FAIRWAY
Farm System Management and Governance
for Good Water Quality and Drinking Water Supplies

FAIRWAY REPORT series

(Short note for the) Database containing harmonized datasets

Authors: Marc Laurencelle, Nicolas Surdyk, Matjaž Glavan, Birgitte Hansen, Claudia Heidecke, Hyojin Kim, Susanne Klages

Date: 31 May 2021
Version: 1
Series: Deliverables

Deliverable 3.3

This report was written in the context of the FAIRWAY project

www.fairway-project.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 727984

DOCUMENT SUMMARY	
Project Information	
Project Title	Farm systems management and governance for producing good water quality for drinking water supplies
Project Acronym	FAIRWAY
Call identifier	H2020-RUR-2016-2
Topic	RUR-04-2016 Water farms – improving farming and its impact on the supply of drinking water
Grant agreement no	727984
Dates	2017-06-01 to 2021-05-31
Project duration	54 months
Website addresses	www.fairway-project.eu www.fairway-is.eu
Project coordination	Stichting Wageningen Research, NL
EU project representative & coordinator	Lara Congiu (REA)
Project scientific coordinator	Gerard Velthof
EU project officer	Marta Iglesias (DG Agri)
Deliverable information	
Title	(Short note for the) Database containing harmonized datasets
Authors	Marc Laurencelle, Nicolas Surdyk ; Matjaž Glavan, Birgitte Hansen, Claudia Heidecke , Hyojin Kim, Susanne Klages
Author email	m.laurencelle@brgm.fr
Deliverable number	D3.3
Work package	WP3
WP Lead	Nicolas Surdyk
Type and dissemination level	Database and report
Editor	Gerard Velthof
Due date	2021-05-31
Publication date	2021-05-31
Copyright	© FAIRWAY project and partners

Version History		
Number & date	**Author**	**Revision**
Final version 1, 31.05.2021	Laurencelle, Surdyk et al.	Other co-authors & Gerard Velthof

CONTENTS

1. SUMMARY

This document presents the development of the database as due for FAIRWAY project deliverable 3.3 (DL3.3) as well as the main features of this database. This database, which was a part of WP3 task 3, was used to calculate indicators selected in the previous task.
The database development has mainly been driven by existing datasets coming from each of the case studies of the FAIRWAY project. The database contains near 390,000 rows of data from the 13 case study sites, about >65 parameters and >500 sub-parameters.
One of the challenges throughout the task of database development has been to find ways to harmonize as much as possible the datasets obtained from those various sources.

This document first describes the database in terms of its general structure (Chapter 3), development process (Chapter 4), and detailed structure (Chapter 5). Possible uses of the database are then mentioned along with examples of some interesting data series and instructions on using the database efficiently. Finally, the major problems and limitations encountered throughout this work are discussed.

Some major challenges identified throughout this work are that:

1) Definitions of a ‘boundary’ are different from the pressure and state perspectives. The catchment area defines the hydrogeological boundary, but the agricultural boundary is an administrative boundary (at least are displayed as that). Moreover, there is generally a lag time (delay) between pressure and state indicators. Consequently, pressure data and state data do not overlap in most cases, and thus they cannot be linked directly.

2) Because of the difference in those definitions, the scale of the collected data is also different. The state data (mainly hydrogeochemical data on water quality) can be point or catchment scale while the pressure data is ideally at the field plot scale but actually most often at administrative levels, i.e., municipal, regional, or even national level.

3) Therefore, it is time-consuming to collect these large sets of data and process the data to a comparable form between state and pressure for a case study site and among the case study sites.

The Database is hosted on the restricted area part of the FAIRWAY server. The Database can thus be accessed only once logged in, via this links:

https://fairway-is.eu/index.php/farm-management/workpackages/harmonised-indicator-database/340-indicator-database

2. INTRODUCTION

The overall aim of the Fairway project is to review approaches for the protection of drinking water resources against pollution by pesticides and nitrate, and to identify and further develop innovative measures and governance approaches for a more effective drinking water protection, together with relevant local, regional and national actors. Fairway overall is structured in eight work packages.

Work Package 3 (WP3), entitled “Monitoring and indicators”, has among its tasks and expected deliverables the preparation of harmonized datasets for water quality monitoring of drinking water resources, and the development of a readily usable database from these harmonized datasets. The database containing harmonized datasets constitutes deliverable number 3.3 of FAIRWAY’s WP3.

A large range of environmental indicators has been considered for this purpose. Yet the focus has mainly been on indicators related to the monitoring of nitrate and pesticide application, transport and fate in the hydrogeological system and in drinking water (Klages et al., 2018).

Development of the database has mainly been driven by existing datasets coming from each of the 13 case studies of the Fairway project, as well as various other datasets on national or EU level that could be found online. Mining of monitoring data on agricultural pressure and drinking water quality resources and most of the site-specific data was conducted with the support of the partners involved in the 13 case studies of Fairway. One of the challenges throughout the task of database development has been to find ways to harmonize as much as possible the datasets obtained from those various sources.

This document first describes the database in terms of its general structure, development process, and detailed structure (these paragraphs about the detailed structure are a short handbook on the database utilisation). Possible uses of the database are then mentioned along with examples of some interesting data series and instructions on using the database efficiently. Finally, the major problems and limitations encountered throughout this work are discussed.

The Database is hosted on the restricted area part of the FAIRWAY server. The Database can thus be accessed only once logged in, via this link:

https://fairway-is.eu/index.php/farm-management/workpackages/harmonised-indicator-database/340-indicator-database

3. General structure of the database

3.1 Database Architecture

A simple architecture was chosen for the database:

- An Excel file consisting of one sheet of data and one sheet of summary per case study.
- A zipped folder containing GIS files from all case studies.

The choice of developing an “Excel database”, rather than a more rigid-structured SQL relational database, was made soon after inspection of the datasets provided by Fairway case-study leaders, which revealed a very large disparity in the types and amounts of data provided between the case studies. That is indeed the main reason why it is considered more appropriate to present the data in separate data tables (Excel “sheets”) and separate subfolders of GIS data for each case study. Other reasons notably include the advantages that an Excel database is easier to disseminate, use, and modify locally for instance to add more recent monitoring data for a given case study. That being said, it remains possible to merge all data sheets into one large table, thanks to the uniform scheme used for the main columns in all Case Study data sheets (more details in Chapters 5 and 6).

3.1.1 Overall structure of the Excel database

The Excel database contains all “tabular” (non-GIS) data related to the 13 case studies of Fairway that was gathered for the purposes of WP3. It is structured as one “data sheet” and one “summary sheet” per case study. Specific names are assigned to the sheets to indicate the case study they are related to:

- The “**data sheets**” are named as follows: CS_[case study number] [abbreviated country name](_[suffix]). Moreover, a few additional characters (suffix) are appended to the abbr. country name to better distinguish sites better when there are more than one CS for the same country. For instance, datasheet “CS_4 FR” contains the data from case study #4 from France (one site only), while datasheets “CS_1 DK_it” and “CS_2 DK_a” respectively contain the data from the “Island Tunoe” and “Aalborg” case studies in Denmark.
- The “**summary sheets**”, which list the available parameters that are available for a given CS, are named simply by inserting “list_” before the name of its corresponding datasheet. For instance, the summary sheet “list_CS_1 DK_IT” lists the parameters that are available in the datasheet “CS_1 DK_IT” for the “Aalborg” case study from Denmark.

Figure 1 shows screenshots of the tabs for the first (a) and last (b) summary and data worksheets of the Excel database.

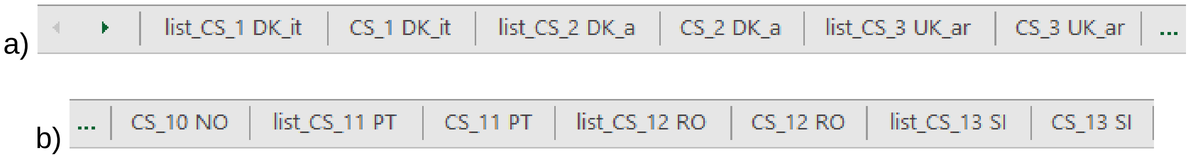
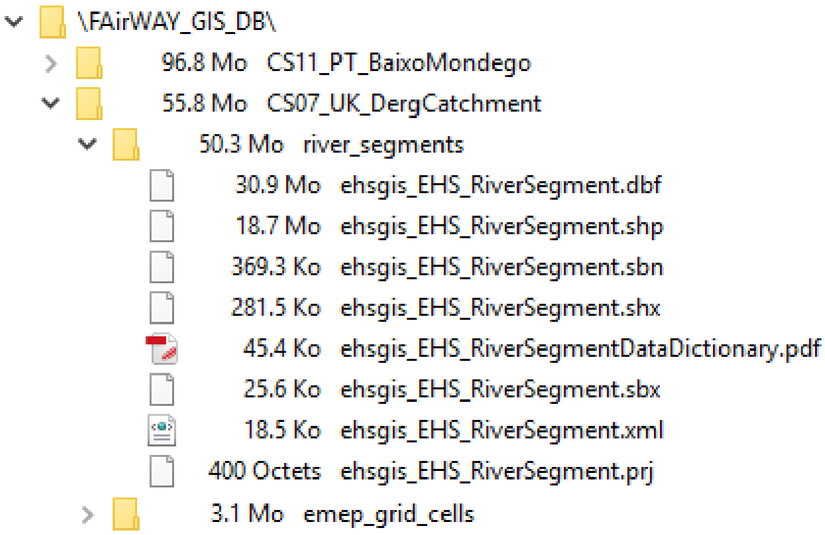


Figure 1: Preview of the names of the first (a) and last (b) summary and data sheets in the Excel database

3.1.2 Organization of the GIS data files

In parallel to the Excel database, a zipped folder contains all “GIS” data gathered for the purposes of WP3[1]. The GIS files are grouped in subfolders, by case study, and then by keywords describing the nature of the spatial data. For instance, a GIS file delineating river segments within the area of interest of Case Study #7 is stored in a subfolder “river_segments” placed in the case study’s subfolder “CS07_UK_DergCatchment” (Figure 2). Moreover, basic QGIS[2] project files (.qgz) offer an easy and straightforward way to explore the GIS data from each case study.



```
\FAirWAY_GIS_DB\
      96.8 Mo  CS11_PT_BaixoMondego
      55.8 Mo  CS07_UK_DergCatchment
         50.3 Mo  river_segments
            30.9 Mo  ehsgis_EHS_RiverSegment.dbf
            18.7 Mo  ehsgis_EHS_RiverSegment.shp
           369.3 Ko  ehsgis_EHS_RiverSegment.sbn
           281.5 Ko  ehsgis_EHS_RiverSegment.shx
            45.4 Ko  ehsgis_EHS_RiverSegmentDataDictionary.pdf
            25.6 Ko  ehsgis_EHS_RiverSegment.sbx
            18.5 Ko  ehsgis_EHS_RiverSegment.xml
         400 Octets  ehsgis_EHS_RiverSegment.prj
          3.1 Mo  emep_grid_cells
```

Figure 2: Preview of the organization of the GIS data files (partial example)

3.1.3 Linking between the Excel database and the GIS data

The vast majority of the tabular data in the Excel database properly refers to a geographical object. The information required to determine which spatial feature a tabular data belongs to is provided in the Excel database and consists of:

(In the case study’s datasheet)

- The text value in the field “Sampling_Point_ID”, which gives the unique identifier to the geographical object that was sampled (or characterized by other means) to produce the value that is reported in the “Results” field (of a given row);

(In the case study’s summary sheet)

- The “GIS_Related_Files” column, which gives the relative path and name of the GIS file containing spatial features related to the parameters, and;
- The “GIS_Key_Field” column, which specifies the name of the field of the GIS file’s attribute table that should contain the text IDs, for a given parameter;
- (And the optional “GIS_Key_Value” column, which is also required in some cases to target a specific spatial feature of the GIS file.)

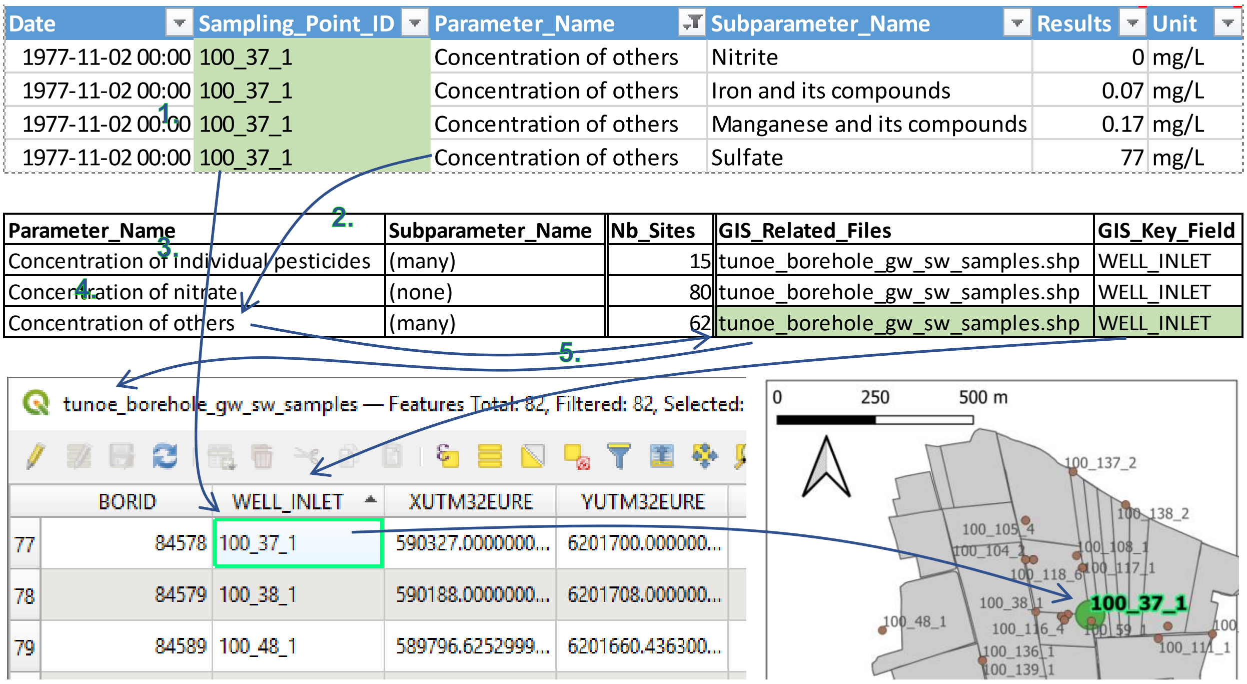
[1] That zip file must first be unzipped (decompressed) to a local folder or server before use, in most cases.
[2] QGIS is a free and open source geographic information system. For more information: https://qgis.org/

The steps one needs to follow to trace back the link from a “Results” tabular data to its corresponding geographical object consist of (Figure 3):

1. Finding the row describing the parameter in the Summary sheet;
2. Reading the “GIS_Related_Files” and “GIS_Key_Field” column values of that row;
3. Opening the GIS file(s) corresponding to [“GIS_Related_Files”]…;
4. Finding the spatial feature in the opened GIS file(s), where the attribute [“GIS_Key_Field”] is equal to [“Sampling_Point_ID”]. That can be done in several ways depending on the GIS software that is used. In a manual approach, it can be carried out using selection/filtering tools with the attribute table of the opened GIS file visible on screen;
5. Viewing the result of the feature selection/filtering on a map (if relevant).

Note that once the linking process has been explored and tested for the parameters of interest, it can be automated quite easily in a custom computer program. For instance, if one wants to relate all tabular data of a given case study to their corresponding geographical objects (from different GIS files). In this process, links will be made between the content of the Excel sheets of the case study of interest and the attribute tables of the relevant GIS files. Only in the end, the spatial features themselves are considered, for instance, to get their X and Y coordinates and/or calculated areas (in case of polygon features).

Note that linking the data of interest to their corresponding geographical objects is not always necessary, in practical contexts. It is especially the case when catchment-scale data is considered. Generally, one may select and extract the tabular data from a couple of related parameters in the Excel database, calculate general statistics, and then perform some analyses (such as trend analysis, cross correlation analysis, etc.) without having to worry about the actual location of the sampling sites. Nevertheless, even in such cases, one should still ensure that the data selected for those analyses are consistent in terms of sampling sites, which can be done by carefully checking the “Sampling_Point_ID” identifiers for the selected rows.



Date	Sampling_Point_ID	Parameter_Name	Subparameter_Name	Results	Unit
1977-11-02 00:00	100_37_1	Concentration of others	Nitrite	0	mg/L
1977-11-02 00:00	100_37_1	Concentration of others	Iron and its compounds	0.07	mg/L
1977-11-02 00:00	100_37_1	Concentration of others	Manganese and its compounds	0.17	mg/L
1977-11-02 00:00	100_37_1	Concentration of others	Sulfate	77	mg/L

Parameter_Name	Subparameter_Name	Nb_Sites	GIS_Related_Files	GIS_Key_Field
Concentration of individual pesticides	(many)	15	tunoe_borehole_gw_sw_samples.shp	WELL_INLET
Concentration of nitrate	(none)	80	tunoe_borehole_gw_sw_samples.shp	WELL_INLET
Concentration of others	(many)	62	tunoe_borehole_gw_sw_samples.shp	WELL_INLET

tunoe_borehole_gw_sw_samples — Features Total: 82, Filtered: 82, Selected:

	BORID	WELL_INLET	XUTM32EURE	YUTM32EURE
77	84578	100_37_1	590327.0000000...	6201700.000000...
78	84579	100_38_1	590188.0000000...	6201708.000000...
79	84589	100_48_1	589796.6252999...	6201660.436300...

Figure 3: Illustration of the link between the Excel database and the GIS data (example with 1 spatial point feature)

4. DEVELOPMENT PROCESS

Data collection was carried out first through surveys sent to the leaders of the 13 case studies. The Case Study leaders found data of various types, quality, quantity and formats, and sent us (i.e. the persons involved in WP3 of Fairway) those datasets. An Excel workbook template with detailed instructions was used for this purpose, to propose a uniform data exchange format for all case studies. Yet, many of the datasets we obtained from the Case Study leaders did not follow the template for various reasons. These “raw” datasets could still be processed and added to the database, in most cases. However, as expected, it was much easier and more time-efficient to integrate datasets prepared by the Case Study leaders using the template, versus “raw” datasets, which required special handling.

In 2020 and 2021, the Case Study leaders were contacted again to invite them to send additional data (for the CS that could collect usable data during the first surveys and were already present in the database) or all data (for the case studies that could not send anything earlier).

In parallel, we explored European and national databases online to find relevant data that could be added to the Excel database to enhance a Case Study data sheet when the Case Study leaders could not provide that type of information. This allowed us to add some interesting parameters to the database for some of the case studies, notably: Concentrations of nitrate from reports on water quality published on the EEA-website[3]; Atmospheric deposition of nitrogen from the EMEP MSC-W model[4]; National-average application rates of a given pesticide in Northern Ireland[5], etc. This was done mostly to test the feasibility and relevance of using very large-scale databases; it was not exhaustive. Overall, we found that the content of those databases was often hard to exploit either due to the data formats (too raw, unclear columns or metadata) or to the aggregated nature of the data (often averaged over one to several years and sometimes also over a large area). This is further discussed in Chapter 7.

In the end, the amounts of data that could be added to the database vary considerably between the 13 case studies (from <100 to >100,000 data rows: see Table 1. One of the main limiting factors to the effective availability of data for the purposes of WP3 is protecting personal data enforced by the new (May 2019) General Data Protection Regulation (GDPR) of the EU. In some case studies, the GDPR indeed made it difficult to get access to data, and more especially to plot-scale data (see Fairway Delivrable WP3 3.2 and §7.1).

We thus received, explored, and processed the data provided by Case Study leaders (along with the data from other sources) to gather it all in the database under development. Data reformatting was a major and time-consuming task. The provided data came from various sources and thus had many different formats, data types, and more or less accompanying information. The work to do was important even when the “Excel workbook template” was used properly by the Case Study leader, as there are several ways to copy-paste data from various sources during data collection, and because of the very varied formats of the raw data originally collected by the Case Study leaders, among other reasons.

Data cleaning was done in parallel to data reformatting, when weird/impossible/unusable values were detected in a dataset (e.g., a -999 numeric value for a parameter reporting daily precipitation rate in mm, or an “ND” text value in some numerical parameter). Those values, in most cases, were deleted.

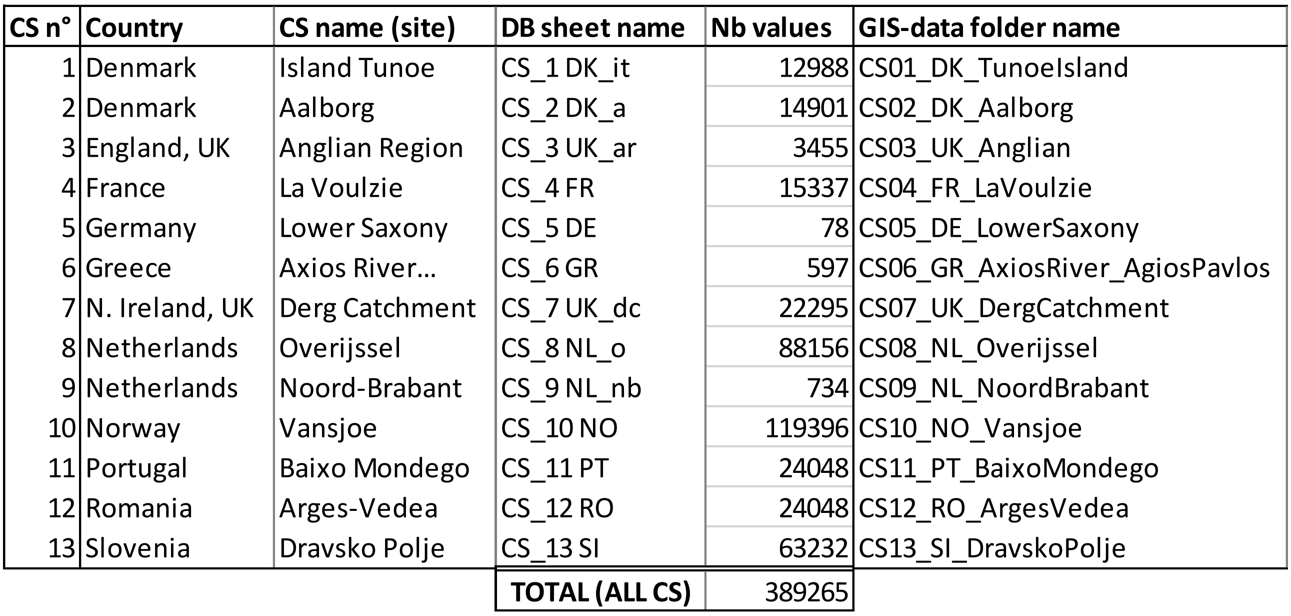
[3] EEA – Eionet – Central Data Repository: http://cdr.eionet.europa.eu/
[4] EMEP MSC-W modelled air concentrations and depositions: https://emep.int/mscw/mscw_moddata.html
[5] AFBI website's Search page: https://www.afbini.gov.uk/search?query=Pesticide+usage+report

Time information required special care, due to the various time scales being used. Dates having lost their proper format due to an unchecked copy-paste or format change of dates in an Excel environment were detected and corrected, for instance. Considering the time scales varied from year to month, day and evenly instant (date & hour of the day) timeframe, we chose to report the time information using two fields: the first giving the format (field “Date_Format”) and the second giving the format-dependent time value (field “Date”). These fields are detailed in Chapter 5.

An effort was made to harmonize the parameter names as much as possible within and between the case studies. Moreover, parameters were organized into groups and divided in sub-parameters when necessary, in order to facilitate exploration and exploitation of the database. Identification of a parameter is thus given in three levels: i) parameter group, ii) parameter name and iii) sub-parameter name (see Chapter 5 for more detail). There were even a few cases where the parameter names were changed (corrected) after we better understood the actual nature of the data provided. For instance, a parameter originally named “crop yield” (with data in kg of N per ha) in a file we received was providing information on the “nitrogen consumption by crops”, which is quite different.

Another step of data processing has been to aggregate the data in time and/or in space, when relevant. This was done most largely with the data of Case Study #1 and #2 of Denmark. They had many data at the farm level (main crop type, nitrogen consumption by crops, mineral fertilisation) and detailed groundwater-quality monitoring data with several observed values per year. Aggregation of those data was done using Python and R scripts, which calculated yearly total or average values at the catchment scale. The aggregated values were then added to the database (in addition to the raw data).

Table 1: Total amount of data rows per case study in the Excel database



CS n°	Country	CS name (site)	DB sheet name	Nb values	GIS-data folder name
1	Denmark	Island Tunoe	CS_1 DK_it	12988	CS01_DK_TunoeIsland
2	Denmark	Aalborg	CS_2 DK_a	14901	CS02_DK_Aalborg
3	England, UK	Anglian Region	CS_3 UK_ar	3455	CS03_UK_Anglian
4	France	La Voulzie	CS_4 FR	15337	CS04_FR_LaVoulzie
5	Germany	Lower Saxony	CS_5 DE	78	CS05_DE_LowerSaxony
6	Greece	Axios River…	CS_6 GR	597	CS06_GR_AxiosRiver_AgiosPavlos
7	N. Ireland, UK	Derg Catchment	CS_7 UK_dc	22295	CS07_UK_DergCatchment
8	Netherlands	Overijssel	CS_8 NL_o	88156	CS08_NL_Overijssel
9	Netherlands	Noord-Brabant	CS_9 NL_nb	734	CS09_NL_NoordBrabant
10	Norway	Vansjoe	CS_10 NO	119396	CS10_NO_Vansjoe
11	Portugal	Baixo Mondego	CS_11 PT	24048	CS11_PT_BaixoMondego
12	Romania	Arges-Vedea	CS_12 RO	24048	CS12_RO_ArgesVedea
13	Slovenia	Dravsko Polje	CS_13 SI	63232	CS13_SI_DravskoPolje
			TOTAL (ALL CS)	389265	

Once parameters corresponding to the Agri-Drinking Water quality Indicators (Hansen et al., 2021) required to calculate a compound indicator such as “Total mineralisation” or “N surplus” became available at a common time and spatial scale in the database for a given case study, the compound was calculated and the resulting time series was added to the database.

Moreover, the area of some geographical objects of interest such as the catchments or field plots was calculated and added to the database as well when possible and relevant.

In the end, the Excel database contains 389,265 “Results” values (data rows) for >65 parameters and >500 sub-parameters. It includes some long time series of pressure and state indicator data that are especially interesting to statistically analyse and compare (see examples in Chapter 6).

5. DETAILED STRUCTURE OF THE DATABASE

This chapter defines all fields that are present in the Excel database.

5.1 DATASHEETS: DETAILED DESCRIPTION OF THE FIELDS (DATA COLUMNS)

The data sheets of the Excel database use a uniform scheme in terms of field (column) names and order. This data scheme is documented in the following two tables, which describe the mandatory fields (Table 2), and the optional fields (Table 3) added to the right end[6] of a few CS datasheets. Note that mandatory fields marked by an asterisk must always be filled with a value.

Table 2 : Description of the mandatory fields in all data sheets of the Excel database. Fields are presented in the order they must follow. The asterisk put on a field name indicates that this field must always have a value (cannot be blank). The leftmost column of the table gives the Excel column letter (index) for that field in an Excel worksheet environment.

	FIELD NAME	DESCRIPTION
A	Case_Study*	Number-based identifier of the Case Study *Example: “CS_1” for case study n°1* Note: This field would be very useful to one who wants to prepare a large table with all data from the Excel database by merging all CS datasheets.
B	Date_Format*	Symbolic format of the “Date” values: several formats are allowed including: “yyyy”: year only (yearly data), “yyyy-mm”: year and month (monthly data), “yyyy-mm-dd”: full date (daily data), “yyyy-mm-dd HH:MM”: full date & time (instant data), “mm”: (average monthly data: e.g., 30-year climatic average monthly precipitation data) Note: The “Date_Format” must be uniform for a given parameter / sub-parameter of a CS.
C	Data_Type*	Tells if the reported “Results” value is of a “numerical” or “categorical” (any text) type This greatly facilitates the reading and processing of the “Results” field values. Note: The “Data_Type” must be uniform for a given parameter / sub-parameter of a CS.

[6] The optional fields (columns) are always placed after the mandatory ones, so that it remains easy to merge several data sheets when necessary.

D	Date*	Time (instant time or period of time) at/during which the reported “Results” value is deemed representative The “Date” value must strictly respect the associated “Date_Format”.
E	Time_Scale*	The representative time scale for the reported value. Can be: “independent” (~assumed constant), “year” (or “multiyear” if the value applies during several consecutive years), “month”, “week”, “day”, or “instant”. To better understand: an “instant” time scale is for a measurement representing an instantaneous state (typically observed at a precise date & time), whereas a “day” time scale is for a result value that is related to a full day (e.g., daily precipitation in mm). Non-instant time scales are most often used for parameter values that are totals, averages, or temporary constants (e.g., the main crop of a field plot for a given year).
F	Sampling_Point_ID*	Identifier of the sampling site location, or more precisely: unique identifier to the geographical object that was sampled to produce the value reported in the “Results” field. The same ID can be found in the corresponding GIS file (in general). Note: The ID for catchment-scale results can be either “Catchment”, the full name of the catchment or study area, or some other variant.
G	Spatial_Scale*	Representative spatial scale for the reported value. Can be: “catchment”, “sub-catchment”, “local…” (plot, piezometer, groundwater well, spring, along with/near stream…), “national”, “international”, etc. International or national scales are for some reference values (e.g., of Crop yields).
H	Parameter_Group*	Group (family) of parameters, which is mostly there to facilitate exploration of a CS’ data or summary sheets *Possible values: “Site data”, “Quality data from farmers”, “Simple quantit farm data”, “Compound or calculated quantit farm data”, “State indicator”, “Link indicator”. (Note: “quantit” means “quantitative”.)*
I	Parameter_Name*	Parameter, which sometimes also acts as a group subdivided into several sub-parameters (see below) Note: Some parameter names are implicitly defining subgroups by using a colon (“:”) to separate the main name (left part) and the subgroup names (right part), in special cases where the “Subparameter_Name” field is already used.

J	Subparameter_Name (*can be blank*)	Subparameter: Most often “” (blank) whenever a parameter name suffices. A Subparameter can either be a subdivision of a Parameter, or provide additional information on the Parameter. Note: Some sub-parameter names implicitly define subgroups by using a colon (“:”) or a comma (“,”) to separate a group name (left part) and its subgroup names (right part), or to provide additional information.
K	CAS_Number (*can be blank*)	CAS Registry Number ideally provided for all non-element chemical substances. The CAS number may prevent some confusion when reading substance names. Search for chemicals using EACH s. engine: https://echa.europa.eu/search-for-chemicals
L	Results*	Measured, reported, or calculated value for the Parameter (and Subparameter) at the given location (Sampling_Point_ID) and time (Date) Value: number or text (consistently with the specified “Data_Type”)
M	Unit*	Measurement unit for the reported “Results” value. Mandatory. Specified as simple text avoiding special characters (e.g., “µ” → ”u”). *Examples: “%”, “ug/L”, “boolean”, “kg N/ha”, “mg/L”, “m3/s”, “uS/cm”, “text”, “ha” …*
N	Below_LQ (*can be blank*)	Boolean (yes/no) values telling whether the reported “Results” value is below the Limit of Quantification. This column contains only the reported information because it is difficult and risky to assume something missing. Therefore, this field is blank most of the time. *Possible values: “Y”, “N” or “” (unknown)*
O	LQ_Value (*can be blank*)	Limit of Quantification value: rarely reported in most of the CS. No attempt was made to fill the blanks (for the same reasons as for the “Below_LQ” field). When it is reported, it is expressed in the measurement unit (“Unit”).

P	Origin (*should not be blank*)	Information about where the data comes from: "as reported" (in most cases) means that the data in the database comes from the datasets initially provided by the CS leaders during the surveys or first effective data exchanges; "added in 20XX…" indicates additional data that was added to the database later on; "reformatted from reported" indicates data that needed major changes in terms of formatting; "translated from reported" indicates data that is not expressed in the language originally used. Note: The "Origin" field was used quite freely to keep a trace of the changes made in the database. It is for information only.
Q	Analytical_Method (*can be blank*)	Analytical method or instrument used for the measurement. This information, primarily used for chemical measurements, is rarely reported. Accordingly, the field was also used to store various other information on the procedures used to prepare/obtain some of the calculated values (*e.g., "Calculated with ArcGIS based on…"*) or on the detailed source of the data.
R S T	Top_of_screen_Depth Bottom_of_screen_Depth Approx_WaterTable_Depth (*can be blank*)	These 3 fields contain additional information about the precise Z location of the sampling: - The top-of-screen depth usually gives this information. - The bottom-of-screen depth is used only to give this specific information. - The approximate water-table depth, used only in CS 6 of Greece, was otherwise not used[7]. The "…screen…" fields were used to store screen-related information in CS 1 and 2 of Denmark only. In CS 10 of Norway, the "Top_of_screen_Depth" field was rather used to store information about the sampling depth in a lake or river (m below the water surface). The 3 fields are completely blank in 9 of the 13 case studies. Unit: meters below ground surface (general), or meters below the water surface (in CS 10)

[7] The fact that "approximate water-table depth" information was almost never provided by the CS leaders most likely means that such information was not readily available to them in any of the datasets they had access to. We think that in many cases this additional information did not already exist in the monitoring data sets, so that obtaining this information would have required a lot of extra work. In other words, this blank field tells us that this information is virtually never stored alongside the water-quality monitoring data, in practice.

U	Place (*can be blank*)	Indicates where the analysis was carried out (mainly for chemical substances), or gives another name for the sampling site, as the place of analysis was very rarely reported. *Typical values: "Field", "Lab", or "Unknown".*
V	Confidentiality	Level of confidentiality of the reported "Results" data. *Possible values: "public" or "restricted".* Only the "public" data can be made publicly available online. The "restricted" data should not be shared.

Table 3: Description of the optional fields that may be present to the right of the mandatory fields (columns) in some of the datasheets of the Excel database. These fields provide extra information that may be useful in specific situations.

OPTIONAL FIELD NAME	**DESCRIPTION**
(…) Sampling_Point_NAME (*optional field*)	This optional field is sometimes used in some CS data sheets to store the long name of the sampling site when the "Sampling_Point_ID" value is a coded identifier. It plays a role similar to the "Place" field described above. *For instance: In CS 10 of Norway, "003-38229" is the coded identifier (in "Sampling_Point_ID") corresponding to the long name "Saebyvannet" (stored in the "Sampling_Point_NAME" field)*
(…) EIONET_Source_URL	This optional field is used in the CS data sheets where public data from the European database EIONET has been imported. It gives a link (URL) to the web page that was accessed in order to download the data.
(…) Orig_Parameter_Name (…) Orig_Sampling_Point_ID (…) Orig_Results (…) Orig_Analytical_Method (*optional fields*)	Extra fields added to a CS datasheet to facilitate backward linking with the previous (more raw) versions of the dataset. They are used notably when numerous texts had to be translated to English, when parameter names changed profoundly in terms of format or language, or when some Sampling_Point_ID had to be corrected for consistency reasons.

5.2 SUMMARY SHEETS

The summary sheets in the Excel database provide an overall picture of the content of each CS datasheet. The different parameters present in the datasheet are listed, along with the measurement units, scales of observation, and data sources. Some basic statistics including the number of reported results and different sampling sites, and the minimum & maximum dates of observation, are given for each parameter. The GIS-related information required to link the tabular data to the GIS data is also provided in two complementary columns (Table 4). An illustration of the linking process between a parameter of a CS in the Excel database and the GIS data is presented earlier in the document (Figure 3).

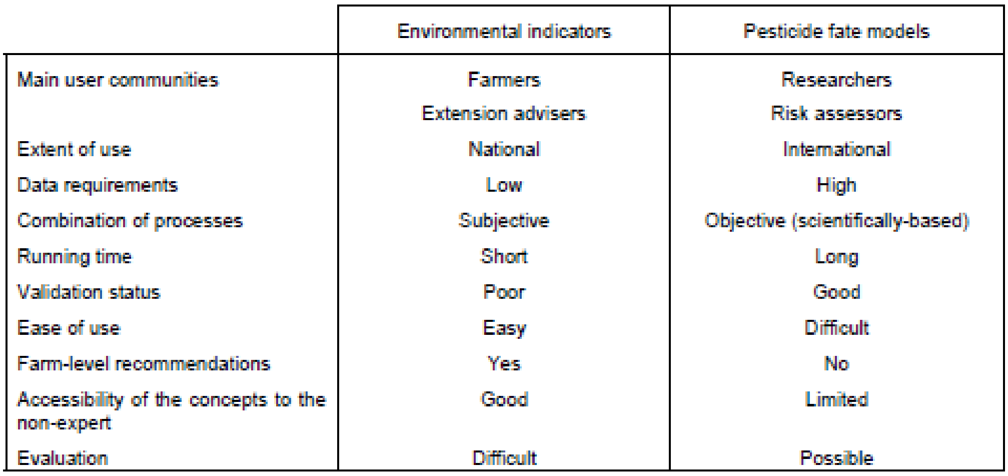
Table 4: Description of the columns providing GIS related information in a summary sheet of the Excel database

OPTIONAL FIELD NAME	DESCRIPTION
GIS_Related_Files	This column gives the filename (and relative path when relevant) of the GIS file containing the spatial features related to the parameter. If no GIS data is linked to the parameter, GIS_Related_Files = “NONE”. Note: If there is more than one GIS file linked to the parameter, the multiple filenames listed in the column are separated by a vertical bar symbol (“ | ”).
GIS_Key_Field	This column specifies the name of the field of the GIS file’s attribute table that should contain the IDs that are present in the datasheet for that parameter.
GIS_Key_Value	This column specifies the ID of a specific spatial feature of the GIS file. Not needed in most cases.

6. USING THE DATABASE

This database can be used in several ways. It may be used to explore data (as presented in §6.2) or to calculate additional indicators. Depending of the case studies interests, the most commonly available State indicators are about nitrate and pesticides concentrations in water. From a practical point of view based on its actual content, the database may notably be used to explore statistical relations between related Pressure and State indicators.

To illustrate our reasoning, Figure 4 compares between selected features of pesticide fate models and environmental risk indicators. Environmental indicators for pesticides are mainly devised to be used by farmers and extension advisers in each set of agro-environmental conditions that are usually specific to one country. Indicators generally have low data requirements, are easy to use, can be calculated quickly, are amenable to the non-expert, but they often suffer from a lack of scientific validation and the fact that the combination of the various processes is done on a subjective basis. Given the very significant efforts put into the evaluation and validation of pesticide fate models over the last 20 years and, at the same time, the opposite and complementary profiles of the two types of tools, a will of the Fairway project has thus been to improve the evaluation of some environmental indicators. That is why WP3 focused on investigating statistical relationships (correlations) between pressure and state indicators.



	Environmental indicators	Pesticide fate models
Main user communities	Farmers Extension advisers	Researchers Risk assessors
Extent of use	National	International
Data requirements	Low	High
Combination of processes	Subjective	Objective (scientifically-based)
Running time	Short	Long
Validation status	Poor	Good
Ease of use	Easy	Difficult
Farm-level recommendations	Yes	No
Accessibility of the concepts to the non-expert	Good	Limited
Evaluation	Difficult	Possible

Figure 4: Comparison between environmental indicators for pesticides and pesticide fate models (from Dubus and Surdyk, 2006)

In WP3, two main families of relations between Pressure and State indicators were thus explored using the database, and also reported in Hansen et al. (2021):

- The relation between gross nitrogen budget (Pressure) indicators (e.g., Nitrogen surplus[8]) and water quality (State) indicators (e.g., Nitrate concentration in groundwater), and
- The relation between pesticide application rate (Pressure) indicators (e.g., the estimated total volume of atrazine applied each year) and concentration of pesticide (State) indicators (e.g., the concentration of atrazine and DEA in spring water).

Note that simpler Pressure indicators may be considered when it is impossible to compute the Nitrogen surplus indicator, for instance: Mineral fertilisation or Main crop type ideally combined with an N surplus (or/and crop fertilisation) indicator. However, these simple indicators remain overall less effective than the compound N surplus indicator (see §6.1.4). Similarly, a pressure indicator not directly quantifying application rate (e.g., N surplus or Mineral fertilisation) may be tried as a proxy for pesticide application rate when no such indicator is directly available (assuming a positive relation between pesticide application rate and N surplus). It may not produce significant correlations in several cases, but it is worth a try still.

6.1 CROSS-CORRELATION ANALYSIS

The exploration of statistical relationships between pairs of Pressure and State time series aims to provide answers to the applied question "How long do we have to wait to see the impacts of changes in agricultural practices?" In WP3 of Fairway, this delay was investigated through cross-correlation analyses of the aggregated data series (by year, for the whole catchment) in the case studies where data were available. These analyses produced "lag time" quantitative information.

6.1.1 Lag time as a Link indicator

- The "lag time" is an estimation of the average delay in the response of an "output" time series to the pulses (fluctuations) of an "input" time series. Technically, it corresponds to

[8] A scientific paper was published specially about the nitrogen surplus indicator, as part of the Fairway project (Klages et al., 2020).

> the delay that (when cancelled in the output time series by subtracting it to its time values) maximises the correlation between the two thus “aligned” time series. From an applied perspective, the lag time provides an efficient means to estimate the travel time of contaminants in the groundwater system up to the sampling points. For instance, a pair of related indicators that were studied using cross-correlation analysis consisted of Nitrogen surplus as the input signal and nitrate concentration in groundwater as the output signal.

That work on lag time estimation and interpretation for three case studies of Fairway (Case Study #1 & #2 of Denmark and Case Study #4 of France) was published in a scientific paper (Kim et al., 2020) and more recently was summarized in another report of Fairway (Hansen et al., 2021). In that work, the computed lag times were also compared to groundwater ages estimated using environmental tracers, which revealed that calculated lag times are generally shorter than tracer-based groundwater age estimates (Kim et al., 2020).

Those lag-time and groundwater-age data sets were added to the database as “Link” indicators when available.

6.1.2 Lag time related to nitrate contamination of groundwater

Figure 5 shows an example of graphical results from a cross-correlation analysis of pressure and state indicators related to nitrate contamination of groundwater. The top graph in the figure shows how the correlation coefficient r varies as a function of the considered lag (delay removed from the “NO_3 conc.” time series). The dotted blue line indicates the r threshold below which the r values are not statistically significant. This graph thus informs that the (peak) correlation (r = 0.79) found at a lag = +19 years is significant. However, the found lag cannot be deemed precise since several r values are very close to the peak r value. Nonetheless, the bottom graph confirms that the two-time series show a similar variation over time once the delay of the response of the “NO_3 conc.” series is removed. In this particular case, it will be interesting to repeat this cross correlation analysis in 5–15 years to see if the NO_3 concentrations continue to decline, as to suggest in this present-day finding.

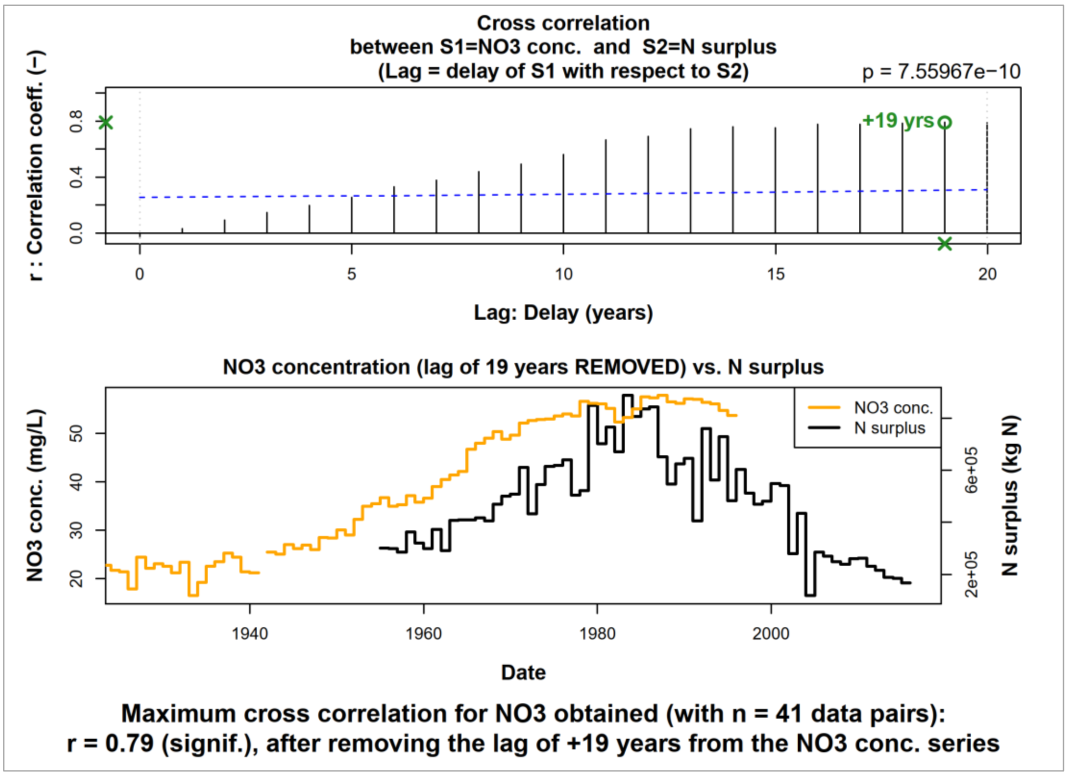


Figure 5: Cross-correlation analysis results for nitrate and superimposed pressure and state curves after removing the lag time of +19 years from the 'NO$_3$ conc.' (state) time-series, for bottom spring of La Voulzie Case Study #4 of France

Figure 6 shows another example of cross correlation analysis of NO$_3$ concentration and N surplus. In this case, the r threshold (below which r values are not statistically significant) is very high, there is only one r-value > 0.5, and it is not significant. This is because the two times series are very short in duration and do not share enough common features. This example highlights that it is often impossible to assess the lag time when the pressure and/or state time series are too short compared to the time interval it would take to fully capture the transient feature we wish to study in the two signals.

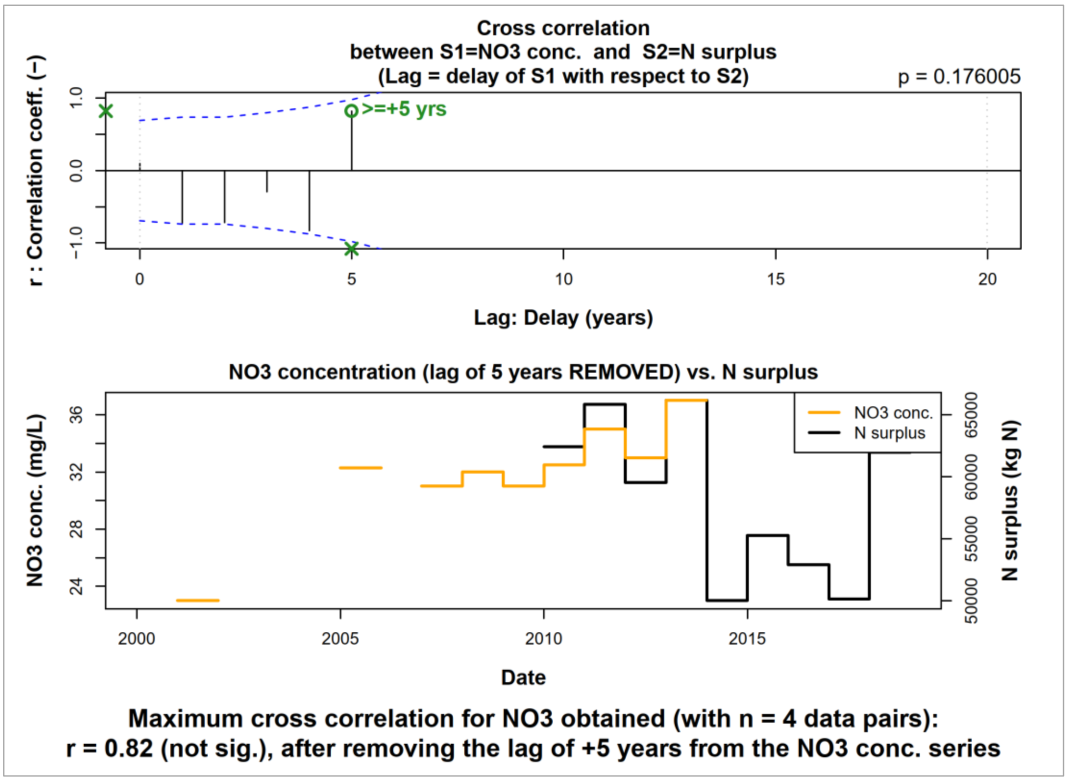


Figure 6: Cross-correlation analysis results for nitrate and superimposed pressure and state curves after removing the lag time of +8 years from the 'NO$_3$ conc.' (state) time-series, for sampling point "34_2582_1" of Aalborg Case Study #2 of Denmark

6.1.3 Lag time related to pesticide contamination of groundwater

Figure 7 shows an example of results from a cross correlation analysis of pressure and state indicators related to atrazine (pesticide) contamination of groundwater. The top graph shows that a (peak) correlation (r = 0.94) found at a lag = +22 years is significant. However, as for the nitrate related analysis, the lag identified cannot be deemed precise since the r values around it are very close to the peak value. Nonetheless, the bottom graph shows that the two series fit quite well visually once the delay of the response of the state series is removed. Furthermore, this particular analysis reveals the importance of having longer time series of both pressure and state indicators. Indeed, the seemingly satisfying fitting of the two series after the lag of 22 years was removed may not be the best answer in this case. Maybe a lag time of 15 years, 9 years, or even less would be a closer estimate of the actual "age" of the atrazine contaminant since it was introduced in the groundwater system. Having longer and more continuous and detailed time series would likely help narrow down the uncertainty in the lag time estimate.

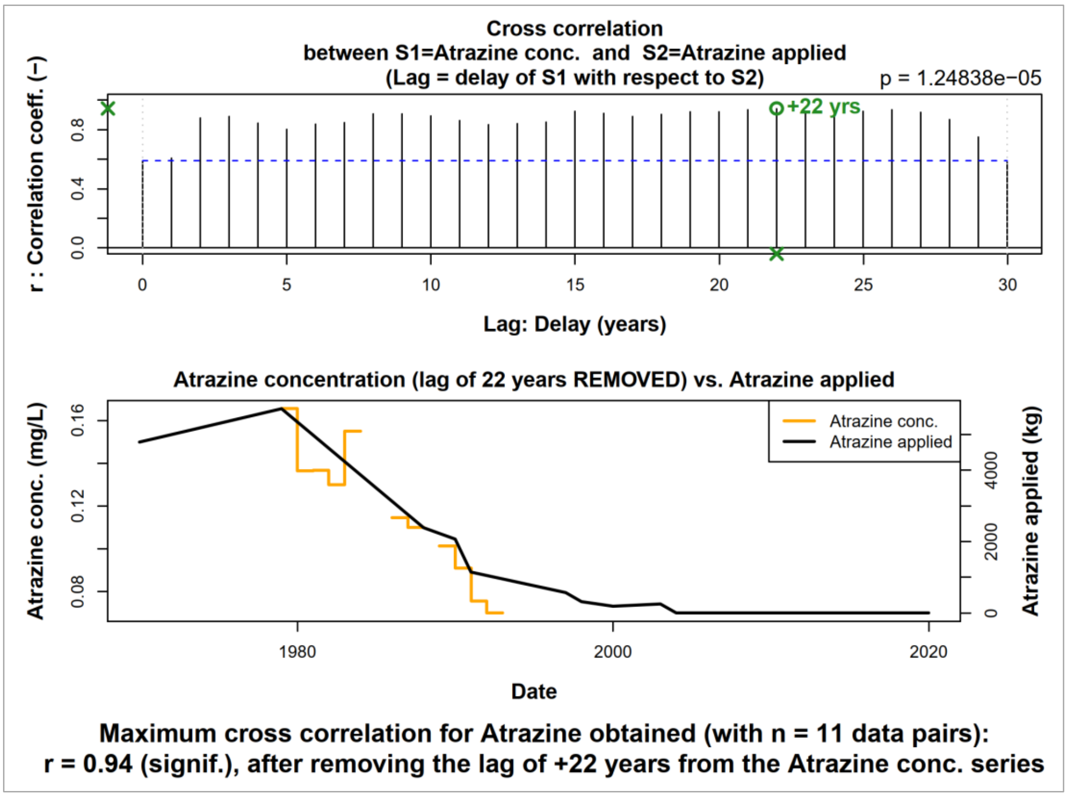


Figure 7: Cross-correlation analysis results for atrazine and superimposed pressure and state curves after removing the lag time of +22 years from the 'Atrazine conc.' (state) time series, for top spring of La Voulzie Case Study #4 of France

Overall, the above interpretations of three examples of cross-correlation analyses underscore the importance of carefully examining the detailed results of such analysis, especially when the length or quality of the time series is limited.

6.1.4 Lag time estimation using other indicators

In addition, to the indicators presented above, some other indicators could be tested (as previously presented in Hansen et al, 2021). In the "La Voulzie" case study #4 of France, lag times were also estimated using the main crop area (i.e. wheat in that case) as the pressure indicator. Thus, the correlation coefficients were smaller than those obtained with other N pressure indicators (N budget and Mineral fertilisation). For example, the estimated lag time was 0 year at the main spring with a weak (r = 0.54) yet statistically significant correlation (Table 5). Using soil occupation directly as an indicator was shown insufficient but the connection between crop area and the presence of specific herbicides in surface waters is documented (e.g Brown et al, 2007).

Maybe that no correlation was found in "La Voulzie" case study because the wheat-area time series was available from 1994-2014, whereas the mineral fertilizer input and potential N leaching were available from the 1950s to 2018.

In the "La Voulzie" statistically significant correlations were also found between mineral fertilisation and yearly average concentrations of nitrate in the springs. The estimated lag times at the top, main, and bottom springs were 7, 14 and 14 years, respectively. The lag times estimated using the potential N leaching were similar to those obtained with mineral fertilisation (Table 5). The mineral fertilisation seems to be an appropriate pressure indicator at the catchment scale if a lag time is to be calculated. Nevertheless, the mineral fertilisation indicator can only be applied in areas where no organic fertilizers are applied, that is to say, regions with no or negligible animal breeding.

Table 5: Estimated lag times (in years) using main crop surface or mineral fertilisation as pressure indicator in La Voulzie (Case Study #4 of France) (From Hansen et al., 2021)

Monitoring points	Lag time with main crop surface as pressure indicator (correlation coefficient)	Lag time with mineral fertilisation as pressure indicator (correlation coefficient)
Top spring	4(0.65)**	7(0.95)**
Main spring	0(0.54)*	14(0.86)**
Bottom spring	8(0.70)*	14(0.93)**

* Statistically significant at $p<0.05$; **$p<0.005$; † Outer protection zone

Statistically significant correlations were also found between fertilisation and the yearly average concentrations of nitrate in Denmark and Slovenia case studies. In Tunø Island (Case Study #1 of Denmark), the correlation analyses yielded significant results at 15 points out of 24 monitoring points. At the 15 points with significant results, the tested pressure indicator (mineral fertilisation) showed strong correlations with the state indicator (yearly average nitrate concentrations). At the other 9 points, either the time series were too short and/or sparse, or the concentrations were relatively invariant at low concentrations ranges (10 mg/L) near the limit of quantification (LQ); why those other results were not significant nor reliable.

In Dravsko Polje (Case Study #13 of Slovenia), the correlation analyses yielded significant results between the amount of fertilizers applied on the catchment (in tons of nitrogen) and the nitrate concentrations[9]. The best correlation was found for a lag time of 4 years. This lag time could seem short considering that the depth to the water table could reach 15 m in the northern part of the study area. But the Slovenian Case study leader (University of Ljubljana) estimated that this lag time could be correctly calculated. In the upper terrace, the proportion of rock/gravel/pebbles is 40%. The proportion go higher with depth. On lower terrace, soils contain a high share of sand. Lag times have also connection to drought and precipitation periods (dry and wet years) that impacts nitrate leaching trough soil profile.

In the England Case study (Case study #3), no lag times were calculated. Since the catchment is surface, the transfer times seem to be annual, and a relationship was found between the month of year and concentration. More closely, the relationship appears between monthly rainfall and nitrate concentration.

Cooper et al, 2020 proposed an explanation to this phenomenon. During winter, the monitoring point is characterised by high nitrate concentrations resulting from a precipitation-induced leaching through agricultural field (especially in exposed arable land). Conversely, during the summer, the monitoring point is characterised by low nitrate concentrations due to low precipitation rates and also because arable crop growth absorbs excess nitrogen in the soil.

6.2 EXPLORATORY DATA ANALYSIS

More generally, the numerous time series and time-independent data of the database can be explored and analysed in various ways.

At the simplest level, the data can be explored individually, visually, by opening the Excel database and scrolling in one of its datasheets. However, since there are thousands of data rows in the CS datasheets, it is much more efficient and appropriate to explore the data sets by the parameter(s). One easy way to carry this out within Excel is to use the AutoFilter on the "…parameter name" fields to select the indicators you wish to extract. Once the filters are properly set, you can copy the visible dataset to the clipboard, and then paste it to another Excel workbook. You may then create

[9] This is the average concentration of nitrate in groundwater wells in the catchment.

graphs from this specific dataset. Suppose you want to prepare a graph displaying several time series. In that case, you may first create a graph with a first time series, and then use the Excel Paste Special options to add more XY series iteratively to the graph as you select each of the other indicators (X = "Date" and Y = "Results") you want to display in the graph. You can also create histograms, box plots, and many other types of graphs, within an Excel environment. In addition, you may use selections and formulas to calculate various statistics (mean, median, min, max, counts, etc.) for the indicators of interest. Alternatively, all sorts of graphs can be generated using scripting languages (see §6.3).

Figure 8 provides an example of a stacked area chart that was prepared from selecting of four indicators related to Nitrogen inputs in the La Voulzie catchment of France (Case Study #4). This specific graph provides an efficient means to assess the relative contributions of the "input" components of the N-surplus compound indicator.

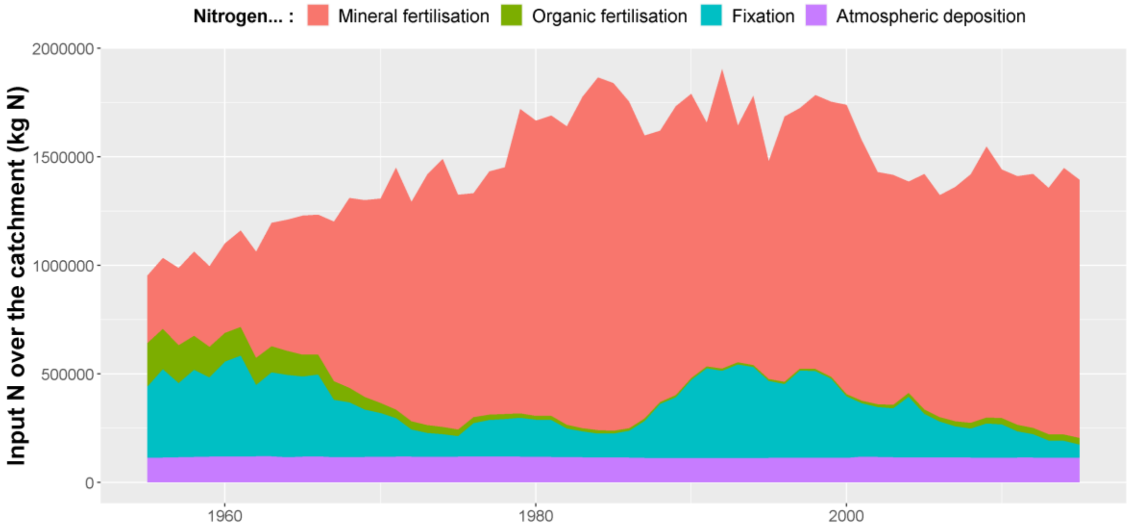


Figure 8: Stacked area chart prepared from catchment-scale yearly N-input indicators data of La Voulzie (Case Study #4)

Figure 9 provides an example graph where six indicators (time series) from the France CS 4 are displayed side by side. Note that the "Recharge" indicator was calculated from the rainfall and potential evapotranspiration ("PET") indicators. This type of graph is very useful when exploring the datasets. It helps to detect similarities between the different signals. For instance, in this graph we see notably that the fluctuations in spring discharge are very similar to that of the hydraulic head and that peaks of recharge propagate in the spring discharge and hydraulic head signals after a delay of a few months. Complementary cross-correlation analyses (not shown here) allowed us to assess this hydraulic lag time more precisely: the spring appears to have an average delay of response of about 5 months.

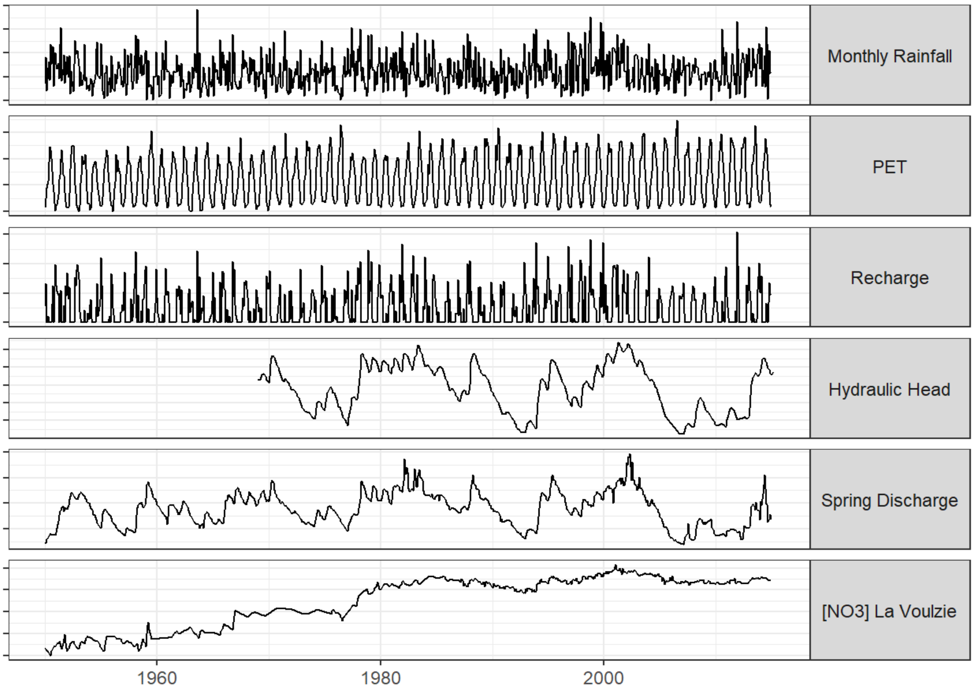


Figure 9: Example graph of several time series put side by side, from indicators data of La Voulzie (Case Study #4)

6.3 Using the database outside Excel

Although it is possible to use this database within the Excel software environment to calculate statistics and produce various graphs, it may be more efficient to use other data analysis tools to exploit this database. In particular, scripting languages such as R or Python may be used[10]. These programming languages and environments include hundreds of packages and functionalities that offer more flexibility in all steps of database exploitation, from data import and processing to data transformation, calculation of statistics and graphical representation. Here are two operations that may be applied to facilitate external use of the database:

- Creation of a single large table containing the data from all case studies. As mentioned earlier, it remains possible to merge all data sheets into one large table, thanks to the uniform scheme used for the columns in all CS datasheets. An easy way to carry this out is to append (copy & paste) all rows of data from every Excel database sheet into a new Excel sheet. The resulting large table, saved as an Excel file, can then be read directly (in Excel format) by software or scripts, while other software and scripts will prefer to read a delimited text file.
- Creation of a text file version of the Excel datasheet of interest. Indeed, it is often simpler and safer (more comprehensive) to read text files rather than Excel files within scripting environments[11]. Any Excel sheet can be exported to a text file. The recommendation is to export as a tab-delimited text file, as this special character (tabulation) is not used in any field of the database.

[10] R: https://cran.r-project.org/ ; Python: https://www.python.org/

[11] In particular, there are often issues when importing (reading) the “Date” field right from the Excel file. That is because it contains a mixture of numbers (e.g., “2005”), actual dates with or without the time of day (e.g., “2010-03-04 15:18” or “2005-01-02”), year-and-months stored as text (e.g., “2015-07”), etc. In the Excel file, dates (with or without time of day) are stored as decimal numbers = number of days since a reference date, years are stored directly as numbers = the year, while other types of time information is stored as text.

Reading and exploiting the database using scripts is not straightforward, however. It requires the development of specific routines to read the data file(s) like (i) process the flexible “Date” information; (ii) extract data for a given parameter, sub-parameter, sampling point (site) and spatial / time scale; (iii) aggregate an indicator at coarser time scales; (iv) perform various other calculations; (v) generate graphs; etc. Although the deliverable does not include such tools, this document will help the reader interested in the programming of such tools. Here is a recommendation on how to read a Case Study datasheet, exported as a tab-delimited text file, efficiently and safely:

1. Read the file using a function that creates a “data frame” from it with all columns treated as text (no automatic conversion of numeric values or dates).
    - Alternatively, if the scripting language allows reading an Excel file with automatic conversion disabled, you may try to read a data sheet directly from the Excel file.
2. Convert fields containing only numeric values to actual numeric values (e.g., “LQ_Value”, “Top_of_screen_Depth”, etc.)
3. Determine the list of all different “indicators” (parameters / sub-parameters) in the imported dataset. Keep a copy of this list in memory.
4. Process the dataset one indicator at a time:
    - Select the data rows that correspond to that indicator.
    - Process the “Date” information according to the “Date_Format”.
    - Convert the “Results” text values to numeric values if “Data_Type” = ‘numerical’.
    - Split the selected data rows by site and by time scale (if there are many) so that each group of rows describes only one parameter or sub-parameter, for only one site (and spatial scale) and one-time scale (in case of time-dependent data).
5. Gather all of these groups of data rows in a named hierarchical list.
6. Convert time-dependent data to time series objects if the scripting language allows it.
7. Keep summary information on every indicator along with the data series, for instance:

```
..$ State indicator.Concentration of nitrate                :List of 9
.. ..$ Parameter_Group  : chr "State indicator"
.. ..$ Parameter_Name   : chr "Concentration of nitrate"
.. ..$ Subparameter_Name: chr ""
.. ..$ Spatial_Scale    : chr "local, groundwater well"
.. ..$ Time_Scale       : chr "instant"
.. ..$ Data_Type        : chr "numerical"
.. ..$ Unit             : chr "mg/L"
.. ..$ MANY_full_params : logi FALSE
.. ..$ Nb_Results       : int 3700
```

8. Program an easy-to-use function allowing its user to specify the indicator he wishes to get (and for which site and time scale if applicable) right before performing the data extraction.
9. Use this function in a script to extract indicators for a specific purpose.
10. Carry out further data processing on the extracted indicators (if needed).
11. Carry out analyses on the extracted indicators.

7. CONCLUSION

As presented earlier, the database was not populated in the same way by all the partners. Several limiting factors and issues were identified throughout this task of WP3.

7.1 DEFINITIONS OF BOUNDARIES

When it is needed to compare a state indicator against a pressure indicator, the first question to ask is generally the scale of work. According to a hydrogeologist, the best scale is the catchment area because it represents a homogeneous zone that collects water flows. A catchment area is a physical extent based on hydrogeological/hydrologic data that cannot easily be linked to administrative or agricultural pre-existing territorial divisions. It is indeed often complicated to have access to farming data matching the catchment area. And thus, data from larger territorial divisions or smaller divisions frequently have to be used.

One possibility is to use data from scales smaller than the catchment area: generally data at the plot scale. However, listing plots is not the most time-consuming task. In a first approach, we asked for data at the plot scale in the Fairway project but collecting data at this scale was not feasible in most case studies. Even if the case study leaders were able to collect the location of plots, collecting the rest of the data we asked at the plot scale was too time consuming. Notably, collecting soil occupation or fertilisation data at plot scale was too time consuming in many case studies, especially in the largest ones such as Arges-Vedea (Case study #12). Some partners had to find alternative ways to define a much smaller sub-catchment to fulfil the task at this scale (Case study #3).

The problem of the plot scale, even if it gives an exact detailed representation of the catchment, is thus that it requires too much effort to collect data at this level of precision for indicators such as crop type, fertilisation and crop yields. The efforts are even more important in a deeper, more inertial aquifer since the data will need to be collected for several years or even decades[12]. The partners that fulfilled this task with the more ease in Fairway started working in the study site before the project and had already started collecting some data.

One might ask why so precise and hardly reachable data was asked. The fundamental reason is that WP3 aimed to select the best (scientifically-sound, fast-responding, and intuitive) indicators to be used by “local” stakeholders in the decision making processes. Therefore, datasets that can precisely illustrate the cause-effect relations at the local level where needed[13]. A second reason why precise data was asked is that most indicators could be reliably tested only if the input data was precisely known.

Plot scale data are, since 2016, even more, problematic to handle. Many member states protect the personal data. This includes the farming practices data since, in many cases, they are linked to a plot and thus to a farmer. Until 2016, this general statement could have been circumvented under certain conditions for a research project. However, in 2016, a new regulation on data protection (EU 2016/679) was enforced. Questions on confidentiality of farm data, therefore, arose in conjunction with the beginning of Fairway. For instance, the German partner (Case study #5) found no possibilities to send data at the plot scale. Note, this is not issue due to data collected at

[12] For the partner in England (Case Study #3), it has been a challenge to collect the data, not helped by all of the people who collected the data c.20 years ago having either retired or died.

[13] The difficulty is that the scale of WP3 is the catchment scale while other WPs of the Fairway project worked at regional or national scales

the plot scale but it could be at the farm scale also. Partner of Case study #8 was not allowed to share private data (including nitrate concentrations) of farmers.

Many case studies are not even catchment areas but are larger (Figure 10). These case studies are larger areas defined for a specific need with a certain homogeneity (farming practices, geology). These large parts of territory include several abstraction wells.

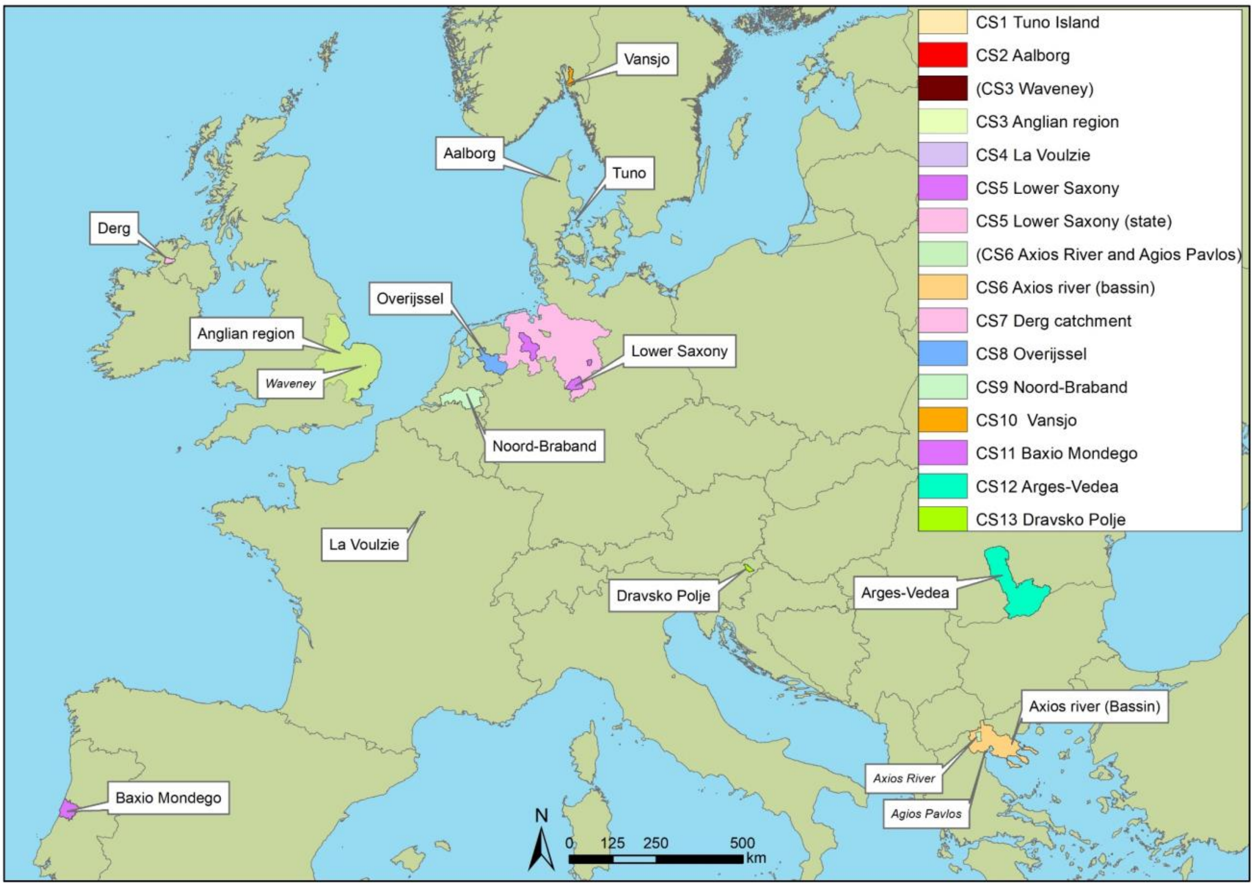


Figure 10: Case studies real extents

For instance, the German Case study (Case study #5) roots in a Lower Saxony (Federal state) project on manure treatment and export from intensive animal production (north-west) regions to regions with arable farming (east). The main partner was the Lower Saxony Chamber of Agriculture. A large territorial scale is needed to exchange manure between areas. The England Case study (Case study #3) comprises three quite small areas within the Anglian water areas within the larger Anglian Water Drinking Water company total area. All areas are independent surface water catchment within the same region geographically so similar altitude, latitude, rainfall, farming pressures. The water company area is one of the reasons why the case study has a large territorial scale.

When having a very large case study, more data (about water quality and farming practices) must be collected to properly calculate pressure indicators. Not all of the abstraction wells within the study area will necessarily have the same trends in pollutants, so different ways to apply indicators in these cases have to be used. This scale of work is designed to get large Multi-Actor-Platforms and the territory needs to be split to understand the impacts on each abstraction well. But in such MAPs, there is no desire to share/divide the territory into smaller parts: the MAPs are built to be large enough to create a *momentum* that can gather local stakeholders' goodwill and significantly impact the environment.

The England Case study (Case study #3) added data for the Waveney micro-catchment area in the database. The Waveney micro-catchment area is not a formal part of the Fairway Project; it just happens to be in the same Drinking Water company area and is a micro-catchment where Adas was able to find the detailed data that was required (Figure 10).

7.2 LOCAL DATA THROUGH EUROPEAN DATABASES

It was envisaged in the early stage of Fairway to use European Databases to fill the Fairway database when local data sources were missing. As explained above, using too-large-scale data is not feasible. The European databases providing data at the national scale were thus discarded.

However, it is possible to use default values at the regional scale. For instance, data from the EMEP MSC-W model database could be used for atmospheric deposition (see Chapter 4). Moreover, the European database also can provide data at the local scale. For instance, reports on water quality are published on the EIONET Central Data Repository website[14]. This data is sometimes public. In Fairway, data available for France (Case study #4), Greece (Case study #6) and Slovenia (Case study #13) were downloaded to test whether this data could easily be used. Those countries were tested because their data was public.

For France, the extraction wells used in the Fairway case study are not part of the EIONET database because the site was not selected to be reported to Europe as part of the Nitrate Directive. Data for a similar site (same groundwater body, same catchment) is available. Concentrations in NO_3 are available for 2015 and 2019. For Slovenia and Greece, concentrations for several drinking water extraction wells are available on the groundwater body below the case studies (Case studies are larger than the French one for these two countries). In Slovenia, concentrations in NO_3 are available from 2008 to 2019. In Greece, concentrations in NO_3 are generally available for 2018 and 2019, and an average concentration is available for 2013-2015.

Thanks to the EIONET database, ten years of NO_3 concentrations dataset with a yearly time step could be obtained for the Slovenian case study. In many other case studies, very little data is available (i.e., few years only) or data is not available. We contacted the EIONET team to have access to more data, but we would have to contact each member state services independently to have more data.

The EIONET / EEA website[15] also provides data for many surface water and groundwater points between 1960 and 2017 for almost every member state. Each point has its own pattern of data. A point could have one data only, whereas another one could have ten data gathered in the 1990s. No regular pattern could be observed. Data covers substances like nitrate or some pesticides, but the database also contains data about pH or conductivity. However, as explained above, inconsistency in data collection patterns often jeopardize the construction of time series.

[14] EIONET Central Data Repository website: http://cdr.eionet.europa.eu/ .

[15] EEA website: https://www.eea.europa.eu/data-and-maps/data/waterbase-water-quality-icm

7.3 TIME SCALE OF THE MONITORING DATA

Another major issue in the collected data is the limited time span of the indicator data. In some cases the sites are being followed for a long time (e.g., >50 years in Case study #4 of France), whereas in other cases the sites were followed for a couple of years only (e.g. <10 years in Case study #7 of UK). Moreover, for the most effective use of the database, it is further needed that time series share a long period in common. This is notably required to allow cross correlation analyses of pressure and state indicators with the consideration of response lags that can be long (>5 years and even >10 years in several cases) when the studied aquifer system is large and/or inertial (e.g., see the example results in §6.1).

7.4 INSTITUTE WITH DIFFERENT AIMS

Fairway gathers many different institutes, each of them having its own field of work. None of the project partners had in its regular missions to collect both farming practices (Pressure Indicator) and hydrogeology/hydrology data (State Indicator).

In the case study, organisations generally rely on a third party to get their missing data. For instance, BRGM (Case study #4) manages the French national database on groundwater quality, giving easy access to data on contaminant concentrations in groundwater. It still had to ask the catchment manager (*Animateur de bassin*) to access farming practice data. A project on this topic had started earlier (in 2015) in La Voulzie so data had already been exchanged, and data for the missing recent years was easy to acquire. AFBI (Case study #7) and GEUS (Case study #1 and Case study #2) were involved in projects on their case studies that started before the Fairway project. Those organisations began to collect data before the start of the Fairway. They had little difficulty providing data to the Fairway database.

Some other partners did not have a prior project on their case study, so it was a long task to get data outside their usual field of work for their case study. First, a third party had to be identified, and then appropriate requests had to be made. The larger the catchment/case study area, the less likely the answers from the third party were positive since the work of data collection involved would have been too important (see §7.1).

Access to a third party that can provide extra data is a key problem. Institutes (research institutes, universities or private companies) specialised in a given field of environmental sciences or management generally have easier access to the data or the third parties with the data on their area of expertise. Just knowing that a third party has the data is not enough, though, to convince the third party to share the data for a project in which he is not involved.

The German Case Study (Case study #5) in Lower Saxony is involved in the pre-existing project on manure treatment and export. The reason why the partners (University of Thünen and Chamber of Agriculture) could not supply information on the water quality of specific wells without asking a third party outside the project. Two attempts to collect data with the help of the water supplier "OOWV" were made: first, the record time scale of the data (about 15 years) was not sufficient, as water transit time was more than 30 years. A second attempt was thus made with another catchment. But data of this very small catchment did not show a good response to mitigation measures by farmers in the related cooperation area.

For new case studies and new MAPs, especially the largest ones (e.g., Case study #6), it was unachievable to collect farming practices data at the plot scale. Either because they could not collect data themselves or because they did not find a specialised third party that could have carried out the task for them.

As reported, case studies could have difficulties getting access to the data outside of their field of work. On the other hand, they have easy access to data in their field: for instance, they know which national databases contain relevant and readily usable data. Moreover, they have easy access to all data they have already collected. For example, the Overijssel Case study #8 (in the Netherlands) provided us with a complete database of piezometric head data also available online from a Dutch website (only a sample was included in the Fairway database).

8. REFERENCES

Brown C.D, Holmes C., Williams R., Beulke S., van Beinum W., Pemberton E., Wells C. (2007) How does crop type influence risk from pesticides to the aquatic environment? Environmental Toxicology and Chemistry, 26:1818–1826

Cooper RJ, Hiscock KM, Lovett AA, Dugdale SJ, Sünnenberg G, Vrain E. Temporal hydrochemical dynamics of the River Wensum (2020), UK: Observations from long-term high-resolution monitoring (2011-2018). Sci Total Environ. 724:138253. doi: 10.1016/j.scitotenv.2020.138253.

Dubus, I.G., Surdyk N. (2006). State-of-the-art review on pesticide fate models and environmental indicators. Report DL#4 of the FP6 EU-funded FOOTPRINT project, 39p.

Kim, H.; Surdyk, N.; Møller, I.; Graversgaard, M.; Blicher-Mathiesen, G.; Henriot, A.; Dalgaard, T.; Hansen, B (2020). Lag Time as an Indicator of the Link between Agricultural Pressure and Drinking Water Quality State. *Water* , *12*, 2385. https://doi.org/10.3390/w12092385

Hansen, B., Kim, H., Møller, I., Henriot, A., Laurencelle, M., Dalgaard, T., Graversgaard, M., Klages, S., Heidecke, C. Surdyk, N. 2021 Evaluation of ADWI's: Agri-Drinking Water quality Indicators in three case studies. Fairway Delivrable WP3 3.2

Klages, S., Surdyk N., Christophoridis C., Hansen, B., Heidecke, C., Henriot, A., Kim, H., Schimmelpfennig, S. (2018). Review report of Agri-Drinking Water quality Indicators and IT/sensor techniques, on farm level, study site and drinking water source. Fairway Delivrable WP3 3.1

Klages, S.; Heidecke, C.; Osterburg, B.; Bailey, J.; Calciu, I.; Casey, C.; Dalgaard, T.; Frick, H.; Glavan, M.; D'Haene, K.; Hofman, G.; Leitão, I.A.; Surdyk, N.; Verloop, K.; Velthof, G. (2020) Nitrogen Surplus—A Unified Indicator for Water Pollution in Europe? *Water* , *12*, 1197. https://doi.org/10.3390/w12041197

9. ACKNOWLEDGEMENTS

We would like to thank everyone who contributed to the Database or to this short note by providing data sets or help of any kind. Special thanks to:

A. de Vries, H. Kim, I. Møller Balling, J. Williams, I. Wright, J.F. Vernoux, S. Klages, C. Heidecke, B. Hansen, C. Christophoridis, L. Farrow, R. Cassidy, C. van den Brink, M. Hoogendoorn, F. Clayer, L. Jackson-Blake, I. Amorim Leitão, I. Calciu, M. Curk.

Although most are not the authors of this short note, they still are full contributors to the database itself, and thus to the Delivrable WP3 3.3 of Fairway. So, thank you all very much. And sincerely sorry for those we may have forgotten to mention here.