the drop-down menu to select this.

- **Solute Conc. Unit** allows the units to be changed.
- **Plot Type** allows the contour colours to be changed or the display to be focussed on the wells through the size and colour of plotted circels.
- **Plot Options** gives control over a variety if different annotations which add detailed information to the map.
- **Groundwater Flows** gives control over the display of this information.

Time Series

Trends & Thresholds

More

Reporting

In addition to providing visualisation of groundwater monitoring data, GWSDAT is also able to export displays in a variety of file formats, for inclusion in reports. The map displays in the *Spatial Plot* tab of the *Analyse* section are used to illustrate this. At the foot of the main map display there are three buttons.

- **Image format** provides a drop-down menu of the file type used when a spatial plot at a particular time point is created. The available file types `png`, `jpg`, `pdf`, `ps` and `pptx`.
- **Save Plot** creates and downloads a file in the specified format.
- **Generate PPT Animation** creates a downloads a sequence of plots which display concentration maps across the whole time course. Paging through these slides provides a very simple but effective means of animation.

The statistical models behind GWSDAT

Spatiotemporal Solute Concentration Smoother

The spatiotemporal solute concentration smoother is estimated using a non parametric regression technique known as Penalised Splines (P-Splines). It is beyond the scope of this document to give a full and detailed explanation of this technique here. However, the following outlines some of the most important aspects for the purposes of GWSDAT. For a more detailed explanation the reader is referred to (Eilers and Marx (1992)) and (Eilers, Rijnmond, and Marx (1996)).

Let y_i be the solute concentration at $\boldsymbol{x_i} = (x_{i1}, x_{i2}, x_{i3})$ where x_{i1} and x_{i2} stand for the spatial coordinates of the well and x_{i3} represents the corresponding time point for the *i-th* observation with $i = 1, \ldots, n$. We start by modelling the solute concentration as

$$y_i = \sum_{j=1}^{m} b_j(\boldsymbol{x_i})\alpha_j + \epsilon_i \tag{1}$$

where the b_j, $j = 1, \ldots, m$ are m functions (known as *basis functions*) conveniently chosen to achieve smoothness (generally a particular kind of polynomial of order 3). The first term in equation (1) is a linear combination of the basis functions b_j, each evaluated at $\boldsymbol{x_i}$, and aims at capturing the deterministic part of the *i-th* observation, generally known as 'signal'; the second term, ϵ_i, accounts for the variability in the measurement due to randomness and is usually termed as 'noise'. The behaviour of ϵ_i is described in terms of a convenient probabilistic model; such a model guarantees that the value of ϵ_i fluctuates around zero conveying the idea that we do not expect to make any systematic error in the measurement. This model also comprises the notion that the expected spread of ϵ_i is given by σ^2, %a non-negative parameter σ; its squared %value is known as the *variance* of the random component ϵ_i. By using the matrix notation

$$\boldsymbol{B}(\boldsymbol{x}) = \begin{pmatrix} b_1(x_1) & \cdots & b_j(x_1) & \cdots & b_m(x_1) \\ b_1(x_2) & \cdots & b_j(x_2) & \cdots & b_m(x_2) \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ b_1(x_i) & \cdots & b_j(x_i) & \cdots & b_m(x_i) \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ b_1(x_n) & \cdots & b_j(x_n) & \cdots & b_m(x_n) \end{pmatrix}$$

equation (1) can be written in a more compact fashion as $\boldsymbol{y} = \boldsymbol{B}(\boldsymbol{x})\boldsymbol{\alpha} + \boldsymbol{\epsilon}$. Because, as mentioned earlier, we expect the ϵ_i's to oscillate around zero, a sensible choice for the regression parameters $\boldsymbol{\alpha}$ is the one that minimises the norm of the vector $\boldsymbol{\epsilon}$ defined as $\mathrm{S}(\boldsymbol{\alpha}) = \|\boldsymbol{\epsilon}\|^2 = \|\boldsymbol{y} - \boldsymbol{B}(\boldsymbol{x})\boldsymbol{\alpha}\|^2$. A large value of basis functions is generally chosen to allow the model to capture most of the signal. The downside of this approach is that it tends also to overfit, that is to fit the noise in the observations, with the consequent loss of smoothness. To overcome this hurdle, the objective function %to be optimised $\mathrm{S}(\boldsymbol{\alpha})$ is modified with the addition of a term that penalises the lack of smoothness of the fit.

The objective function now takes the form $\mathrm{S}(\boldsymbol{\alpha}) = \|\boldsymbol{y} - \boldsymbol{B}(\boldsymbol{x})\boldsymbol{\alpha}\|^2 + \lambda\|D\alpha\|^2$ where λ is a non-negative smoothing parameter and D is the $(m-2) \times m$ matrix

$$\boldsymbol{D} = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & & \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 \end{pmatrix}$$

The additional term in the objective function

$$\|D\alpha\|^2 = (\alpha_1 - 2\alpha_2 + \alpha_3)^2 + \ldots + (\alpha_{m-2} - 2\alpha_{m-1} + \alpha_m)^2$$

controls the smoothness of the fit by applying penalties over adjacent coefficients. By minimising the new objective function for a given value of λ, we obtain the least squares estimator of the parameters

$$\hat{\boldsymbol{\alpha}} = \left(\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D}\right)^{-1} \boldsymbol{B}'\boldsymbol{y}.$$

Consequently, the fitted values are given by:

$$\hat{\boldsymbol{y}} = \boldsymbol{B}\hat{\boldsymbol{\alpha}} = \boldsymbol{B}\left(\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D}\right)^{-1} \boldsymbol{B}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$$

When $\lambda = 0$, the expression for the estimator of the parameters $\{\ \hat{\boldsymbol{\alpha}}\ \}$ boils down to the classical solution in linear models theory. As $\boldsymbol{\lambda} \rightarrow \boldsymbol{\infty}$, the fitted function tends to a linear function. The Figure below shows the effect of penalisation: it forces the coefficients to yield a smooth pattern. The fitting process of a function using B-Splines is pictured with and without penalisation, together with the basis functions (the columns of the $\boldsymbol{B}$ matrix). The left plot results from not penalising ($\boldsymbol{\lambda = 0}$) the term in the objective function that accounts for the smoothness; it can be noticed that it yields a rather wiggly regression function. In the right plot, a suitable choice for $\boldsymbol{\lambda}$ constrains the optimisation method to find values for the coefficients $\hat{\boldsymbol{\alpha}}$ which result in a smoother regression curve.

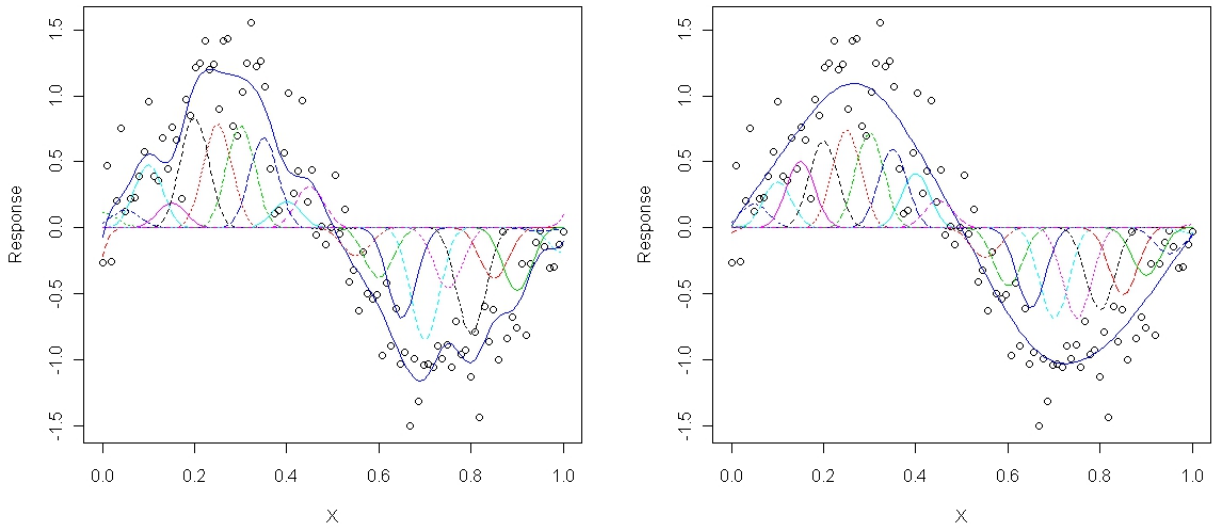


fig.cap: Curve based on 20 nodes in the basis, without penalisation (left), with penalisation (right).

Prior to fitting the regression coefficients $\boldsymbol{\alpha}$ the observed solute concentration values are natural log transformed. This avoids the possibility of predicting negative concentration values and also helps the model cope with data which often spans several orders of magnitude. Furthermore, the uncertainty in the measured concentrations can reasonably be expected to be proportional to the magnitude of the value, e.g. the uncertainty around a measured value of 10ug/l would be expected to be very much less than the uncertainty surrounding a measured value of 10000ug/l. The natural log transformation stabilises the variance.

The choice of the penalisation parameter $\boldsymbol{\lambda}$ is a crucial matter as a too small value would result in 'overfitting' (tracking the noise) whereas an extremely large value would lead to 'underfitting' (producing a flat estimated function as a result of loss of signal). 'several criterias have been proposed, such as those described by (Hurvich and Simonoff 1998) and (Wood 2006), but we tackled the issue by a *Bayesian* approach; see (Denison et al. 2002), (Raftery, Madigan, and Hoeting 1997) and (Wood 2011).

Under this paradigm, $\boldsymbol{\lambda}$ is not considered to be a fixed unknown quantity to be estimated but rather a random variable whose value may vary within a given range. This behaviour is described in probabilistic terms which assign a measure of confidence or *probability* to each of the values $\boldsymbol{\lambda}$ may take on.

The Bayesian framework allows to compute the probability that the random variable $\boldsymbol{\lambda}$ may take a particular value, conditional on the fact that $\boldsymbol{y}$ has already been observed. This probability, indicated as $\boldsymbol{f(\lambda|y)}$, is known as the *posterior distribution* of $\boldsymbol{\lambda}$.

Bayes' rule states that $\boldsymbol{f(\lambda|y) \propto f(y|\lambda)f(\lambda)}$ where $\boldsymbol{\propto}$ stands for "proportional to". $\boldsymbol{f(y|\lambda)}$ is known as the *likelihood function* and expresses the conditional probability of observing data $\boldsymbol{y}$, given that the true value of the parameter is $\boldsymbol{\lambda}$; $\boldsymbol{f(\lambda)}$ is known as the *prior distribution* of the random variable $\boldsymbol{\lambda}$ and comprises our prior beliefs on its uncertainty.

The optimal value of $\boldsymbol{\lambda}$ is the one that maximises the posterior distribution and is computed using numerical methods.

References

Denison, D., C. Holmes, B. Mallick, and A. Smith. 2002. *Bayesian Methods for Nonlinear Classification & Regression.* John Wiley & Sons, New York.

Eilers, Paul H. C., Dcmr Milieudienst Rijnmond, and Brian D. Marx. 1996. "Flexible Smoothing with B-Splines and Penalties." *Statistical Science* 11: 89–121.

Eilers, P., and B. Marx. 1992. *Generalized Linear Models with P-Splines in Advances in Glim and Statistical Modelling (L.fahrmeir et Al. Eds.).* Springer, New York.

Hurvich, C., and J. Simonoff. 1998. "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60: 271–93.

Raftery, A., D. Madigan, and J. Hoeting. 1997. "Bayesian Model Averaging for Linear Regression Models." *Journal of the American Statistical Association* 92: 179–91.

R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Wood, S. N. 2006. *Generalized Additive Models - an Introduction with R.* Chapman & Hall/CRC.

———. 2011. "Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73: 3–36.