OSPAR
COMMISSION

Assessment on statistical techniques applicable to the OSPAR Radioactive Substances Strategy

Radioactive Substances Series

2009

OSPAR Convention

The Convention for the Protection of the Marine Environment of the North-East Atlantic (the “OSPAR Convention”) was opened for signature at the Ministerial Meeting of the former Oslo and Paris Commissions in Paris on 22 September 1992. The Convention entered into force on 25 March 1998. It has been ratified by Belgium, Denmark, Finland, France, Germany, Iceland, Ireland, Luxembourg, Netherlands, Norway, Portugal, Sweden, Switzerland and the United Kingdom and approved by the European Community and Spain.

Convention OSPAR

La Convention pour la protection du milieu marin de l'Atlantique du Nord-Est, dite Convention OSPAR, a été ouverte à la signature à la réunion ministérielle des anciennes Commissions d'Oslo et de Paris,
à Paris le 22 septembre 1992. La Convention est entrée en vigueur le 25 mars 1998.
La Convention a été ratifiée par l'Allemagne, la Belgique, le Danemark, la Finlande, la France, l’Irlande, l’Islande, le Luxembourg, la Norvège, les Pays-Bas, le Portugal, le Royaume-Uni de Grande Bretagne et d’Irlande du Nord, la Suède et la Suisse et approuvée par la Communauté européenne et l’Espagne.

Acknowledgement

This report has been prepared by Mr Bruno Fievet, France with the support of Ms Karine Beaugelin, France.

Contents

Executive summary

This report presents the conclusions of the work by the Intersessional Correspondence Group on statistical techniques applicable to the OSPAR Radioactive Substances Strategy (ICG-Stats), which was established under the OSPAR Radioactive Substances Committee. The report includes guidance for statistical analysis to be used in the scope of the OSPAR periodic evaluations of progress towards the objective of the Radioactive Substances Strategy. This report specifically addresses the issue of an appropriate methodology to assess changes in concentration of radionuclides in the marine environment, as compared with an agreed baseline. This report also investigates trend analysis techniques applied to changes in discharges from nuclear installations and the offshore oil and gas industry. The report was adopted by the Radioactive Substances Committee in 2009.

Statistical techniques for concentrations

For datasets with no values below detection limits, the Second Periodic Evaluation report chose to aggregate the original data as annual means making it possible to derive two means corresponding to the baseline and to the assessment period, with their associated standard deviations. Those two means are then compared using statistical tests, with or without any assumption regarding the distribution of data around the means. This strategy was primarily designed so that the processing of discharge data on an annual basis could be maintained. Chapter 3 of the First Periodic Evaluation deals with to the statistical methods used to compare the assessment period with the agreed baseline and for consistency purpose the ICG-Stats recommends using the same methods as in the Second Periodic Evaluation:

Both parametric and non-parametric tests are run in parallel

- Welch-Aspin (heteroscedastic form of Student t test). Data are supposed to be normally distributed but no assumption is made regarding homogeneity of variances.
- Wilcoxon-Mann-Whitney (rank test). No assumption is made regarding data distribution.

When both tests show either evidence for a significant difference or for no significant difference (5% threshold level), the conclusion is "There is a significant difference" or "There is no significant difference", respectively. When one test shows evidence for a significant difference whilst the other one does not, the conclusion is "There is some evidence for a significant difference".

When more than 80% of values are below detection limits, no statistical method is proposed because the reliability of conclusions drawn from such datasets would be tenuous and controversial.

For datasets including up to 80% of non-detects values (<DL), which are largely left unexplored in the Second Periodic Evaluation report, statistical methods, which are relevant, consistent, published and commonly accepted, can be proposed. These methods are presented and illustrated in Chapter 3. They make it possible to better use some datasets, in particular in French coastal monitoring areas. The decision flowchart in Figure 7.1 illustrates the general data processing.

. The decision flowchart in Figure 7.1 illustrates the general data processing.

Further considerations on trend detection techniques

In addition to the statistical methods used in the First and Second Periodic Evaluation reports, which are based on comparison of means against the baseline and ranking test, trend analysis techniques have been explored for discharges of radioactive substances into the marine environment. A number of tests have been studied and applied to two examples: Sellafield and La Hague.

Trend analysis tests make no distinction between the baseline and assessment period. This clearly disagrees with the Programme for More Detailed Implementation of the Strategy with regard to Radioactive Substances (the “RSS Implementation Programme”, Chapter 3 of the First Periodic Evaluation. However,

provided they are applied on an evaluation period from 1995 to present, it would coincide with the addition of the baseline period and the assessment period selected in the RSS Implementation programme. Trend analysis techniques have therefore been studied as a possible complementary approach.

Ten statistical tests representing four main types of techniques have been studied. None of them have proven robust, intuitive and meaningful in all situations. However, statistical test of three types have been found informative provided that their results are interpreted with care. Most statistical tests will be more valid when additional data becomes available. It should be noted that trend analysis techniques have not been tested on data on doses to man and biota. It is recommended that a more detailed assessment on the implementation of trend analysis techniques on OSPAR data be carried out, particularly with regard to concentrations and doses.

Provided trend analysis tests prove sufficiently robust, intuitive and meaningful, they might be used for future evaluations, when more data are available, as complementary to the statistical tests used in the Periodic Evaluations to evaluate progress.

Récapitulatif

Le présent rapport a pour but de présenter les conclusions du Groupe intersessionnel par correspondance sur les techniques statistiques applicables à la Stratégie OSPAR substances radioactives (ICG-Stats), mis sur pied dans le cadre du comité substances radioactives. Ce rapport comprend des directives pour les analyses statistiques à mener dans le cadre des évaluations périodiques de la progression dans le sens de la réalisation de l’objectif de la Stratégie substances radioactives OSPAR. Ce rapport aborde en particulier la question d’une méthodologie appropriée à l’évaluation des changements observés pour les teneurs des radionucléides dans le milieu marin, par rapport à une ligne de base déterminée. Il examine également les techniques d’analyse de tendance appliquées aux changements observés pour les rejets provenant des installations nucléaires et de l’industrie pétrolière et gazière en mer. Le comité substances radioactives a adopté ce rapport en 2009.

Techniques statistiques utilisées pour les concentrations

Pour les séries de données ne comportant aucune valeur inférieure aux limites de détection (<DL), le rapport sur la deuxième évaluation périodique a choisi d’agréger les données d’origine, sous forme de moyennes annuelles, permettant de calculer deux moyennes, correspondant à la ligne de base et à la période d’évaluation, avec leurs déviations standard correspondantes. Ces deux moyennes sont alors comparées grâce à des tests statistiques, avec ou sans hypothèse sur la distribution des données autour de ces moyennes. Cette stratégie a été adoptée principalement afin de reprendre la méthode de traitement des données de rejets, sur une base annuelle. Le chapitre 3 de la première évaluation périodique est consacré aux méthodes statistiques appliquées à la comparaison de la période d’évaluation avec la ligne de base et dans un souci de cohérence, l’ICG-Stats recommande d’utiliser les mêmes méthodes que celles de la deuxième évaluation périodique:

Les tests paramétriques et non paramétriques sont réalisés en parallèle

- Welch-Aspin (forme hétéroscédastique du test t de Student). On suppose que les données sont distribuées selon une loi normale mais on ne fait pas d'hypothèse sur l’homogénéité des variances.
- Wilcoxon-Mann-Whitney (test de rang). On ne fait aucune hypothèse sur la distribution des données.

Lorsque les deux tests démontrent, soit une différence significative, soit aucune différence significative, (seuil 5%), on conclut que "il existe une différence significative" ou "il n’existe aucune différence significative", respectivement. Lorsque l’un des deux tests révèle une différence significative mais pas l’autre, on conclut que "Il existe certaines indications d’une différence significative".

Lorsque plus de 80% des valeurs sont inférieures aux limites de détection, on ne propose aucune méthode statistique car des conclusions tirées de ces séries de données seraient ténues et discutables.

Pour les séries de données comportant jusqu'à 80% de valeurs inférieures aux limites de détection, qui n'ont, pour la plupart, pas été exploitées dans la seconde évaluation périodique, des méthodes statistiques pertinentes, cohérentes, publiées et communément acceptées, sont proposées. Ces méthodes sont présentées et illustrées dans le chapitre 3. Elles permettent de mieux exploiter certaines séries de données, en particulier dans les régions correspondant aux côtes françaises. L'organigramme décisionnel en Figure 7.1 illustre le traitement général des données.

Considérations supplémentaires sur les techniques de détection des tendances

Au delà des méthodes statistiques employées dans les rapports de la première et seconde évaluations périodiques qui se fondent sur la comparaison des moyennes par rapport à la ligne de base et un test de rangs, des techniques d'analyse de tendance ont été examinées pour les rejets de substances radioactives dans le milieu marin. Un certain nombre de tests ont été étudiés et appliqués à deux exemples: Sellafield et La Hague.

L'analyse de tendance ne fait aucune distinction entre la ligne de base et la période d'évaluation. Ceci va à l'encontre du Programme pour une mise en œuvre plus détaillée de la Stratégie substances radioactives ("Programme de mise en œuvre de la RSS", chapitre 3 de la première évaluation périodique (OSPAR, 2006)). Cependant, s'ils sont appliqués sur une période d'évaluation allant de 1995 au présent, **ceci coïncide avec le prolongement de la période de ligne de base par la période d'évaluation définies dans le programme de mise en œuvre de la RSS**. **Les techniques d'analyse de tendance ont donc été étudiées à titre d'approche complémentaire possible**.

Dix tests statistiques, représentant quatre types principaux de techniques ont été étudiés. Aucun d'entre eux ne s'est avéré robuste, intuitif et révélateur dans toutes les situations. Cependant trois types de tests se sont révélés efficaces à condition que leurs résultats soient interprétés avec précaution. La plupart des tests statistiques seront cependant plus fiables, lorsqu'avec le temps, davantage de données seront disponibles. **Il est recommandé d'entreprendre une évaluation plus détaillée sur la mise en œuvre de techniques d'analyse de tendance sur les données OSPAR, en particulier sur les teneurs et les doses.**

Si les tests d'analyse de tendance s'avèrent robustes, intuitifs et révélateurs, ils pourraient être utilisés lorsque davantage de données seront disponibles, en complément des tests statistiques utilisés dans les évaluations périodiques, pour apprécier les progrès réalisés.

1. Introduction

This report serves to strengthen the statistical analysis of radioactive substances and considers the applicable statistical techniques on trend analysis and on the treatment of results where a relatively large number of values are below the detection limit. Such analyses are essential to show what progress the Contracting Parties to the OSPAR Convention (OSPAR, 1992) are making in reducing anthropogenic inputs of radioactive substances to the North-East Atlantic, in line with the commitments that they have made in the OSPAR Radioactive Substances Strategy.

The possibility of harm to the marine environment and its users (including the consumers of food produced from the marine environment) from inputs of radionuclides caused by human activities was always a subject with which the 1972 Oslo and 1974 Paris Conventions were concerned – a concern taken over by the 1992 OSPAR Convention and taken forward in the work of implementing it. When international action to protect the marine environment from all kinds of pollution was first agreed in 1972 (OSPAR, 1972), the Oslo Convention acknowledged that radioactive substances were one of the forms of wastes and other matter to be addressed, and committed the Contracting Parties to working in the appropriate UN specialised agencies and other international bodies to promote measures to protect the marine environment against them. When the Paris Convention (OSPAR, 1974) was adopted in 1974, in order to provide for international action against land-based sources of marine pollution, the Contracting Parties undertook “to adopt measures to forestall and, as appropriate, eliminate pollution of the maritime area from land-based sources by radioactive substances”[1].

When the Oslo and Paris Conventions were up-dated and unified in 1992 to form the OSPAR Convention, stringent restrictions were included not merely on the dumping of any radioactive waste or matter (which was then temporarily halted under an international moratorium) but also on any possibility of resuming such dumping, and radioactivity was included as one of the factors against which the need for control measures on discharges from land-based sources would be judged.

When the first Ministerial Meeting under the 1992 Convention of the OSPAR Commission was held in 1998 at Sintra, Portugal, agreement was reached on both:

a. a complete and permanent ban on all dumping of radioactive waste and other matter; and

b. a strategy to guide the future work of the OSPAR Commission on protecting the marine environment of the North-East Atlantic against radioactive substances arising from human activities.

This strategy was revised and confirmed by the second Ministerial Meeting of the OSPAR Commission at Bremen in 2003. The OSPAR Radioactive Substances Strategy thus now provides that:

> “In accordance with the general objective [of the OSPAR Convention], the objective of the Commission with regard to radioactive substances, including waste, is to prevent pollution of the maritime area from ionizing radiation through progressive and substantial reductions of discharges, emissions and losses of radioactive substances, with the ultimate aim of concentrations in the environment near background values for naturally occurring radioactive substances and close to zero for artificial radioactive substances. In achieving this objective, the following issues should, *inter alia*, be taken into account:
>
> a. legitimate uses of the sea;
>
> b. technical feasibility;
>
> c. radiological impacts on man and biota.”

The Strategy further provides that:

> “This strategy will be implemented in accordance with the Programme for More Detailed Implementation of the Strategy with regard to Radioactive Substances (OSPAR, 2001) in order to achieve by the year 2020 that the Commission will ensure that discharges, emissions and losses of radioactive substances are reduced to levels where the additional concentrations in the marine environment above historic levels, resulting from such discharges, emissions and losses, are close to zero.”

The Programme for More Detailed Implementation of the Strategy with regard to Radioactive Substances (the “RSS Implementation Programme”) (OSPAR, 2001) and the agreements made at the second OSPAR Ministerial Meeting, in effect, provide that

a. the Contracting Parties will each prepare a national plan for achieving the objective of the Strategy,

b. they will monitor and report on progress in implementing those plans, and

c. the OSPAR Commission will periodically evaluate progress against an agreed baseline.

Under Annex IV to the OSPAR Convention, OSPAR is required to produce periodic assessments of the quality status of the maritime area covered by the Convention. A general assessment of the whole of the

[1] Article 5(1).

North-East Atlantic was produced in 2000, supported by five sub-regional reports. A further general assessment is planned to be produced in 2010, which will concentrate on the extent to which the aims of the thematic strategies of the OSPAR Commission have been delivered. In preparation for this, in relation to the OSPAR Radioactive Substances Strategy the following thematic assessments, it has been produced:

2006: RA-1 First Periodic Evaluation of Progress towards the Objective of the Radioactive Substances Strategy (concerning progressive and substantial reductions in discharges of radioactive substances, as compared with the agreed baseline)

2007: RA-2 Second Periodic Evaluation of the Progress towards the Objective of the Radioactive Substances Strategy (concerning concentrations in the environment as compared with the agreed baseline and including an assessment (for those regions where information is available) of the exposure of humans to radiation from pathways involving the marine environment.

2008: RA-3 An assessment (for those regions where information is available) of the impact on marine biota of anthropogenic sources (past, present and potential) of radioactive substances.

2009: RA-4 Third Periodic Evaluation of the Progress towards the Objective of the Radioactive Substances Strategy (being an overall assessment of radionuclides in the OSPAR maritime area).

2. Identification of problems

The OSPAR report entitled "Revised First periodic evaluation of progress towards the objective of the OSPAR Radioactive Substances Strategy " (OSPAR, 2006) hereafter referred to as "1PE") shows what progress the Contracting Parties to the OSPAR Convention are making to reducing anthropogenic inputs of radioactive substances to the North-East Atlantic, in line with the commitment that they made in the OSPAR Radioactive Substances Strategy. The OSPAR report entitled "Second Periodic Evaluation of Progress towards the Objective of the OSPAR Radioactive Substances Strategy" (OSPAR, 2007) hereafter referred to as "2PE") specifically address the changes in radionuclide concentrations in the marine environment, as compared with an agreed baseline. Data for ^{137}Cs, ^{99}Tc, $^{238+239}$Pu and ^{3}H in seawater and a few biological compartments are reported.

As required by the Programme for More Detailed Implementation of the Strategy with regard to Radioactive Substances (the “RSS Implementation Programme”), the main statistical data processing used in reports 1PE and 2PE consists of comparing the assessment period (2002 - 2005) with baseline elements corresponding to the reference period (1995 - 2001). Both periods are characterized by a mean value with the associated standard deviation, and statistical tests are performed to compare the two mean values at a given level of confidence. Chapter 3 of 1PE is devoted to the statistical methods used to compare the assessment period with the agreed baseline and for consistency purpose the same methods are used in 2PE. This is briefly illustrated with the following example (where values below the detection limits were purposely discarded from the dataset ^{137}Cs in seaweed from Monitoring region 2).

Starting with raw data (Figure 2.1A: individual measurement results), data were provided by Contracting Parties as annual means (Figure 2.1B). Then, two samples were built: a baseline corresponding to the mean and standard deviation of annual means from 1995 to 2001 and an assessment period corresponding to the mean and standard deviation of annual means from 2002 to 2005 (Figure 2.1C). And those two samples (the baseline and the assessment period) were compared statistically.

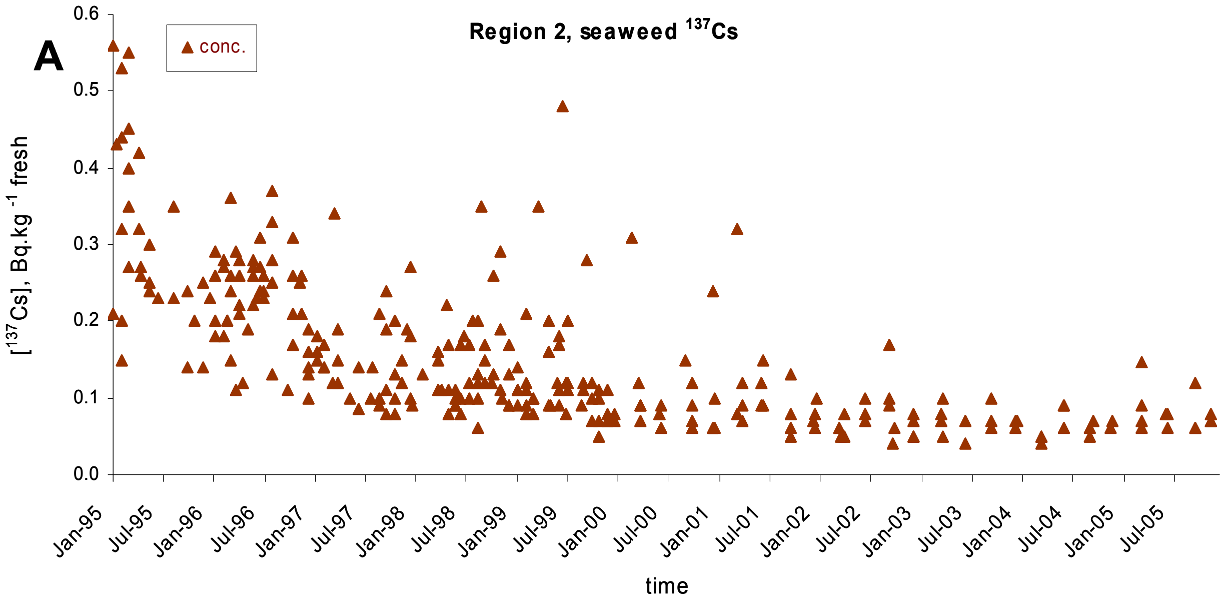


Figure 2.1. *(see legend below)*

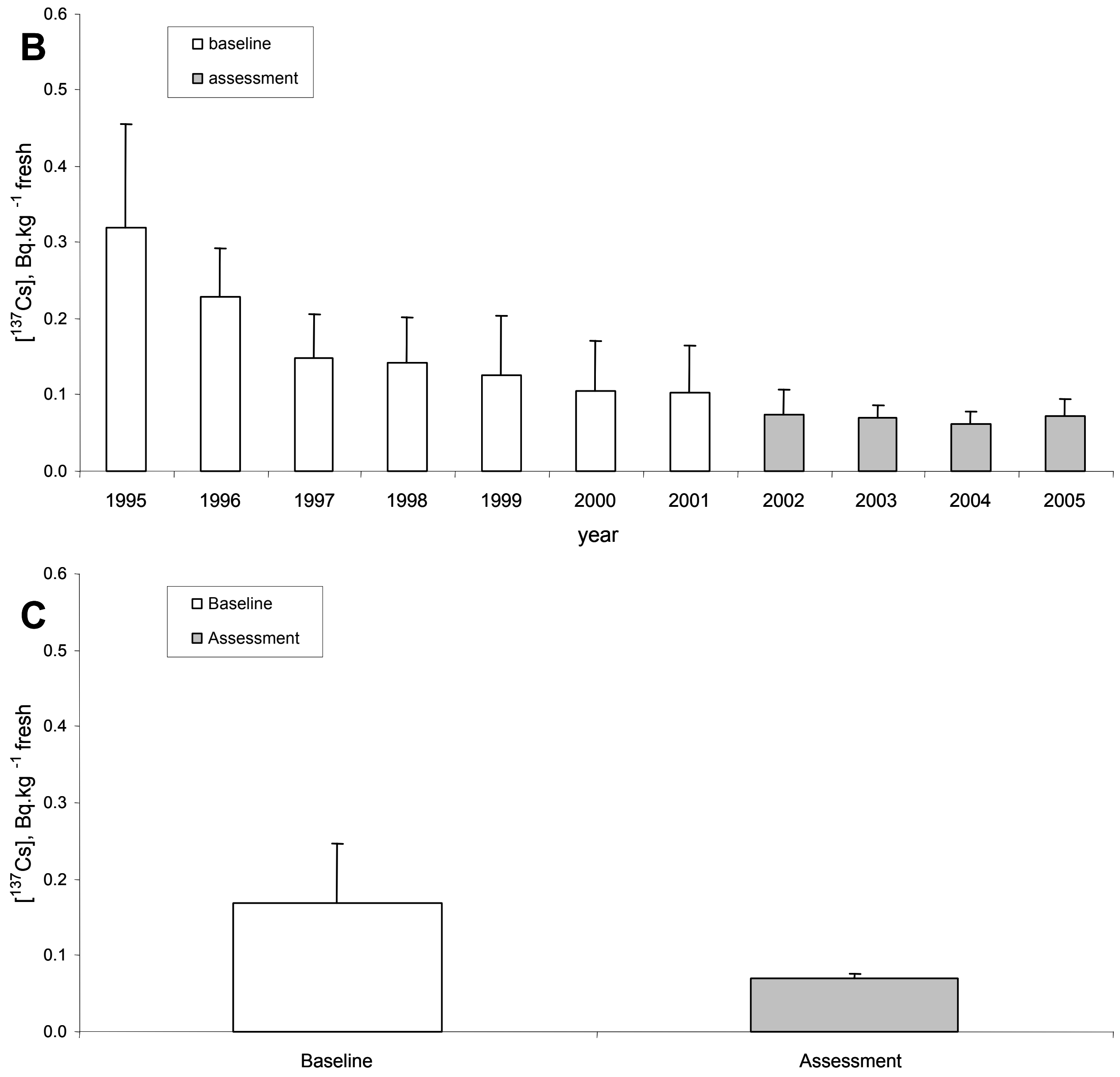


***Figure 2.1. A**: time series measurements of ^{137}Cs concentrations (Bq.kg^{-1} fresh) in seaweed from Monitoring Region 2; **B**: Data provided by Contracting Parties as annual means; **C**: Baseline and Assessment period corresponding to the mean and standard deviation of annual means from 1995 – 2001 and 2002 - 2005, respectively.*

2.1 Difficulties associated with the interpretation of means in certain instances

Time series measurements of radionuclide concentrations in compartments of the marine environment provided by Contracting Parties (CPs) may include indeterminate values when the concentrations are below the measurement detection limits[2] (DL). Those data are reported as "< DL value" (the DL value being determined for each measurement), which means that the actual radionuclide concentration value is somewhere between zero and the DL value. Such data are referred to as "non-detects values" in this

[2] The term "Detection Limit" commonly used in nuclear metrology refers to the "limit of quantification" *stricto sensu.* (US Environment Protection Agency, 2005; Currie, 2005)

document. When datasets[3] include non-detects values (<DL), annual means are derived by substituting the non-detects value by the value of the DL, itself. The resulting annual means were then reported preceded by a "lower than" sign (<) without any component for variability. This precluded any statistical analysis to assess the significance of observed differences. Dealing with those datasets is further complicated because DL values are not constant within each dataset and between different datasets. Last but not least, the proportion of non-detects values in datasets is highly variable, depending on locations, compartments and radionuclides.

A summary of strategies proposed by United States Environmental Protection Agency (EPA) and International Council for the Exploration of the Sea (ICES WGSAEM Report 2007) to deal with non-detect data is presented and discussed in Annex 1. In summary, these strategies generally recommend, either to discard non-detect values when they carry little information because they are far above actual concentrations, or to substitute non-detect values by the DL values (or DL/2 value). OSPAR should not follow these recommendations because they have been shown to introduce some important bias in many circumstances (see Annex 1).

In the 2PE, when present in datasets, non-detects values are substituted by the LD values itself and the resulting mean values are preceded by the "<" sign (lower-than) with no component to describe variability. Beside the considerable bias introduced when LD values are far above actual concentration levels, such mean values cannot be compared statistically. For example, data provided for Monitoring Regions 1, 2 and 3, corresponding to French coasts do not allow estimating the changes in radionuclide concentrations between the baseline and the assessment periods. This methodology, which consists in substituting non-detects values by LD values, is highly controversial (see Annex 1) and more relevant and consistent methods are available, as presented hereafter. A proposed methodology to deal with datasets including non-detect data is a major recommendation of this report (see Chapter 3).

2.2 Acknowledgment that, with time, additional data points make trend analysis valid

Trend analysis is a statistical approach that could be applicable to radioactive substances, in addition to comparison of means and ranking test. This approach is investigated and documented in Chapter 6 of this report. However it should be noted that trend analysis could consider data with a time resolution of a year (for discharge data) or less (month if such data are available for discharges or exact day of sampling for concentrations). It may require using all available data from 1995 to 2005. The method would explore the presence, or otherwise, of trends in the data. There would be no distinction made between the baseline and assessment period. **Trend analysis would form a complementary approach which could be carried out in parallel with the comparison of means**.

As there would be no distinction made between the baseline and assessment period, and no evaluation against the agreed baseline, it should be pointed out that such methods do not agree with the RSS Implementation Programme. On the other hand, it should be recognised that an evaluation period from 1995 would coincide with the addition of the baseline period and the assessment period selected by OSPAR in the RSS Implementation programme.

Trend analysis could form a complementary approach which could be assessed in more detail by the OSPAR Radioactive Substances Committee (RSC) in order to be proposed at the next Ministerial meeting for implementation within the Implementation Programme for future evaluations when additional data points may make it possible.

[3] A dataset corresponds to measurement results for one radionuclide, in one compartment of the marine environment (seawater, seaweed, mollusc, fish), from one geographical area (Monitoring areas 1-15), as defined in the OSPAR report 2PE.

2.3 Insufficient data

There are some cases where the quantity of data is not sufficient to perform any assessment. They include for example, regions where there are no data available for the baseline or datasets where the vast majority of data are non-detects values (measurement result below the detection limit).

There are cases where data are not sufficient to date to perform an assessment of radionuclide concentration changes with time. They include the inputs of radioactive substances to the sea from the offshore oil and gas industry de-scaling operations. Data are being collected which will make assessments possible in the future.

There are cases where data are not detailed enough to perform an assessment but where the addition of realistic assumptions make assessments possible. They include the inputs of radioactive substances to the sea from the offshore oil and gas industry from discharges of produced water and displacement water. For the latter discharges, estimated average daily quantities of discharges of produced water and displacement water have been published for each year from 1996 (OSPAR, 2006). Assuming that the average concentrations of the U-238 and Th-232 decay chains (e.g. the longer lived radionuclides Pb-210, Po-210, Ra-226 and Ra-228) remains fairly constant at long term (as it is suggested by measurements reported by Norway in the First Periodic Evaluation), statistical techniques may be used to assess the trend of the inputs of radioactive substances to the sea from the offshore oil and gas industry from discharges of produced water and displacement water. **This should also be assessed in more detail by the RSC.**

3. Selection of acceptable methodologies

3.1 Number of data points needed to achieve statistical significance

This point is only partially dealt with in the present report (see Chapters 4 and 5 for indications on concentrations and Chapter 6 for discharges). This should be covered as part of a recommended wider review of all current data and its spatio-temporal coverage (see Chapter 7).

3.2 Proposed method to deal with datasets including non-detect values

The methods proposed by ICG-Stats to deal with datasets which include non-detects values (< DL) come from recent works by environment scientists, Dr Dennis HELSEL[4] and co-workers, published in the book "Nondetects And Data Analysis: Statistics for Censored Environmental Data" (Helsel, 2005). The statistical techniques are inspired by those widely used in the fields of medical sciences or in systems-engineering (reliability analysis) (Lee and Helsel, 2007).

The authors recommend considering two cases, depending on the proportion of non-detects values present in the dataset:

- up to 80% of non-detects values
- more than 80% of non-detects values (see Chapter 5 for more details on the choice of that 80% cut-off threshold).

These methods can be used to describe the datasets with relevant statistical parameters and to make comparisons amongst datasets, for example the baseline and the assessment period datasets.

3.3 Methodologies appropriate to different contexts

Three clearly identified contexts may now be identified:

A. dataset including NO non-detect values (<DL)
The methodology adopted in the report 1PE and 2PE is kept (comparisons of means from the assessment period with the baseline using both parametric and non-parametric statistical tests)

B. dataset including up to 80% non-detect values (<DL)
The methods published in Helsel (2005) and described in the present report are proposed (see Chapter 4).

C. dataset including more that 80% non-detect values (<DL)
Data are considered as insufficient and no assessment is performed (see Chapter 5)

These methods are presented and illustrated with an example dataset (^{137}Cs in seaweed from Region 2, see Map 1) provided by France, which includes non-detects values (see Chapter 4). Calculations were undertaken using the R software[5] and the NADA add-on package (Lee and Helsel, 2005; 2007).

[4] http://www.practicalstats.com/nada/

[5] http://cran.r-project.org.
The NADA Library can be downloaded from the http://cran.r-roject.org/src/contrib/Descriptions/NADA.html

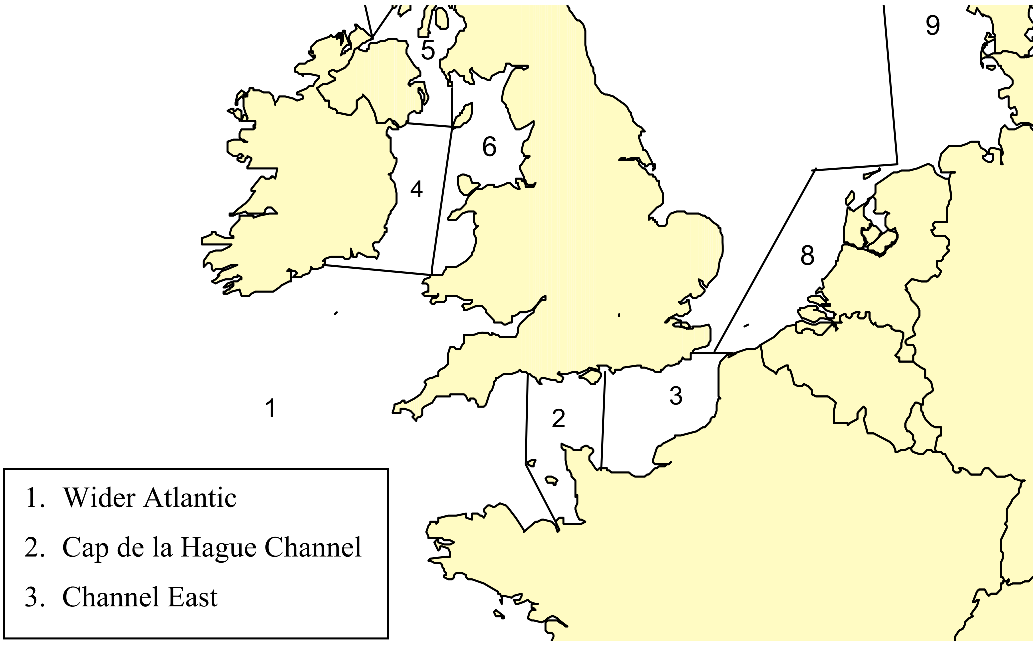


Figure 3.1. *Monitoring regions identified for the establishment of baselines on concentrations of radioactive substances*

Using this same methodology, the results of the following datasets from France are summarized in Annex 2:

- Monitoring region 1: ^{137}Cs in seaweed
- Monitoring region 2: ^{3}H in seawater
- Monitoring region 3: ^{137}Cs in seaweed

3.4 Trend identification techniques

Trend identification techniques have been discussed several times in the context of OSPAR. A number of statistical tests have been identified as possibly being using for the Radioactive Substances Strategy. These include: Kendall's Tau Correlation, Mann-Kendall test, Theil Slope test, Pearson's Correlation, Model Utility Test for Simple Linear Regression Model, Spearman Correlation, Independent two sample heteroscedastic "t" test, Wilcoxon Rank Sum test, Mann-Whitney test, Fryer and Nicholson Lowess test , and Lag 1 autocorrelation test. They can be divided in two categories: comparison of means between two periods, and trend analysis on the whole period. In accordance with the RSS implementation programme, which states that progress should be evaluated against a fixed baseline represented by the period 1995 - 2001, tests based on comparison between the baseline and the assessment period were the only methods used in the 1st evaluation report on discharges and in the 2nd evaluation report, on concentration and dose. This ICG went further in the application of trend analysis and examples of the application of both 'comparison of means' and 'trend analysis' techniques to OSPAR discharge data are presented in Chapter 6. No attempt has been made at this stage of preliminary investigation to apply trend analysis techniques to OSPAR concentration and dose data. This should be done if such techniques were to be included in the RSS Implementation Programme.

4. Trial applications: Concentrations

4.1 Graphical presentation of datasets

Before any statistical processing, **it is proposed that a graphical presentation of the dataset is produced in order to outline its crucial characteristics**. The advantages of this are as follows:

- a clear distinction between data corresponding to actual radionuclide concentrations and to non-detects values (< DL);
- outliers clearly pointing out whether they may correspond to radionuclide concentration (typing errors) or to DL values (far above actual concentrations); and
- dispersion of data, as regards date of sampling and/or concentration level.

An example is proposed on the following graph:

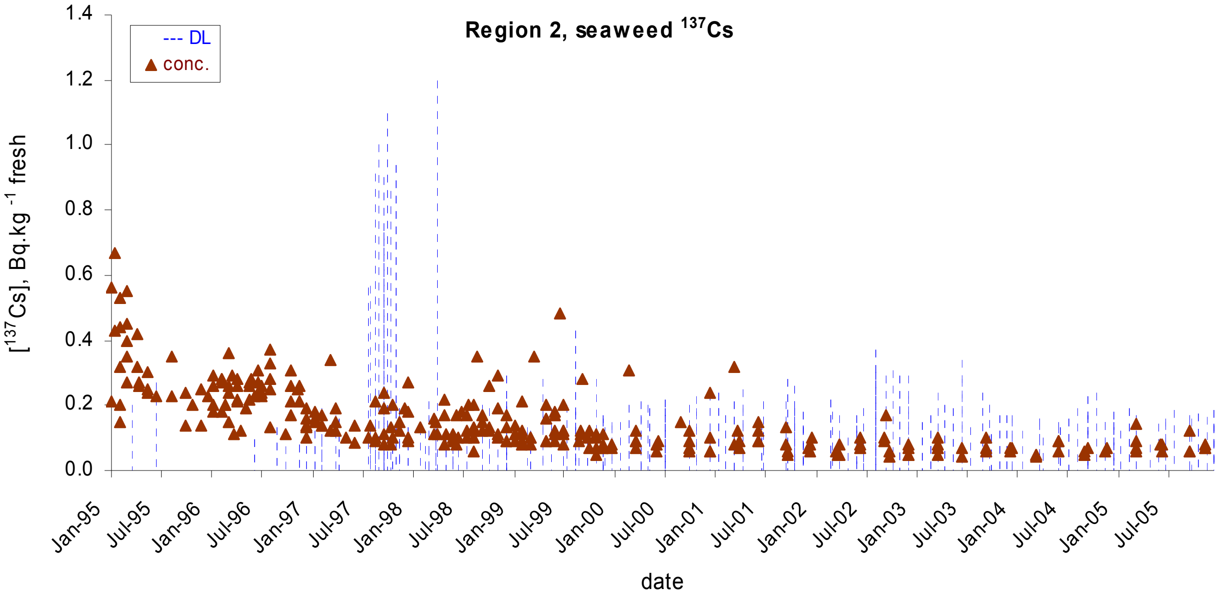


Figure 4.1. *Time series measurements of ^{137}Cs concentrations (Bq.kg^{-1} fresh) in seaweed from Monitoring region 2 (triangle symbols). Non-detects values are represented by a vertical dotted line between zero and the DL value which means that the actual value lies within this interval.*

4.2 Statistical description of datasets

For one Monitoring region, one compartment and one radionuclide (one dataset), the statistical parameters are estimated. Depending on the total number of observations and the proportion of non-detects values, those parameters are estimated using the following methods, cited in Table 4.1, and briefly presented in Annex 3.

Table 4.1. *Methods to estimate statistical parameters for each dataset, depending on the total number of observations and the proportion of non-detects values.*

	< 50 observations	**> 50 observations**
< 50% of non-detects values	Kaplan Meier	Kaplan Meier
50% - 80% of non-detects values	Robust ros (regression on order statistics)	mle (maximum likelihood estimation)
> 80% of non-detects values	See Chapter 5	

Statistical parameters estimated for the example dataset (^{137}Cs in seaweed from Monitoring region 2) are given in Table 4.2. Firstly, two data sets are compared; annual means reported during the baseline period (1995 - 2001) and the assessment period (2002 - 2005).

Table 4.2. *Statistical parameters describing dataset ^{137}Cs in seaweed from Monitoring region 2. (1) Total number of observations; (2) percentage of non-detects values; (3) number of different detection limit values, [min; max] lowest and highest DL values; (4) lowest and highest detected (>DL) values.*

Period	**Tot No.** (1)	**non-detects (%)** (2)	**No. DLs [min; max]** (3)	**Detects [min; max]** (4)	**median**	**mean**	**Standard deviation**
Baseline (1995 - 2001)	337	96 (28.49%)	33 [0.08;1.20]	[0.05 ; 0.67]	0.12	0.16	0.10
Assessment (2002 - 2005)	118	66 (55.93%)	21 [0.07;0.37]	[0.04 ; 0.17]	0.07	0.07	0.02

Alternatively, datasets can be assessed by calculating the statistical parameters on an annual basis. Two mean values with their associated standard deviation can then be derived; annual means and period means for the baseline and assessment periods, as given in Table 4.3. It can be noted that annual means are calculated from series which may include non-detect values (Table 4.3, column 2) whilst means for the baseline and the assessment periods are derived from these annual means as if they were true values .

***Table 4.3:** Statistical parameters describing dataset of ^{137}Cs in seaweed from Monitoring region 2 on an annual basis. (1) Total number of observations; (2) percentage of non-detects values; (3) number of different detection limit values, [min; max] lowest and highest DL values; (4) lowest and highest detected (>DL) values.*

Period	**Tot No** (1)	**non-detects (%)** (2)	**No. DLs [min; max]** (3)	**Detects [min; max]** (4)	**median**	**annual mean**	**Ann. Std. Dev.**	**Period mean**	**Period Std. Dev.**
1995	32	2 (6.3%)	2 [0.20;0.27]	[0.14;0.67]	0.26	0.31	0.14	**0.16**	**0.08**
1996	60	6(10.0%)	4 [0.08;0.17]	[0.10;0.37]	0.23	0.22	0.07		
1997	59	25 (42.4%)	19 [0.08;1.10]	[0.08;0.34]	0.12	0.13	0.06		
1998	60	19 (31.7%)	12 [0.09;1.20]	[0.06;0.35]	0.11	0.13	0.06		
1999	64	18 (28.1%)	11 [0.12;0.43]	[0.05;0.48]	0.10	0.12	0.07		
2000	31	12 (38.7%)	8 [0.15;0.24]	[0.06;0.31]	0.09	0.10	0.06		
2001	32	15 (46.9%)	10 [0.14;0.28]	[0.05;0.32]	0.09	0.10	0.06		
2002	32	16 (50.0%)	13 [0.07;0.37]	[0.04;0.17]	0.04	0.07	0.02	**0.06**	**0.01**
2003	29	17 (58.6%)	10 [0.11;0.34]	[0.04;0.10]	0.07	0.07	0.02		
2004	28	17 (60.7%)	11 [0.10;0.24]	[0.04;0.09]	0.06	0.06	0.02		
2005	28	15 (53.6%)	6 [0.08;0.17]	[0.08;0.15]	0.06	0.06	0.01		

This alternative methodology on an annual basis is consistent with the method used in the report 2PE to compare two mean values for the baseline and the assessment periods **Therefore, it is proposed that statistical parameters are calculated on an annual basis, as set out in Table 4.3, prior to comparison of the baseline period and assessment period means and standard deviations.**

4.3 Further comparison of the baseline and the assessment periods

Further comparison of the data from the baseline and assessment periods can be made using the following statistical techniques:

- Non-parametric generalised Wilcoxon test (add reference);
- Parametric Welch-Aspin method – a heteroscedastic form of the Student t test (add reference); and
- Non-parametric Wilcoxon-Mann-Whitney Rank test (add reference).

In addition, empirical cumulative probability distribution function estimates by the Kaplan-Meier method (add reference) can provide a useful visual representation of the baseline and assessment period datasets.

In the following paragraphs these techniques are applied to the ^{137}Cs in seaweed dataset from Monitoring region 2.

Starting from individual data

Starting with individual measurement results, two data subsets, corresponding to the baseline and the assessment periods, can be statistically compared with the non-parametric generalized Wilcoxon test (see Helsel, 2005), with no assumption regarding data distribution. Hypothesis H_0 that both subsets are distributed according to the same law of probability is tested. In other words, if H_0 is verified, this means that no significant difference exists between the assessment period and the baseline (significance threshold at the 5% level), i.e. no statistical increase or decrease. Conversely, rejection of hypothesis H_0 indicates that the difference (increase or decrease) between the two periods is statistically significant.

Comparison of the two periods with the non-parametric generalized Wilcoxon test gives:

Chisq= 84.7 on 1 degrees of freedom, p≈ 0,

indicating that a significant difference exists between the two periods at the 5% threshold level. So it can be concluded that concentrations have decreased between the baseline and the assessment period.

Figure 4.2 shows the empirical cumulative probability distribution function estimates using the Kaplan-Meier method. This probability distribution graph also indicates that concentrations have decreased between the baseline and the assessment period.

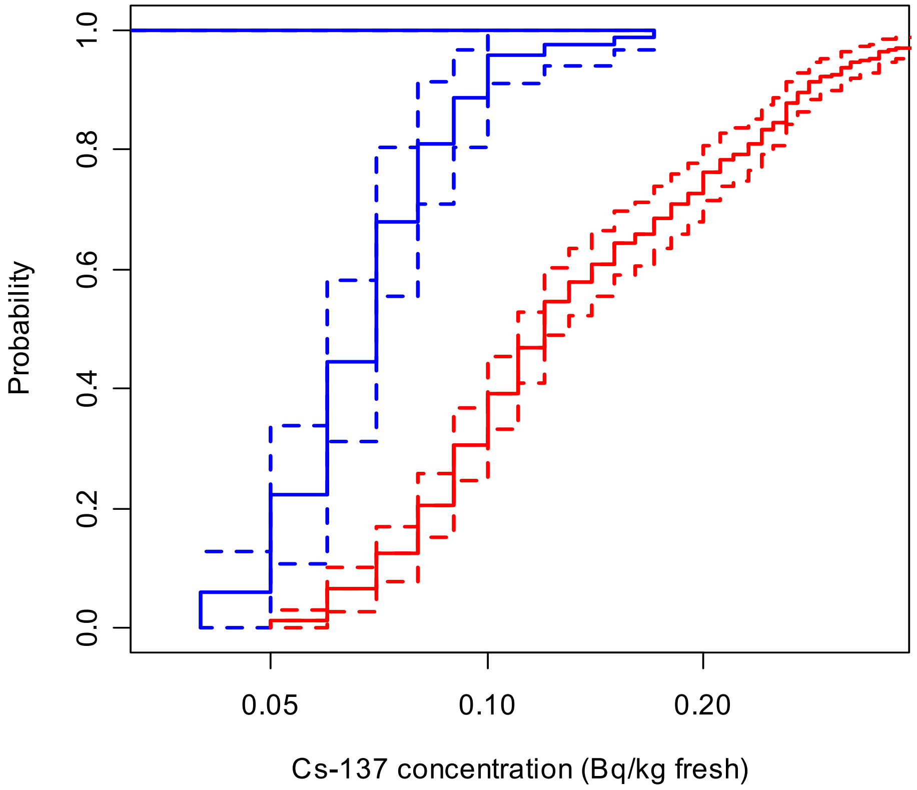


Figure 4.2: *Empirical cumulative probability distribution functions estimates by Kaplan-Meier method (right and red: baseline; left and blue: assessment period). Dotted lines correspond to 95% confidence intervals on probability values. For the assessment period (left and blue), lower values are more probable.*

Starting from annual means

With the two "period means", derived from annual means, the same statistical tests as those used in the report 1PE can be performed:

- Parametric: Welch-Aspin (heteroscedastic form of Student t test)
- Non-parametric (rank test): Wilcoxon-Mann-Whitney

Running Welch-Aspin test gives:

t = 3.0726, df = 6.187, p-value = 0.02103, Running Wilcoxon-Mann-Whitney Rank test gives:

W = *28, p-value = 0.006061,*

Both tests indicate that the difference in the concentrations recorded between the baseline and assessment periods is statistically significant (5% level), confirming that concentrations have decreased between the baseline and the assessment period.

5. Case where the dataset contains more than 80% non-detects values

When datasets include a large majority of non-detects values (<DL), it is common sense that they carry little information and their statistical description and comparison is limited in terms of reliability. Helsel (2005) recommends considering 80% of non-detects values as a cut-off threshold. The rationale for this choice is justified in Helsel (2005) as follows:

From Helsel (2005)…

Several studies found that estimation errors increase dramatically between 60 and 80% non-detects, and that above 80% non-detects any estimates are merely guesses. Therefore at 80% non-detects and above, methods that dichotomize the data into proportions of detect/nondetect should replace attempts to estimate the central location or spread of a non-detects data set.

…end of Helsel (2005) citation

It should be noted that the proportion of 80% of non-detect values is presented by Helsel as the upper limit for he application of the methods. If the estimation errors had to be reduced to ascertain conclusions in the context of the evaluation of the OSPAR strategy, a lower proportion could be recommended; Helsel suggested a proportion of 60% of non-detect values as the threshold beyond which the estimation error increase dramatically.

Since the statistical description of datasets is poorly reliable when more than 80% of non-detects are present, it is considered more consistent not to perform any statistical analysis in this case. Several datasets from OSPAR CPs may correspond to this category such as the following, provided by France (more than 80% of non-detects values):

- Monitoring region 1 : ^{137}Cs and ^{3}H in seawater
- Monitoring region 2 : ^{137}Cs in seawater
- Monitoring region 3 : ^{137}Cs and ^{3}H in seawater

Therefore, it is proposed that when more than 80% of non-detect values are present in a dataset, no statistical analyses will be undertaken. This will also be taken to be true in cases where annual data is used (as set out in Table 4.3) and only data from one specific year has more than 80% non-detect values.

6. Trial applications: Discharges (trend analysis)

This chapter outlines the application of various trend detection techniques to annual discharge data for individual radionuclides[6] from Sellafield and La Hague (nuclear sector). In addition the methods are applied to the total-α and total-β discharges from the nuclear sector installations. The approach taken is to plot and visually assess the data, assess the data for normality and finally to identify if there is a downward trend[7] in the data using the trend detection methods. Further detail on these tests can be found in Annex 4.

6.1 Sellafield discharge data

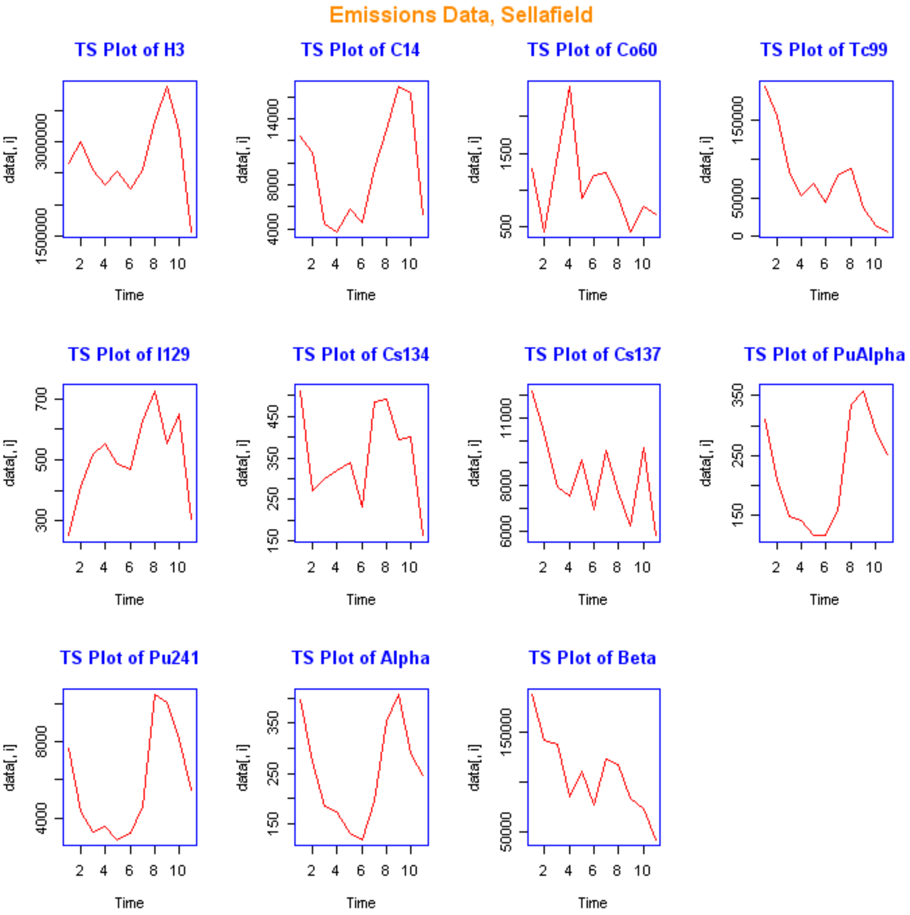


Figure 6A.1. *Time Series (TS) plots for Sellafield discharges 1995 - 2005*

Although it is necessarily qualitative and subjective, a visual examination of the discharges reported can be performed as a first step of the analysis. The visual examination of the plots in Figure 6A.1 of the Sellafield discharge data indicates that ^{99}Tc, ^{137}Cs and total-β discharges show a decreasing trend over the period under investigation. Following an initial increase, ^{60}Co also appears to display a decreasing trend.

The results of some more formal trend analysis tests to establish the presence or absence of a trend among the eleven Sellafield series are reported. The first test conducted is an independent samples t-test to establish whether discharges for the post 2001 period have decreased compared with the baseline period 1995 - 2001.

[6] Most of these nuclides have not been selected by OSPAR as objects of an individual assessment.

[7] Note: the tests applied will not identify the presence or otherwise of upward/increasing trends in the data

Tests of Normality

The t-test assumes that the underlying observations are normally distributed. The normal probability plots contained in Figures 6A.2 and 6A.3 and the formal Kolmogorov-Smirnov tests of normality (Table 6A.1) do not indicate normality. It should be borne in mind that these tests are based on very small samples and consequently are not very powerful tests. So although there is no evidence to reject normality for any of these data sets, the non-parametric Wilcoxon Rank Sum test is also applied which does not rely on the assumption of normality.

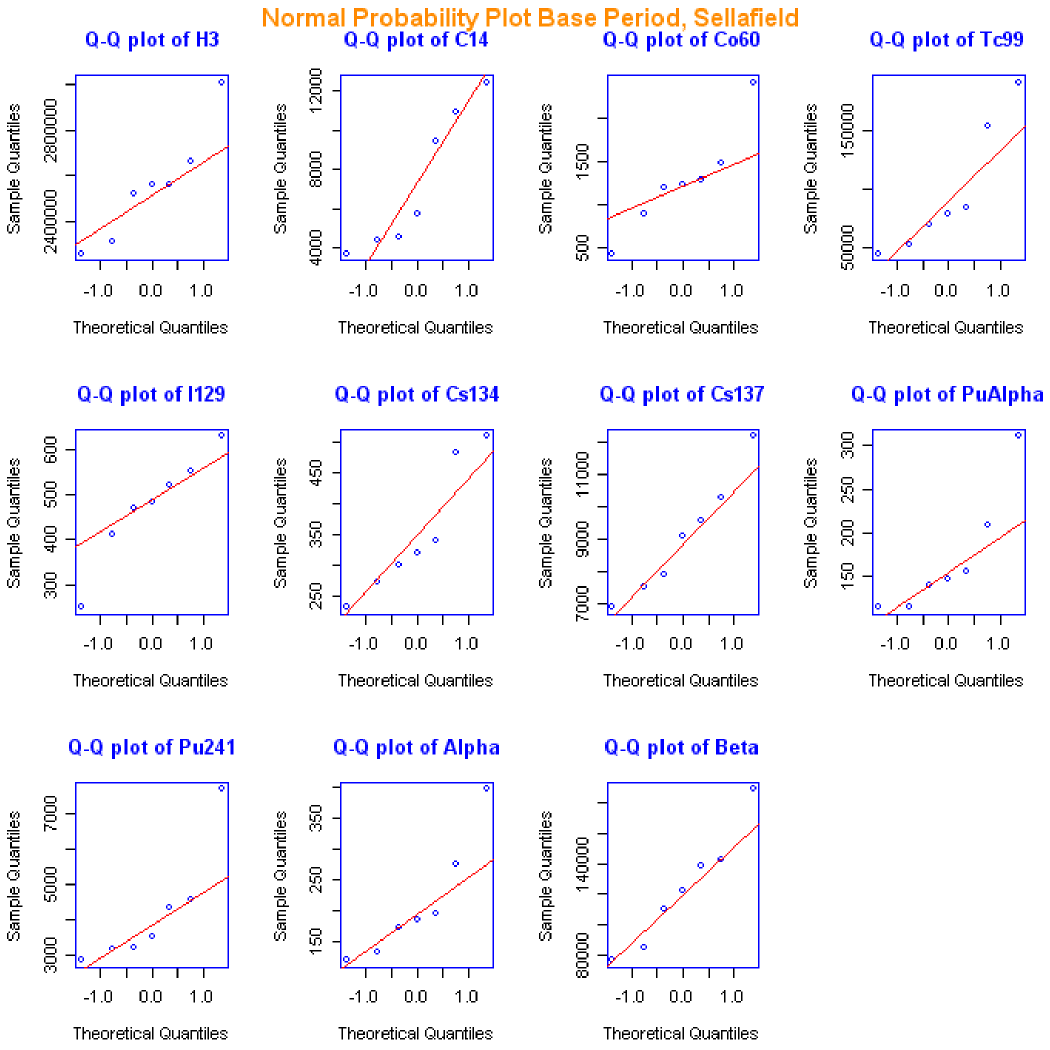


***Figure 6A.2**. Normal probability plots for Sellafield discharges: baseline period 1995 - 2001*

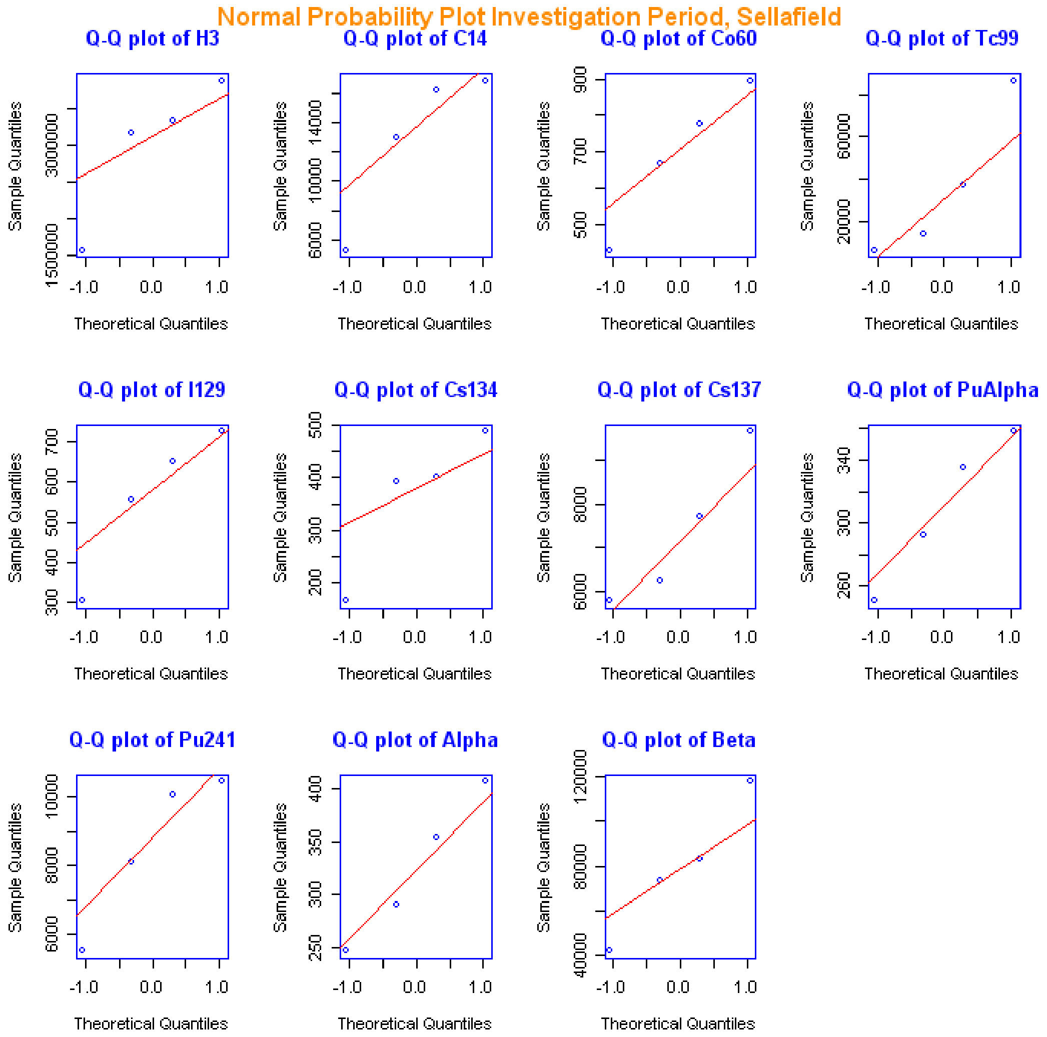


Figure 6A.3. *Normal probability plots for Sellafield discharges: assessment period 2002 - 2005*

Table 6A.1. *P-values for KS tests of normality, Sellafield discharges 1995 - 2005*

	All	**Baseline**	**Assessment**
Nuclide	**1995 - 2005**	**1995 - 2001**	**2002 - 2005**
^{3}H	0.87	0.89	0.69
^{14}C	0.57	0.72	0.88
^{60}Co	0.88	0.78	0.98
^{99}Tc	0.51	0.45	0.93
^{129}I	0.98	0.90	0.93
^{134}Cs	0.95	0.67	0.65
^{137}Cs	0.93	0.97	0.94
Pu-α	0.61	0.46	0.98
^{241}Pu	0.65	0.61	0.91
Total-α	0.88	0.56	0.99
Total-β	0.91	0.97	0.99

Two Sample t-Test and Wilcoxon Rank Sum Test[8]

***Table 6A.2**. P-values for Two Sample t-Test and Wilcoxon Rank Sum Test, Sellafield Discharges 1995 - 2005.*

Nuclide	Two Sample t-test		Wilcoxon rank sum test	
^{3}H	0.77	NO	0.92	NO
^{14}C	0.94	NO	0.98	NO
^{60}Co	*0.02*	*YES*	*0.04*	*YES*
^{99}Tc	*0.03*	*YES*	*0.05*	*YES*
^{129}I	0.77	NO	0.92	NO
^{134}Cs	0.55	NO	0.68	NO
^{137}Cs	0.08	NO	0.12	NO
Pu-α	1.00	NO	0.99	NO
^{241}Pu	0.99	NO	1.00	NO
Total-α	0.97	NO	0.98	NO
Total-β	*0.04*	*YES*	*0.04*	*YES*

Table 6A.2 clearly shows that both the t-test and the non-parametric Wilcoxon test are in agreement in indicating that the levels of the nuclides ^{60}Co, ^{99}Tc and the total-β discharges are lower for the assessment period 2002-2005 compared with the baseline period 1995 - 2001. No significant difference can be established for the other nuclides under investigation.

***Table 6A.3**. P-values for three correlation measures, Sellafield discharges 1995 - 2005*

Nuclide	Pearson's product-moment correlation		Spearman's rank correlation rho		Kendall's rank correlation tau	
^{3}H	0.57	NO	0.59	NO	0.56	NO
^{14}C	0.80	NO	0.84	NO	0.86	NO
^{60}Co	0.11	NO	0.06	NO	0.06	NO
^{99}Tc	*<0.01*	*YES*	*<0.01*	*YES*	*<0.01*	*YES*
^{129}I	0.88	NO	0.93	NO	0.96	NO
^{134}Cs	0.36	NO	0.42	NO	0.56	NO
^{137}Cs	*0.02*	*YES*	*0.04*	*YES*	*0.03*	*YES*
Pu-α	0.86	NO	0.84	NO	0.73	NO
^{241}Pu	0.91	NO	0.93	NO	0.86	NO
Total-α	0.64	NO	0.71	NO	0.62	NO
Total-β	*<0.01*	*YES*	*<0.01*	*YES*	*<0.01*	*YES*

[8] In the following tables, cases where a significant difference can be detected are displayed in red and italics

Table 6A.3 displays P-values for tests of negative correlation in the Sellafield discharge data. Three different correlation measures are used and all are in agreement indicating the presence of a negative correlation for ^{99}Tc, ^{137}Cs and total-β discharges.

Pearson's correlation approach is equivalent to using the model utility test to examine the presence of a negative slope coefficient "b" in the simple linear regression model: Y = a + bT + e, where Y is the discharges and 'T' is time.

It should also be noted that testing Kendall's correlation tau is equivalent to performing the Mann-Kendall trend detection test and the Theil slope test.

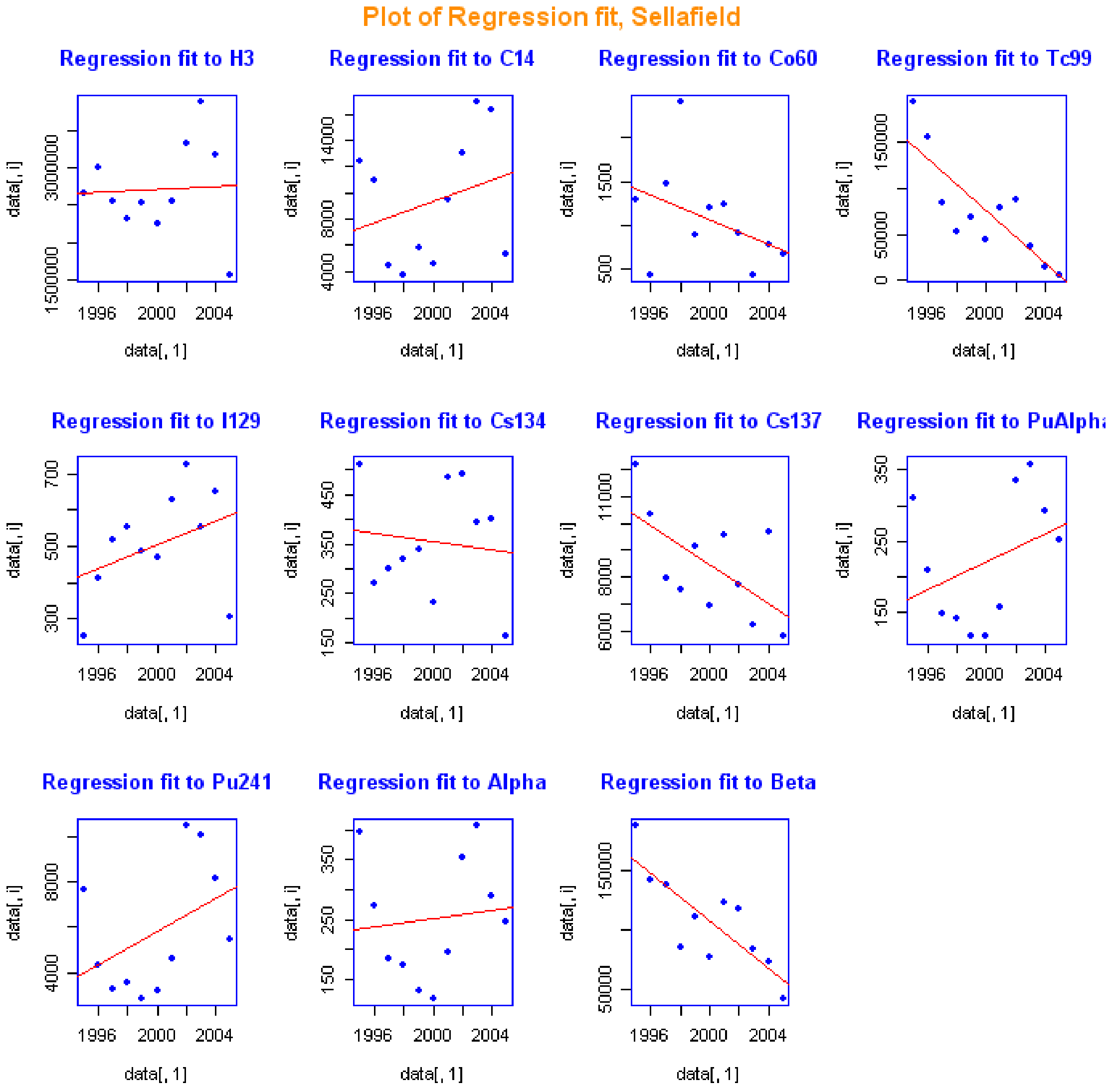


Figure 6A.4. *Linear regression fits to Sellafield discharge data 1995 - 2005*

Trend-Y-Tector

***Table 6A.4**. Results of Trend-Y-Tector test for trend using Loess method of Nicholson and Fryer, Sellafield discharges 1995 - 2005*

	Trend-y-Tector
Nuclide	**LOESS**
^{3}H	NO
^{14}C	NO
^{60}Co	*YES*
^{99}Tc	*YES*
^{129}I	NO
^{134}Cs	NO
^{137}Cs	*YES*
Pu-α	NO
^{241}Pu	NO
Total-α	NO
Total-β	***YES***

The final test applied to the Sellafield data used the Trend-Y-Tector package developed under the auspices of OSPAR which implements the test based on a Loess smoother developed by Nicholson and Fryer. The results of this test are displayed in Table 6A.4 and they indicate the presence of a downward trend in ^{60}Co, ^{99}Tc, ^{137}Cs and total-β. Figure 6A.5 shows graphically the Lowess smoothers for each series. From these plots it is clear that ^{99}Tc, ^{137}Cs and total-β display strongly negatively sloped smoothers while in the case of ^{60}Co the negative trend is not as steeply sloped.

Summary and conclusions regarding Sellafield data

All of the trend detection techniques applied to the Sellafield data suggest that there is a decreasing trend in ^{99}Tc and total-beta discharges.

All of the tests except the three correlation tests indicate that there is evidence of a decreasing trend in ^{60}Co. The correlation tests indicate that there is no trend.

All of the tests with the exception of the t-test and the Wilcoxon test suggest that there is a downward trend of ^{137}Cs. Care should be taken in interpreting the results of this series as it is evident that there is a significant increase in the year to year variability of ^{137}Cs measurements in the latter part of the period under investigation. In the presence of such variability it is impossible to produce accurate forecasts of the next year's level.

The trend detection techniques were also applied to data on other radionuclides, although no clear conclusions on whether the trends were decreasing could be made[9].

[9] Note: the tests applied will not identify the presence or otherwise of upward/increasing trends in the data

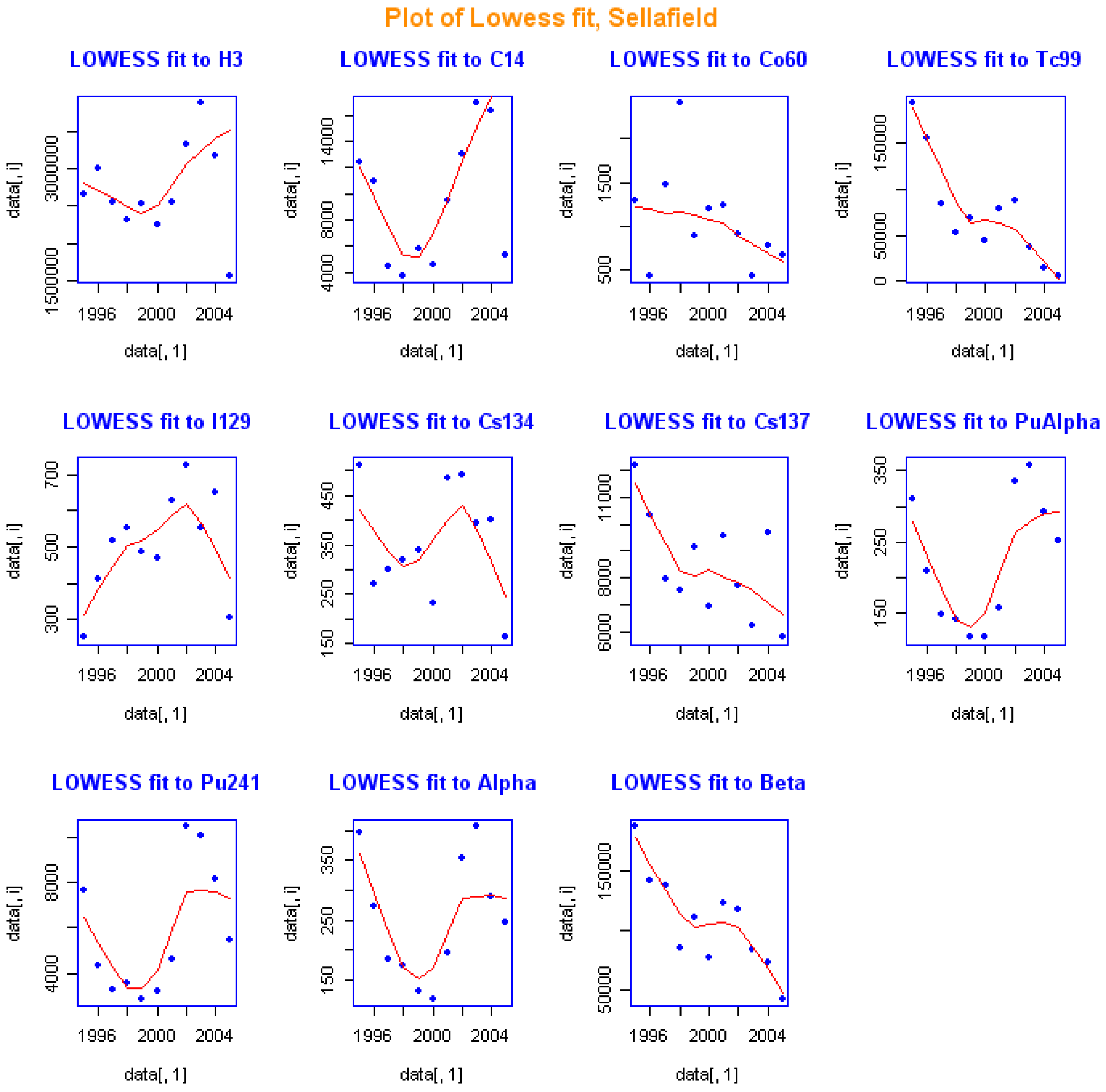


Figure 6A.5. *Lowess fits to Sellafield discharges data 1995 - 2005*

6.2. La Hague discharge data

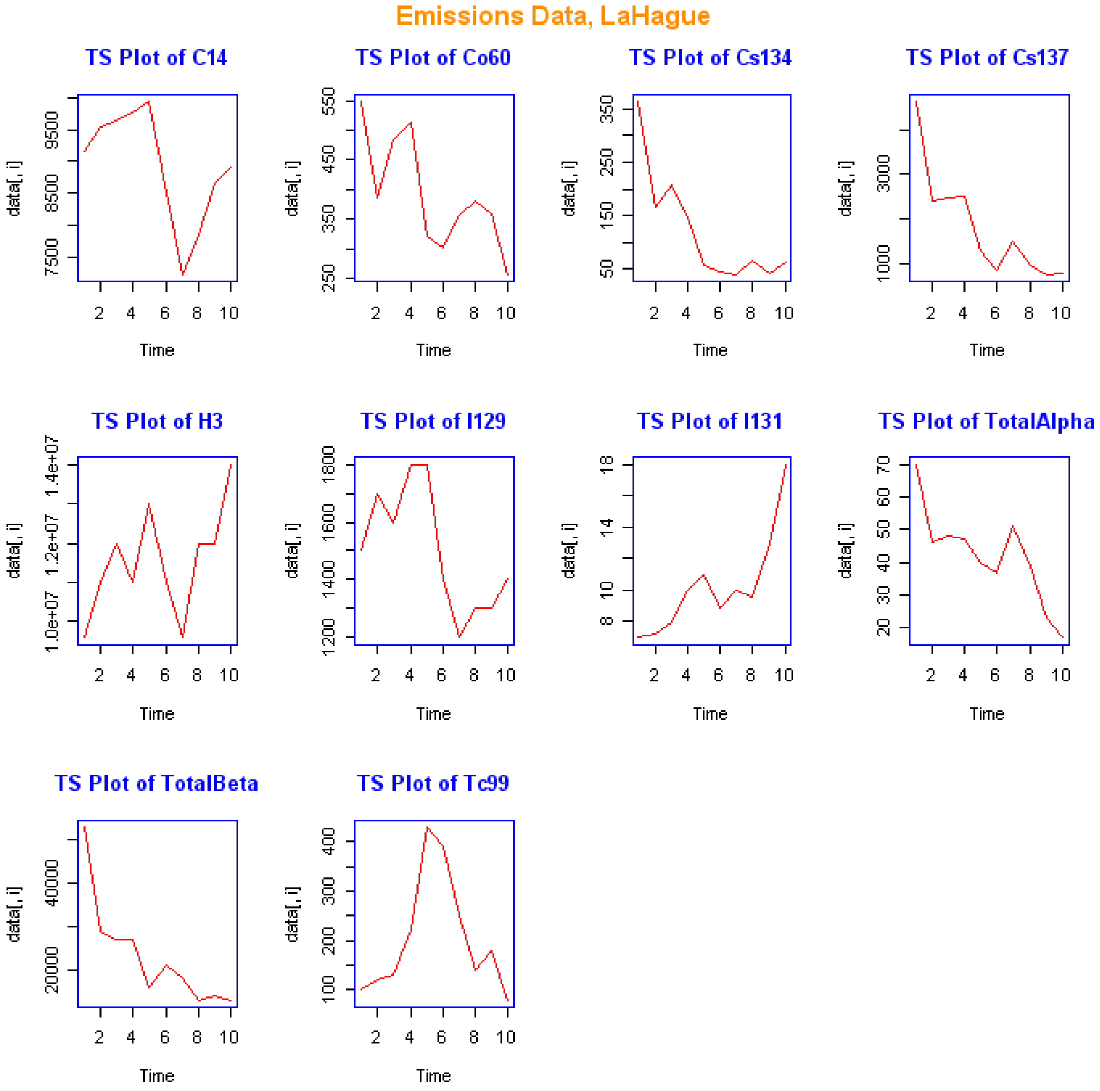


***Figure 6B.1**. Time series plots for La Hague discharges 1995 - 2004*

A visual examination of the plots in Figure 6B.1 of the La Hague discharges data, indicates that ^{60}Co, ^{134}Cs, ^{137}Cs and total α and β discharges show decreasing trend over the period under investigation. Following an initial increase, ^{99}Tc also appears to display a decreasing trend. The situation for ^{14}C and ^{129}I is more complex. The initial increase is followed by a sharp decrease and another increasing trend. However, discharges of ^{14}C and ^{129}I are lower in 2004 than in 1995.

It should be noted that the apparent increase of ^{131}I discharges do not result from a real increase but from the change of reporting procedures in France. The discharges of ^{131}I are so low that many times this radionuclide can not be detected in the effluents. When it is not detected, the old procedure mentioned that a zero discharge should be reported. The actual procedure mentions that a fraction of the detection limit multiplied by the volume should be reported. This results automatically in an increase of reported discharges for radionuclides such as ^{131}I which are not always detected in the effluents. Consequently, it does not make sense to apply any assessment for the discharges of this radionuclide for the La Hague plant.

The results of some more formal tests to establish the presence or absence of a trend among the La Hague discharge data are reported in the following sections[10].

[10] **Note**: the tests applied will not identify the presence or otherwise of upward/increasing trends in the data

Tests of Normality

Normality is assessed by considering the normal probability plots contained in Figures 6B.2 and 6B.3 and formal Kolmogorov-Smirnov tests of normality (Table 6B.1) which are unable to reject the null hypothesis of normality. Again it is important to realise that the very small sample sizes involved here imply low power for these tests.

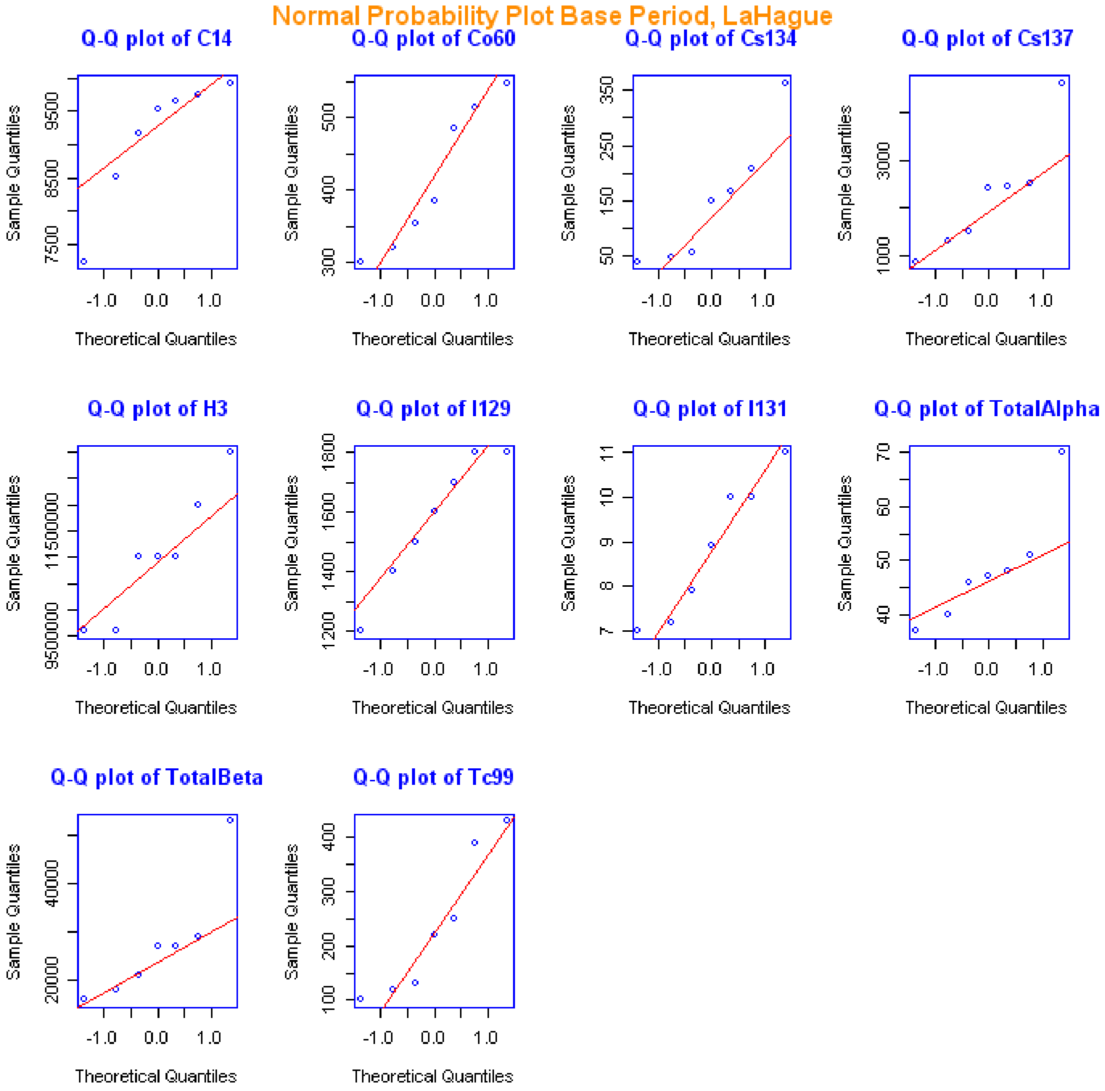


***Figure 6B.2**. Normal probability plots for La Hague discharges 1995 - 2001*

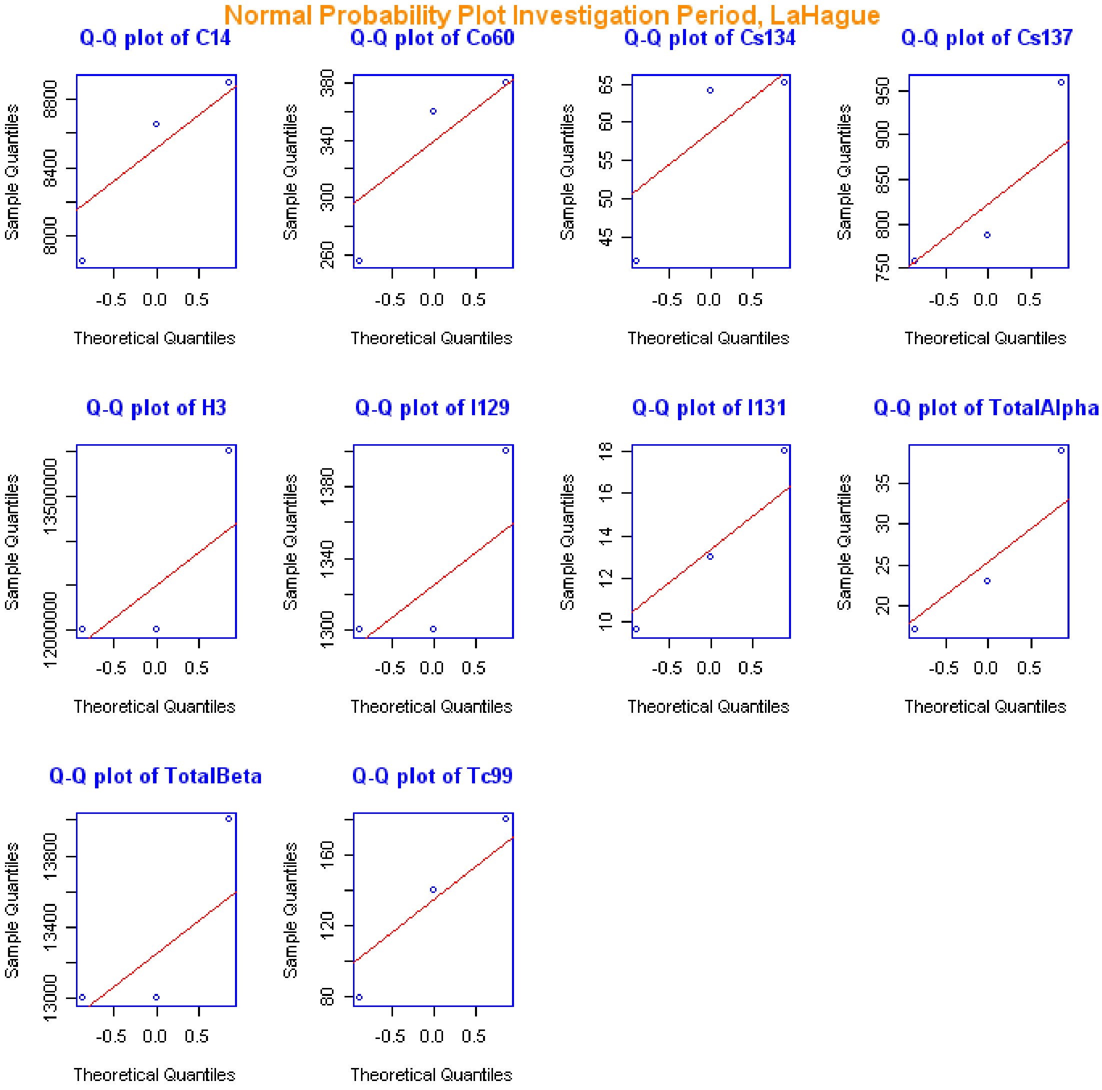


Figure 6B.3. *Normal probability plots for La Hague discharges 2002 - 2004*

Table 6B.1*. P-values for KS tests of normality, La Hague discharges 1995 - 2004*

	All	**Baseline**	**Assessment**
Nuclide	**1995 - 2004**	**1995 - 2001**	**2002 - 2004**
^{14}C	0.94	0.73	0.89
^{60}Co	0.63	0.92	0.79
^{134}Cs	0.27	0.86	0.67
^{137}Cs	0.72	0.60	0.77
^{3}H	0.95	0.87	0.76
^{129}I	0.91	0.99	0.76
^{131}I	0.66	0.94	0.99
Total-α	0.88	0.63	0.92
Total-β	0.75	0.54	0.76
^{99}Tc	0.74	0.85	0.99

Two Sample t-Test and Wilcoxon Rank Sum Test

***Table 6B.2**. P-values for Two Sample t-Test and Wilcoxon Rank Sum Test, La Hague discharges 1995 - 2004.*

Nuclide	Two Sample t-test		Wilcoxon rank sum test	
^{14}C	0.11	NO	0.13	NO
^{60}Co	0.09	NO	0.19	NO
^{134}Cs	*0.04*	*YES*	0.26	NO
^{137}Cs	*0.01*	*YES*	*0.02*	*YES*
^{3}H	0.94	NO	0.97	NO
^{129}I	*0.02*	*YES*	0.07	NO
^{131}I	0.91	NO	0.97	NO
Total-α	*0.03*	*YES*	*0.02*	*YES*
Total-β	*0.01*	*YES*	*0.01*	*YES*
^{99}Tc	0.06	NO	0.19	NO

Table 6B.2 clearly shows that both the t-test and the non-parametric Wilcoxon test are in agreement in indicating that the levels of ^{137}Cs, total-α and total-β discharges are lower for the period 2002 - 2005 compared with the baseline period 1995 - 2001. The t-test indicates a difference can be established for the ^{134}Cs and ^{129}I but the Wilcoxon test disagrees. The discrepancy is caused by the lack of power of the Wilcoxon test compared to the t-test when the variable is normally distributed. The discrepancy is caused by the lack of power of the Wilcoxon test with such a small size of the evaluation sample (size 3 for La Hague, when Sellafield evaluation sample size is 4). Simulation showed however that in some cases with samples designed to be non normal, WMW could prove more powerful than Welch-Aspin Approximate test. Moreover, when WMW is less powerful than WAA, this is offset by a higher protection against Type I error.

***Table 6B.3**. P-values for three correlation measures, La Hague discharges 1995 - 2004.*

Nuclide	Pearson's product-moment correlation		Spearman's rank correlation rho		Kendall's rank correlation tau	
^{14}C	*0.05*	*YES*	*0.05*	*YES*	0.24	NO
^{60}Co	*0.01*	*YES*	*0.01*	*YES*	*0.02*	*YES*
^{134}Cs	*<0.01*	*YES*	*0.01*	*YES*	*0.01*	*YES*
^{137}Cs	*<0.01*	*YES*	*<0.01*	*YES*	*<0.01*	*YES*
^{3}H	0.94	NO	0.94	NO	0.95	NO
^{129}I	*0.03*	*YES*	*0.02*	*YES*	0.07	NO
^{131}I	>0.99	NO	>0.99	NO	>0.99	NO
Total-α	*<0.01*	*YES*	*0.01*	*YES*	*<0.01*	*YES*
Total-β	*<0.01*	*YES*	*<0.00*	*YES*	*<0.01*	*YES*
^{99}Tc	0.55	NO	0.63	NO	0.70	NO

Table 6B.3 displays P-values for tests of negative correlation in the La Hague discharge data. Three different correlation measures are used and all three are in agreement indicating the presence of a negative correlation for ^{60}Co, ^{134}Cs, ^{137}Cs, total-α and total-β discharges. Furthermore the Pearson's and Spearman's correlation tests indicate negative correlations for ^{14}C and ^{129}I which are not strongly indicated by Kendall's test.

The plot showing a regression fit to ^{14}C in Figure 6B.4 is quite informative here, the regression line clearly displays a negative slope and noting that this is equivalent to the negative Person's correlation. These further prove a decreasing trend of ^{14}C discharges over the period. However a visual examination shows that for the last four years there has actually been an increasing trend in ^{14}C levels.

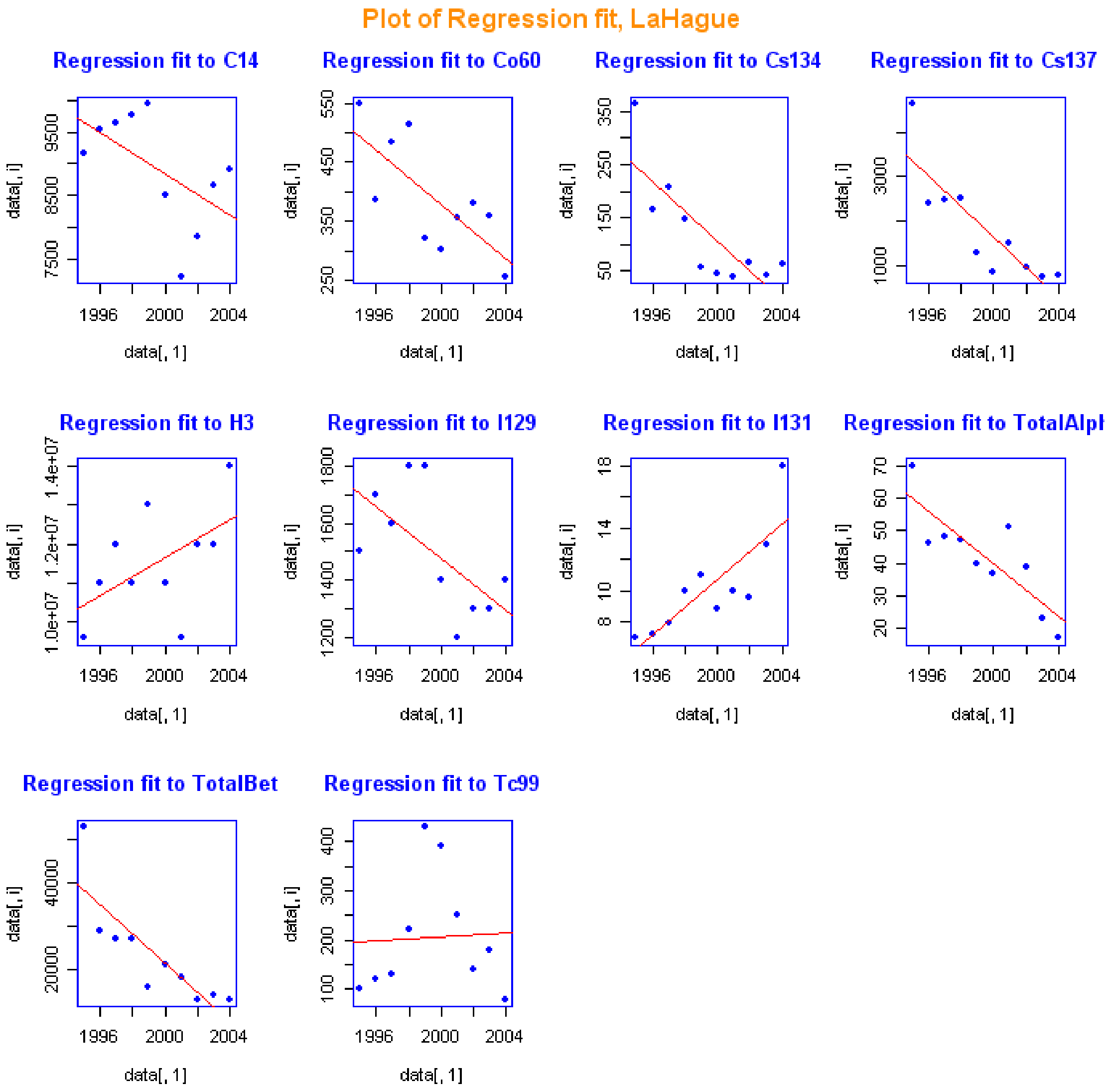


__Figure 6B.4__. Linear regression fits to La Hague discharges data 1995 - 2004

Trend-Y-Tector

***Table 6B.4**. Results of Trend-Y-Tector test for trend using Lowess method of Nicholson and Fryer, La Hague discharges 1995 - 2004*

	Trend-y-Tector
Nuclide	**LOESS**
^{14}C	*YES*
^{60}Co	*YES*
134Cs	*YES*
^{137}Cs	*YES*
^{3}H	NO
^{129}I	*YES*
^{131}I	NO
Total-α	*YES*
Total-β	*YES*
^{99}Tc	NO

The final test applied to the La Hague discharge data used the Trend-Y-Tector package developed under the auspices of OSPAR which implements the test based on a Lowess smoother developed by Nicholson and Fryer. The results of this test are displayed in Table 6B.4 and they indicate the presence of a downward trend in ^{14}C, ^{60}Co, ^{134}Cs, ^{137}Cs, ^{129}I, total-α and total-β.

Figure 6B.5 shows graphically the Lowess smoothers for each discharge series. From these plots it is clear that that the Lowess smoothers for ^{14}C and ^{129}I both display increasing trends in the most recent years. The formal Trend-Y-Tector test does not allow this behaviour to alter its conclusion of an overall decreasing trend for the entire period under investigation.

Summary and Conclusions regarding La Hague discharges

- All of the Trend Detection techniques suggest there is a decreasing trend in the ^{137}Cs, total-α and total-β discharges.
- There is agreement on ^{134}Cs with only the Wilcoxon test not detecting a trend (but this latter test lacks power).
- Several tests (Pearson, Spearman, regression fit, trend-y-tector) show a decreasing trend for ^{14}C and ^{129}I as well as ^{60}Co (same tests plus Kendall) while other tests are not able to detect such trend.
- The two series ^{14}C and ^{129}I display similar behaviour as can be seen from Figure 6B.1. The discharges increase from 1994 to 1999 and then a decrease is observed from 1999 to 2001 followed by a smaller increasing trend since 2001[11].
- The Trend-Y-Tector, Pearson and Spearman tests are confused by the sharp drop around 2000 into detecting a downward trend followed by a smaller increasing trend. This is a weakness of these tests which have difficulties with short term variability such as sharp drops.
- With regard to the ^{60}Co data while it is not detected by the t-test or the Wilcoxon test the other tests indicate the presence of a decreasing trend. In this case, while it appears a trend is present

[11] Note: the tests applied will not identify the presence or otherwise of upward/increasing trends in the data.

there is also substantial year-to-year variability which is sufficient to cause the t-test and the Wilcoxon test to be incapable of detecting a decrease in levels from the Baseline period prior to 2001 to the latter period from 2001 to 2004.

- The apparent increase of ^{131}I is artificial[12]. The increase in reported data does not results from increase of actual discharges but from a change in the reporting procedure which has been modified in line with the European Commission recommendation of 18 December 2003 on standardised information on radioactive airborne and liquid discharges into the environment from nuclear power reactors and reprocessing plants in normal operation.

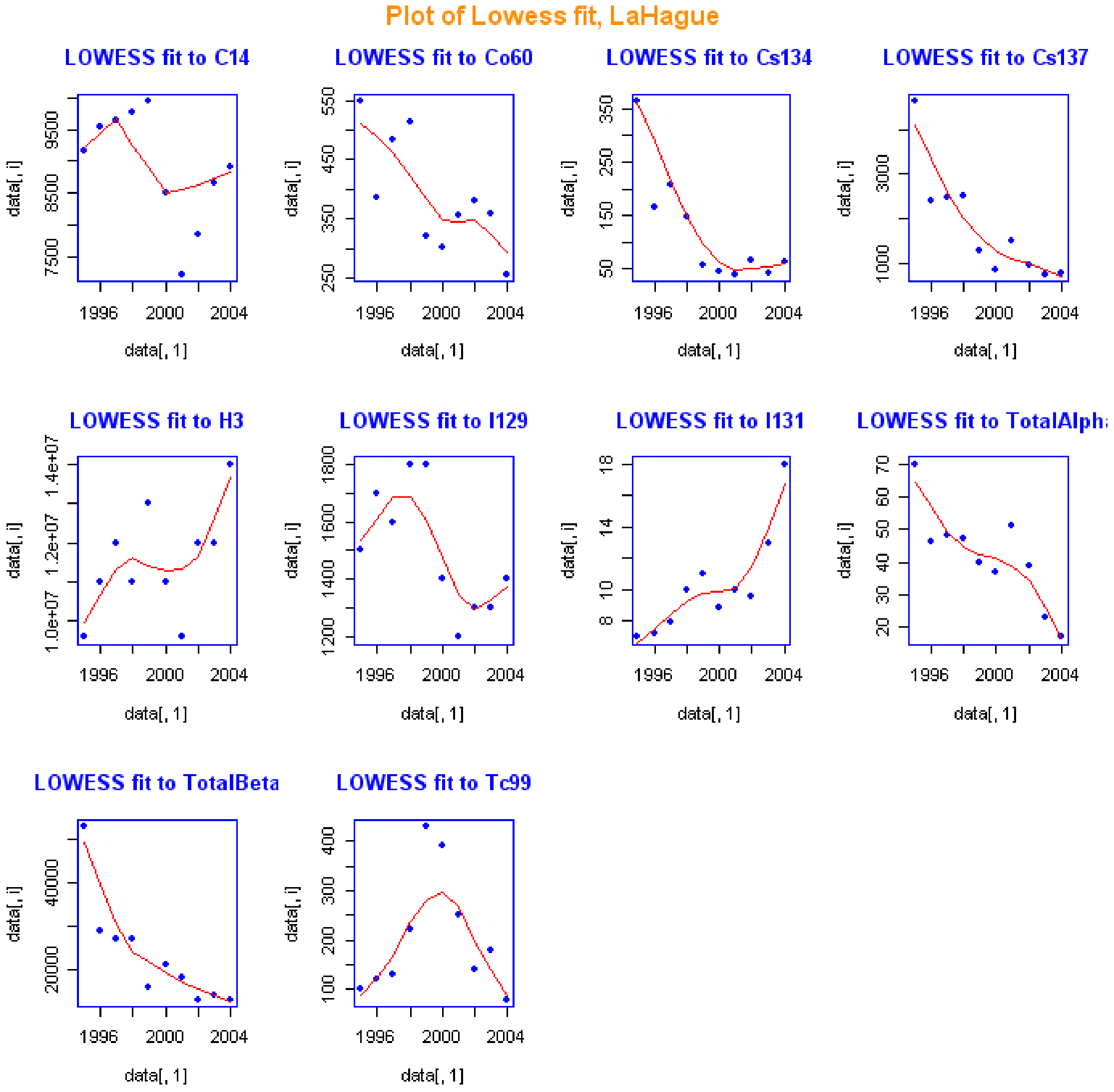


Figure 6B.5. *Lowess fits to La Hague discharges data 1995 - 2004*

[12] Note: the tests applied will not identify the presence or otherwise of upward/increasing trends in the data.

6.3 Total-α & Total-β

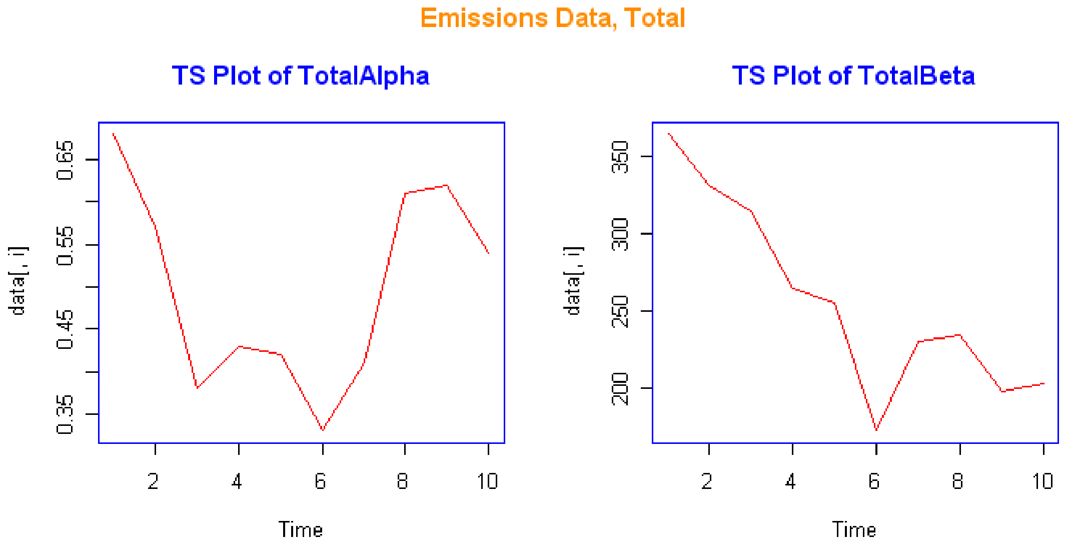


Figure 6C.1*. Time series plots for overall discharges 1995 - 2004*

The visual examination of the plots in Figure 6C.1 of the overall OSPAR discharges data from the nuclear sector, suggests that the total-β discharges show decreasing trend over the period under investigation.

Tests of Normality

The Normal Probability plots contained in Figure 6C.2 and formal Kolmogorov-Smirnov tests of normality (Table 6C.1) are unable to reject the null hypothesis of normality. As before it should be noted that these tests are based on very small samples and consequently are not very powerful tests.

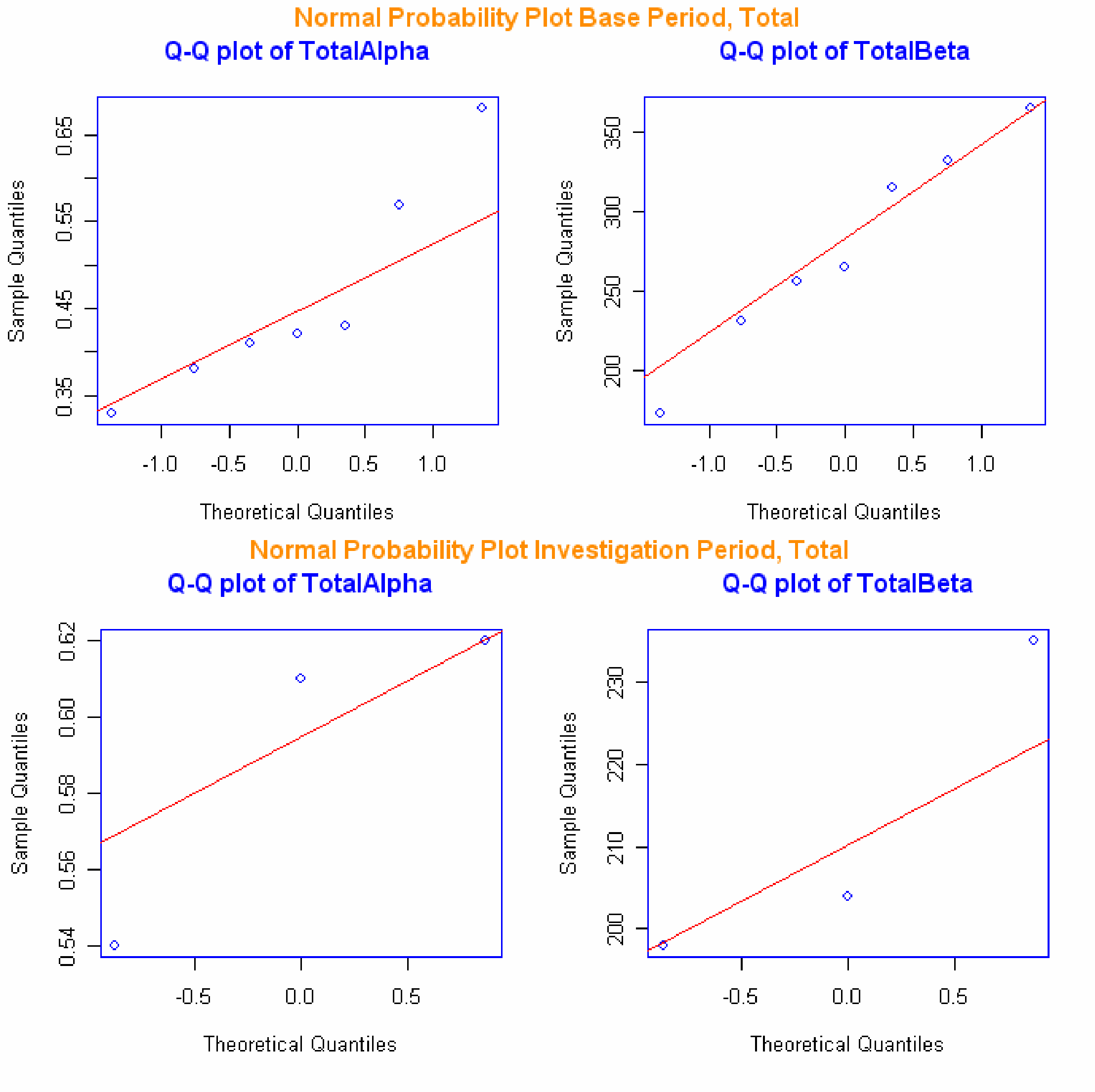


***Figure 6C.2**. Normal probability plots for overall discharges 1995 - 2001 and 2002 - 2004*

***Table 6C.1.** P-values for KS tests of normality, overall discharges 1995 - 2004*

	All	**Base**	**Investigation**
Nuclide	**1995 - 2004**	**1995 - 2001**	**2002 - 2004**
Total-α	0.650669	0.42	0.75
Total-β	0.9498465	0.99	0.79

Two Sample t-Test and Wilcoxon Rank Sum Test

***Table 6C.2**. P-values for Two Sample t-Test and Wilcoxon Rank Sum Test, overall discharges 1995 – 2004.*

Nuclide	Two Sample t-test		Wilcoxon rank sum test	
Total-α	0.981	NO	0.942	NO
Total-β	*0.023*	*YES*	0.092	NO

Table 6C.2 clearly shows that the t-test indicates the total-β levels are lower for the period 2002 - 2005 compared with the baseline period 1995 - 2001. The non-parametric Wilcoxon test is significant at a significance level α=0.10 but not at α=0.05. We recall however again that the Wilcoxon test has lower power than the t-test. Both tests are in agreement with regard to the total-α discharges and indicate no decreasing trend.

***Table 6C.3**. P-values for three correlation measures, overall discharges 1995 - 2004*

Nuclide	Pearson's product-moment correlation		Spearman's rank correlation rho		Kendall's rank correlation tau	
Total-α	0.515	NO	0.500	NO	0.500	NO
Total-β	*0.001*	*YES*	*0.002*	*YES*	*0.001*	*YES*

As can be seen in Table 6C.3 the three correlation measures used are all in agreement indicating the presence of a negative correlation for total-β discharges but not for total-α discharges. Figure 6C.3 displays the same information visibly.

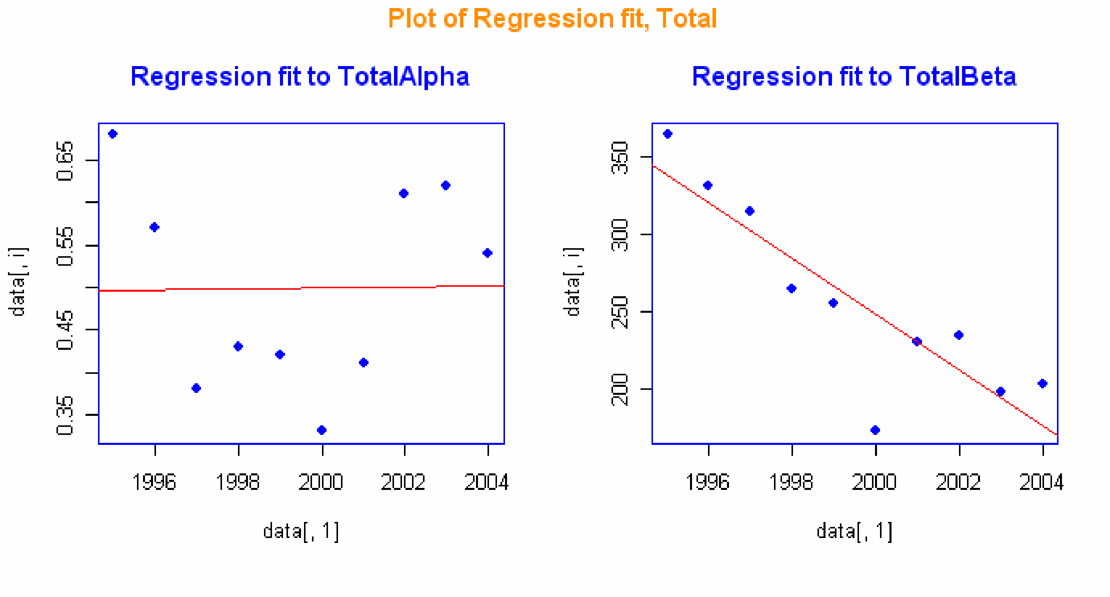


***Figure 6C.3**. Linear Regression fits to Overall Discharges data 1995 - 2004*

Trend-Y-Tector

Table 6C.4. *Results of Trend-Y-Tector test for trend using Loess method of Nicholson and Fryer, overall discharges 1995 - 2005*

	Trend-y-Tector
Nuclide	**LOESS**
Total-α	NO
Total-β	*YES*

The results of the Trend-Y-Tector Lowess test are displayed in Table 6C.4 and indicate the presence of a downward trend in total-β but not for total-α discharges. Figure 6C.4 shows graphically the Lowess smoothers for each series.

Summary and Conclusions regarding overall data on nuclear sector discharges

The conclusions here are straightforward: a negative trend is present in the total-β series but not in the total-α series.

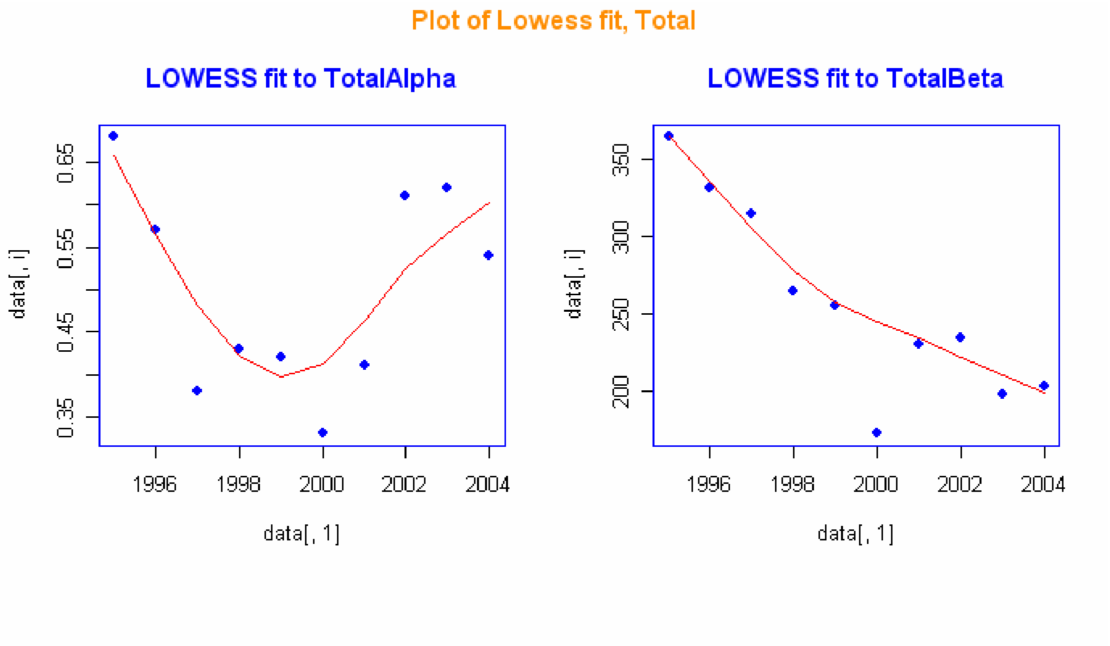


Figure 6C.4. *Lowess fits to Overall Discharges data 1995 - 2004*

6.4 General Observations on Trend Detection Procedures

According to previous statistical studies within the context of OSPAR, it is a requirement that the statistical methods used for Periodic evaluations should respond to the need to be:

a. ***robust*** – that is, to be both routinely applicable to many data-sets and as insensitive as possible to statistical assumptions;

b. ***intuitive*** – that is, for the results of the analysis to be understandable without a detailed understanding of statistical theory;

c. ***revealing*** – that is, to provide easy access to several layers of information about the major features of the data.

Four different types of formal tests have been explored for discharges:

Type 1(monotonous trend analysis):

- Kendall's Tau Correlation
- Mann-Kendall test
- Theil Slope test
- Pearson's Correlation
- Model Utility Test for Simple Linear Regression Model
- Spearman Correlation

Type 2 (comparison of means and ranking test, applied in the 1st and 2nd evaluation report):

- Independent two sample heteroscedastic "t" test
- Wilcoxon Rank Sum test equivalent to Mann-Whitney test

Type 3 (non monotonous trend analysis):

- Fryer and Nicholson Lowess test as implemented by Trend-Y-Tector software

Type 4:

- Lag 1 autocorrelation test

Type 1: These tests all measure some form of correlation in the data. Pearson's test is equivalent to the model utility test for linear regression and it tests only for the presence of a linear trend. Such tests are not robust when the trend is not linear. The other tests test for non linear trends. In general these tests are all quite informative but they can cause misleading results if not interpreted properly. In other words, there are not intuitive. A perfect example of this is La Hague carbon14 data above. The decreasing trend in this data results from an increasing trend for the first few years, a sharp fall and then another increasing trend.

Type 2: The Type 2 tests both examine if there is a decrease in levels post 2001 compared with prior to 2001. The t-test relies on an assumption of Normality in the data. The Wilcoxon test is non-parametric and does not rely on this assumption. It is important to realise that with small datasets such as these it is not possible to accurately determine whether the underlying data really is Normal. One difficulty with these tests is that they depend upon the choice of baseline period in a quite sensitive way. Some significant decreasing trends in the data may not be picked up by these tests. In particular if the trend has caused the data to already reach lower levels prior to the end of the baseline period the tests may not be sensitive enough to pick up this. Also if there is substantial variability in the data this will cause the tests to fail to report a significant trend. It should be noted however that this lack of sensitivity is counterbalanced by a higher protection against type I errors (concluding that there is a difference when in truth there is no difference)

Type 3: The main justification for the use of these tests is, as mentioned, an assumption that the underlying process is better represented by a smoother f(t) rather than the original measurements. This may be true on the long run, with sufficient data over many years the smoother may have an advantage at detecting the long term behaviour of the time series. However in the short term the smoother can smooth out real variations in the data and so disguise some aspects of the data. .In the short term, these tests can be not revealing enough. The smoother is sensitive to certain parameters which determine the amount of smoothing undertaken. With the choice of parameters suggested by Fryer and Nicholson which is implemented in the software packages above, there is a significant degree of smoothing. So large short run variations in trend can be smoothed out and consequently missed. Again the Carbon 14 data from La Hague displays this effect.

Type4: The Lag 1 Autocorrelation test cannot be relied upon as a trend detection tool. This can be seen from the attached appendix which contains the results of the test applied to the three data sets. Such tests are therefore not relevant.

7. Conclusions, recommendations and guidelines for the statistical analysis of future assessments

The purpose of this report is to present the conclusions by the Intercessional Correspondence Group established under the OSPAR Radioactive Substances Committee (ICG-Stats) to consider statistical techniques applicable to the OSPAR Radioactive Substances Strategy. Here are some proposed guidelines for statistical analysis to be used in the scope of the OSPAR Periodic Evaluations. This report specifically addresses the issue of an appropriate methodology to assess changes in concentration of radionuclides in the marine environment, as compared with an agreed baseline. This report also investigates trend analysis techniques applied to changes in discharges of nuclear installations and offshore oil and gas industry.

7.1 Statistical techniques for concentrations

Datasets provided by Contracting Parties fall into three categories:

- datasets with all radionuclide concentrations above detection limits
- datasets including less than 80% of values below detection limits
- datasets with more than 80% of values below detection limits.

For datasets with no values below detection limits, the choice was made in the report 2PE to aggregate original data as annual means prior to deriving two means corresponding to the baseline and the assessment period, with their associated standard deviations. Those two means are then compared using statistical tests, with or without any assumption regarding the distribution of data around the means. This strategy was primarily designed to stick with the yearly basis of data processing for discharges (report 1PE). Chapter 3 of 1PE is devoted to the statistical methods used to compare the assessment period with the agreed baseline and for consistency purpose the ICG-Stats recommends to use the same methods as in 2PE.

Both parametric and non-parametric tests are run in parallel

- Welch-Aspin (heteroscedastic form of Student t test). Data are supposed to be normally distributed but no assumption is made regarding homogeneity of variances.
- Wilcoxon-Mann-Whitney (rank test). No assumption is made regarding data distribution.

When both tests show either evidence for a significant difference or for no significant difference (5% threshold level), the conclusion is "There is a significant difference" or "There is no significant difference", respectively. When one test shows evidence for a significant difference whilst the other one does not, the conclusion is "There is some evidence for a significant difference".

When more than 80% of values are below detection limits, no statistical method is proposed because the reliability of conclusions drawn from such datasets would be tenuous and controversial.

For datasets including up to 80% of non-detects values (<DL), which are largely left unexplored in the report 2PE, statistical methods (Helsel, 2005), which are relevant, consistent, published and commonly accepted, can be proposed. These methods are presented and illustrated in this (Chapter 3). They make it possible to better use some datasets, in particular in Monitoring regions corresponding to the coasts of France. The decision flowchart presented below (Figure 7.1) illustrates the general data processing.

The proposed methodology could also be submitted for validation to the ICES group on Statistics in Environmental Monitoring which provides advice to other OSPAR Committees.

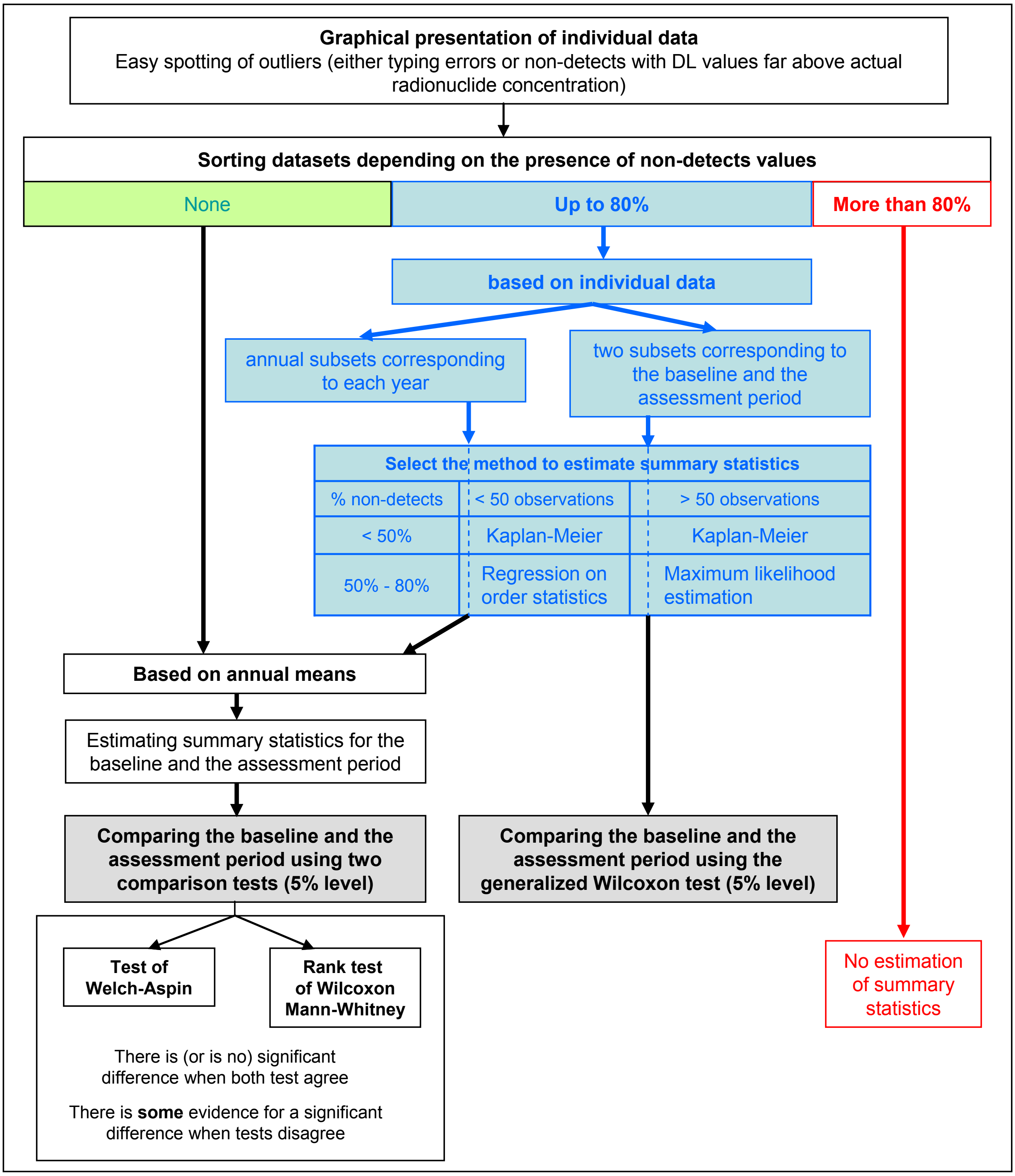
7.2 Further considerations on trend detection techniques

In addition to the statistical methods used in the 1st and 2nd evaluation report which are based on comparison of means against the baseline and ranking test, trend analysis techniques have been explored for discharges of radioactive substances into the marine environment. A number of tests have been studied and applied to two examples: Sellafield and La Hague.

Trend analysis tests make no distinction between the baseline and assessment period. This clearly disagrees with the Programme for More Detailed Implementation of the Strategy with regard to Radioactive Substances (the “RSS Implementation Programme”, Chapter 3 of 1PE (OSPAR, 2006). However, provided they are applied on an evaluation period from 1995 to present, **it would coincide with the addition of the baseline period and the assessment period selected in** the RSS Implementation programme. **Trend analysis techniques have therefore been studied as a possible complementary approach**.

Ten statistical tests representing four main types of techniques have been studied. None of them have proven robust, intuitive and revealing in all situations. However, statistical test of three types have been found informative provided that their results are interpreted with care. Most statistical tests will be more valid when additional data will be available with time. It should be noted that trend analysis techniques have not been tested to man and biota. **It is recommended to perform a more detailed assessment on the implementation of trend analysis techniques on OSPAR data, particularly on concentrations and doses.**

Provided trend analysis tests prove enough robust, intuitive and revealing, they might be used for future evaluation, when more data are available, as complements to statistical tests used in the Periodic Evaluations to evaluate progress.


Graphical presentation of individual data
Easy spotting of outliers (either typing errors or non-detects with DL values far above actual radionuclide concentration)
Sorting datasets depending on the presence of non-detects values
None
Up to 80%
More than 80%
based on individual data
annual subsets corresponding to each year
two subsets corresponding to the baseline and the assessment period
Select the method to estimate summary statistics
Based on annual means
Estimating summary statistics for the baseline and the assessment period
Comparing the baseline and the assessment period using two comparison tests (5% level)
Comparing the baseline and the assessment period using the generalized Wilcoxon test (5% level)
Test of Welch-Aspin
Rank test of Wilcoxon Mann-Whitney
There is (or is no) significant difference when both test agree
There is **some** evidence for a significant difference when tests disagree
No estimation of summary statistics

% non-detects	< 50 observations	> 50 observations
< 50%	Kaplan-Meier	Kaplan-Meier
50% - 80%	Regression on order statistics	Maximum likelihood estimation

Figure 7.1. *Decision flowchart depicting the general data processing for datasets.*

References

Currie L.A., 2004. Detection and quantification limits: basic concepts, international harmonization, and outstanding ("low-level") issues. Applied Radiation and Isotopes 61(2-3):145-149

Helsel D., 2005. "Nondetects And Data Analysis: Statistics for Censored Environmental Data", ISBN 0-471-67173-8, Ed Wiley, New-York, 250p

ICES WGSAEM, 2007. Report of the Working Group on the Statistical Aspects of Environmental Monitoring. ICES WGSAEM Report 2007; CM 2007/MHC:02

Klein J.P. and Moeschberger M.L., 2003. Survival Analysis: Techniques for censored and truncated data, Second Edition. Springer, New York, 536p

Lee L. and Helsel D., 2005. Statistical analysis of water-quality data containing multiple detection limits: S-language software for regression on order statistics, Computers & Geosciences 31, 1241-1248.

Lee L. and Helsel D., 2007. Statistical analysis of water-quality data containing multiple detection limits II: S-language software for nonparametric distribution modeling and hypothesis testing, Computers & Geosciences 33, 696-704.

OSPAR, 1972. Convention for the Prevention of Marine Pollution by Dumping from Ships and Aircraft, Oslo, 15 February 1972.

OSPAR, 1974. Convention for the Prevention of Marine Pollution from Land-Based Sources, Paris, 4 June 1974.

OSPAR, 1992. OSPAR Convention for the Protection of the Marine Environment of the North-East Atlantic, Paris, 22 September 1992.

OSPAR, 2001. Programme for the More Detailed Implementation of the OSPAR Strategy with regard to Radioactive Substances. OSPAR Agreement 2001-03.

OSPAR, 2006. Revised First Periodic Evaluation of Progress towards the Objective of the OSPAR Radioactive Substances Strategy (JAMP product RA-1). Publication 302/2006. ISBN 1-905859-40-6. http://www.ospar.org/documents/dbase/publications/p00302_Revised First Periodic Evaluation.pdf

OSPAR, 2007. Second Periodic Evaluation of Progress towards the Objective of the OSPAR Radioactive Substances Strategy, Publication 338/2007. ISBN No. 978-1-905859-77-1. http://www.ospar.org/documents/dbase/publications/p00338_Second periodic evaluation.pdf

US Environment Protection Agency (2005) Federal Advisory Committee on Detection and Quantitation for Uses in Clean Water Act Programs. http://www.epa.gov/waterscience /methods/det/faca/mtg20050929/defintionoptions.html

Annex 1: Brief discussion of previous strategies dealing with non-detect values

Summary of recommendation by International Council for the Exploration of the Sea

In its report of the Working Group on the Statistical Aspects of Environmental Monitoring (ICES WGSAEM report 2007), ICES recommends that:

"OSPAR RSC should note the way in which LOD (Limit Of Detection) values in radioactivity concentration could be treated. Data below LOD in radioactivity concentration measurements can carry poor information (when LOD are largely above the real value) or rich information (when measurements below LODs are mixed up with detected values at the same concentration levels). In the first case, all <LOD values should be discarded, in the second case, they may considered as detected values. In addition, an analytical way to take into account <LOD values in an exact way is suggested (based on the substitution of <LOD values by LOD values in some particular cases). Furthermore, it is important to have a close look at the data and to have information about the way labs work."

Summary of recommendation by US Environmental Protection Agency

Table A1.1*. Guidelines for Analyzing Data with Nondetects coming from EPA (Guidance for Data Quality Assessment: Practical Methods for Data Analysis; EPA/600/R-96/084; U.S. EPA, Office of Research and Development: Washington, DC, 1998).*

Percentage of Nondetects	Statistical Analysis Method
< 15%	Replace nondetects with DL/2, DL, or a very small number.
15% - 50%	Trimmed mean, Cohen's adjustment, Winsorized mean and standard deviation.
> 50% - 90%	Use tests for proportions

Comments about these methods

Here are the comments of Dennis R. Helsel about these methods, they are extracted from his paper: More than obvious: better methods for interpreting nondetect data. *Environ Sci Technol*. 2005 Oct 5;39(20):419A-423A.

From Helsel

Computing descriptive statistics

Current environmental guidance recommends three methods for computing descriptive statistics of data with nondetects: substituting one-half (or another fraction) of the RL (reporting level); the delta-lognormal method (D-LOG), which was originally known as Aitchison's method; and Cohen's method (*6–12*). However, all three methods are considered old technology that exhibit either bias or higher variability than other methods now available. Numerous studies have found that substituting one-half of the RL is inferior to other methods. Helsel and Cohn stated that the method "represents a significant loss in information" compared to other, better methods (*13*). Singh and Nocerino reported that it produced "a biased estimate of mean with the highest variability" (*14*), and Lubin et al. showed that it "results in substantial bias unless the proportion of missing data is small, 10 percent or less" (*15*). Resource Conservation and Recovery Act (RCRA) guidances recommend substitution only when data sets contain <15% nondetects, in which case the method is "satisfactory" (*8*, *12*). However, that judgment appears to be based only on opinion rather than on peer reviewed science. The U.S. EPA's 2004 Local Limits Development Guidance Appendices break from this pattern by not recommending substitution methods (*16*). Instead, this guidance recognizes that substitution results in a high bias when the mean or standard deviation is calculated and that performance worsens as the proportion of nondetects increases. Substitution introduces more problems today than 15 years ago, because most data today have multiple RLs. Several factors cause multiple RLs, including levels that change over time, samples with different dilutions, interferences from other constituents, different data interpretations for samples sent to multiple laboratories, or variations in RLs because methods for setting them have changed. Regardless of the cause, substituting a fraction of these changing limits for nondetects introduces a signal unrelated to the concentrations present in the samples. Instead, the signal represents the pattern of RLs used. In the end, false trends may be introduced—or real ones cancelled out.

Cohen's method assumes that data follow a normal distribution and is developed for a single censoring threshold or RL. Both assumptions are important limitations to how the method is applied today. Few modern data sets have only one RL, so data must be re-censored at the highest level before the tables can be used. For example, with RLs of 1 and 10 units, all detected observations between 1 and 10 (and all nondetects) must be designated as <10 units before the tables can be used. This assumption causes information to be lost, introducing error. Today, the lognormal distribution is considered more realistic than the normal distribution for most environmental data. Cohen's method is often computed with the logarithms of data, and estimates of mean and standard deviation of logarithms are transformed back into original units. This approach introduces a bias for data with <50 observations (*13*, *21*). Cohen's method is now totally unnecessary. Today, statistical software can easily handle multiple RLs and provide more accurate solutions to maximum likelihood equations.

Current methods

Modern MLE software, imputation (ROS for example), and the Kaplan–Meier method are three more accurate methods for computing statistics on data with nondetects. Each is now available in the survival analysis or reliability analysis sections of commercial statistics software. **MLE** solves a "likelihood equation" to find the values for mean and standard deviation that are most likely to have produced both nondetect and detected data. To begin, the user must choose a specific shape for the data distribution, such as the lognormal. Both detected observations and the proportion of data falling below each RL are used to fit the curve. MLE does not work well for data sets with <50 detected values, where 1 or 2 outliers may throw off the estimation, or situations in which insufficient evidence exists for one to know whether the assumed distribution fits the data well (*13*, *14*, *21*). **Imputation methods** fill in values for censored or missing observations without assigning them all the same value. The distribution of data, and perhaps other characteristics, must be specified. For example, regression on order statistics (ROS) is a simple imputation method that fills in nondetect data on the basis of a probability plot of detects (*13,21*). Multiple RLs can be incorporated. MLEs of mean and standard deviation can also be used to impute missing values (*22*). Because detected observations are used as measured, imputation methods depend less on assumptions of distributional shape than the MLE approach. As a result, imputation methods generally perform better than MLE with small sample sizes or when the data do not exactly fit the assumed distribution. For example, robust ROS estimates of mean and standard deviation performed better than MLE for sample sizes of <50 (*13*, *21*). EPA (*16*) and the state of Colorado (*23*) have incorporated ROS methods into recent environmental guidance documents. In medical and industrial statistics, Kaplan–Meier is the standard method for computing descriptive statistics of censored data (*2*, *3*). It is a nonparametric method designed to incorporate data with multiple censoring levels and does not require specification of an assumed distribution. It estimates the percentiles, or cumulative distribution function (CDF), for the data set. The mean equals the area beneath the CDF (*2*). Kaplan–Meier is also a counting procedure. A percentile is assigned to each detected observation, starting at the largest detected value and working down the data set, on the basis of the number of detects and nondetects above and below each observation. Percentiles are not assigned to nondetects, but nondetects affect the percentiles calculated for detected observations. The survival curve, a step function plot of the CDF, gives the shape of the data set. The Kaplan–Meier method has been used primarily for data with "greater thans", such as time until a disease recurs. For this method to be applied to "less thans", such as low-level chemical concentrations, data values must be individually subtracted from a large constant, or "flipped" (*4*), before the software is run. Flipping data is necessary only because of the way commercial software is now coded; it may become unnecessary with future versions as Kaplan–Meier becomes more widely used for analysis of "less-than" data. One caution is that estimates of the mean, but not percentiles, will be biased high with this method when the smallest value in the data set is a nondetect.

Testing hypotheses

Little guidance has been published for testing differences among groups of data with nondetects. The most frequently recommended method is the test of proportions, also called contingency tables (*7*, *8*). This test is most appropriate for data with only one RL, because all the data will be placed into one of two categories: below or above the RL. Thus, the approach tests for differences in the proportion of detected versus nondetected data. Information is lost on the relative ordering between detected values; this is captured and used by nonparametric tests such as the rank–sum test. Moreover, the use of the test of proportions on data with multiple RLs requires that values must be re-censored and reported as either below or above the highest RL. Compared with methods that handle multiple limits, this approach loses information. Nevertheless, the primary advantages of the test of proportions are its simplicity and its availability in familiar software. Unfortunately, the most commonly used test procedure is substituting one-half (or another fraction) of the RL before running standard tests such as the *t*-test. For data with one RL, Clarke demonstrated the significant errors produced by this procedure and by cited 15 years ago (*1*) and have not yet been adopted in environmental guidance documents. Now, however, much more detail is available (*4*). Parametric methods use MLE to perform tests equivalent to the *t*-test and analysis of variance (ANOVA) on data with multiple RLs. No substitution of fabricated values is required. Instead, likelihood-ratio tests determine whether splitting the data into groups explains a significant proportion of the overall variation. If so, the means differ among the groups. Millard and Deverel pioneered the use of nonparametric score tests for environmental data in 1986 (*25*). These tests, sometimes called the generalized Wilcoxon or Peto–Prentice tests, extend the familiar Wilcoxon rank–sum and Kruskal–Wallis tests to data with multiple RLs. No values are substituted, and no re-censoring is necessary. The tests are used to compare the CDFs among groups of data and to determine whether their percentiles differ. Even if lower percentiles are indistinguishable because they are all nondetects, differences in higher percentiles will be seen if they are significant.

…end of Helsel citation

Moreover, Helsel makes the demonstration than substituting values for nondetect produces poor estimates of statistics in: "Helsel, DR. Fabricating data: how substituting values for nondetects can ruin results and what can be done about it. *Chemosphere*.2005, 65, 2434-2439".

References (used in Helsel citation)

(1) Helsel, D. R. Less Than Obvious: Statistical Treatment of Data Below the Detection Limit. *Environ. Sci. Technol*. **1990**, *24*, 1766–1774.

(2) Klein, J. P.; Moeschberger, M. L. *Survival Analysis: Techniques for Censored and Truncated Data*; Springer: New York, 2003.

(3) Meeker, W. O.; Escobar, L. A. *Statistical Methods for Reliability Data*; Wiley: New York, 1998.

(4) Helsel, D. R. *Nondetects and Data Analysis: Statistics for Censored Environmental Data*; Wiley: New York, 2005.

(6) *Technical Support Document for Water Quality-Based Toxics Control*; EPA/505/2-90-001; U.S. EPA, Office of Water: Washington, DC, 1991.

(7) *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites*; EPA 540-R-01-003; U.S. EPA, Office of Emergency and Remedial Response: Washington, DC, 2002.

(8) *Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities: Addendum to Interim Final Guidance*; U.S. EPA, Office of Solid Waste: Washington, DC, 1992; http://www.epa.gov/epaoswer/hazwaste/ca/resource/guidance/sitechar/gwstats/gwstats.htm.

(9) *Guidance for Data Quality Assessment: Practical Methods for Data Analysis*; EPA/600/R-96/084; U.S. EPA, Office of Research and Development: Washington, DC, 1998.

(10) *Assigning Values to Non-Detected/Non-Quantified Pesticide Residues in Human Health Food Exposure Assessments*; Item 6047; U.S. EPA, Office of Pesticide Programs: Washington, DC, 2000; www.epa.gov/opp fod01/trac/science/trac3b012_nonoptimized.pdf.

(11) *Development Document for Proposed Effluent Limitations Guidelines and Standards for the Concentrated Aquatic Animal Production Industry Point Source Category*; EPA-821-R-02-016; U.S. EPA, Office of Water: Washington, DC, 2002.

(12) *RCRA Waste Sampling Draft Technical Guidance*; EPA-530- D-02-002; U.S. EPA, Office of Solid Waste: Washington, DC, 2002.

(13) Helsel, D. R.; Cohn, T. Estimation of Descriptive Statistics for Multiply Censored Water Quality Data. *Water Resour.Res*. **1988**, *24*, 1997–2004. (Note: In this paper, ROS is called MR.)

(14) Singh, A.; Nocerino, J. Robust Estimation of Mean and Variance Using Environmental Data Sets with Below Detection Limit Observations. *Chemom. Intell. Lab. Syst.* **2002**, *60*, 69–86.

(15) Lubin, J. H.; et al. Epidemiologic Evaluation of Measurement Data in the Presence of Detection Limits. Environ. Health Perspect. 2004, 112, 1691–1696.

(16) *Local Limits Development Guidance Appendices*; EPA 833- R-04-002B; U.S. EPA, Office of Wastewater Management: Washington, DC, 2004.

(21) Shumway, R. H.; et al. Statistical Approaches to Estimating Mean Water Quality Concentrations with Detection Limits. *Environ. Sci. Technol*. **2002**, *36*, 3345–3353.

(22) Kroll, C. N.; Stedinger, J. Estimation of Moments and Quantiles Using Censored Data. *Water Resour. Res.* **1996**, *32*, 1005–1012.

(23) *Determination of the Requirement to Include Water Quality Standards-Based Limits in CDPS Permits Based on Reasonable Potential: Procedural Guidance*; Colorado Water Quality Control Division: Denver, CO, 2003; www.cdphe.state.co.us/wq/PermitsUnit/wqcdpmt.html#RPGuide.

(25) Millard, S.; Deverel, S. Nonparametric Statistical Methods for Comparing Two Sites Based on Data with Multiple Nondetect Limits. *Water Resour. Res. 1988, 24, 2087–2098.*

Annex 2: Data processing of French datasets

1. **Monitoring region 1, seaweed, ^{137}Cs**
2. **Monitoring region 3, seaweed, ^{137}Cs**
3. **Monitoring region 2, seawater, ^{3}H**

Monitoring region 1, seaweed, ^{137}Cs

1. *Graphical presentation of data*

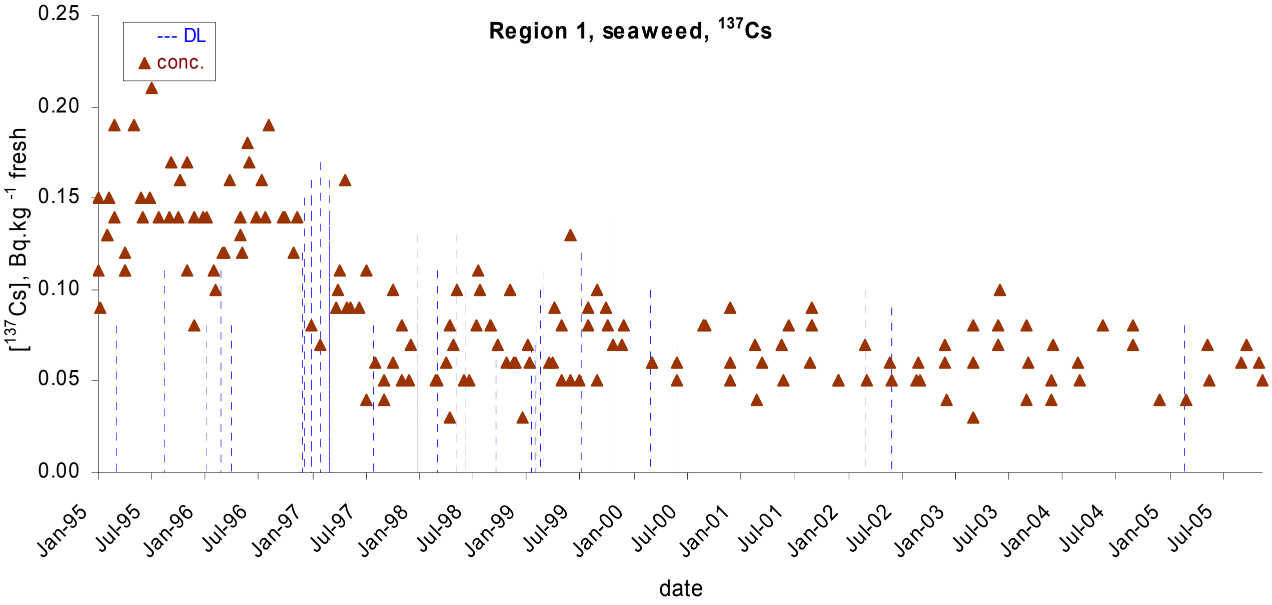


Figure A2.1. *Time series measurements of ^{137}Cs concentrations (Bq.kg^{-1} fresh) in seaweed from Monitoring region 1 (triangle symbols). Non-detects values are represented by a vertical dotted line between zero and the DL value which means that the actual value lies within this interval.*

2. *Statistical description of the dataset*

Individual data

Table A2.1*. Statistical parameters describing dataset 137Cs in seaweed from Monitoring region 1. (1) Total number of observations; (2) percentage of non-detects values; (3) number of different detection limit values, [min; max] lowest and highest DL values; (4) lowest and highest detected (>DL) values.*

Period	**Tot No.** (1)	**non-detects (%)** (2)	**No. DLs [min; max]** (3)	**Detects [min; max]** (4)	**median**	**mean**	**Standard deviation**
Baseline (1995 - 2001)	160	33 (20.63%)	12 [0.05;0.17]	[0.03 ; 0.23]	0.09	0.08	0.042
Assessment (2002 - 2005)	39	3 (7.69%)	3 [0.08;0.10]	[0.03 ; 0.10]	0.06	0.06	0.016

a. On a yearly basis

Table A2.2. *Statistical parameters describing dataset 137Cs in seaweed from Monitoring region 1 on a yearly basis. (1) Total number of observations; (2) percentage of non-detects values; (3) number of different detection limit values, [min; max] lowest and highest DL values; (4) lowest and highest detected (>DL) values.*

period	**Tot No.** (1)	**non-detects (%)** (2)	**No. DLs [min; max]** (3)	**Detects [min; max]** (4)	**median**	**annual mean**	**Ann. Std. Dev.**	**Period mean**	**Period Std. Dev.**
1995	28	2 (7.14%)	2 [0.08 ; 0.11]	[0.08;0.23]	0.140	0.140	0.037	**0.088**	**0.034**
1996	25	5 (20%)	4 [0.08 ; 0.15]	[0.10;0.19]	0.140	0.133	0.026		
1997	28	8 (28.57%)	7 [0.08 ; 0.17]	[0.04;0.16]	0.080	0.077	0.031		
1998	27	8 (29.63%)	6 [0.05 ;0.13]	[0.03;0.11]	0.060	0.067	0.025		
1999	28	8 (28.57%)	6 [0.07 ; 0;14]	[0.03;0.13]	0.07	0.070	0.024		
2000	12	2 (16.67%)	2 [0.07 ; 0.1]	[0.05;0.09]	0.060	0.067	0.015		
2001	12	0	0	[0.04;0.09]	0.060	0.063	0.015		
2002	12	2 (16.67%)	2 [0.09 ; 0.10]	[0.04;0.07]	0.050	0.056	0.011	**0.059**	**0.003**
2003	12	0	0	[0.03;0.10]	0.060	0.063	0.021		
2004	7	0	0	[0.04;0.08]	0.060	0.060	0.017		
2005	8	1(12.5%)	1 [0.08]	[0.04;0.07]	0.060	0.057	0.012		

3. *Comparison of the baseline and the assessment periods*

a. Starting from individual data (Table A2.1)

Comparison of the two periods with the non-parametric generalized Wilcoxon test gives:

Chisq= 17.8 on 1 degrees of freedom, p= 2.44E-05,

indicating that a significant difference exists between the two periods at the 5% threshold level. So it can be concluded that concentrations have decreased between the baseline and the assessment period.

b. With the two "period means" derived from annual means (Table A2.2)

i) Testing the hypotheses of normal distribution of data and homogeneity of variances.

Whether data are log-transformed or not, the test of Shapiro-Wilk rejects the hypothesis of normality and the test of Fisher rejects the hypothesis of homogeneity of variance. So the suitable test to compare the two "period means" is the non-parametric test of Wilcoxon-Mann-Whitney. However, for the purpose of consistency, the same statistical tests as those used in the report 1PE can be performed:

ii) Test of Welch-Aspin

t = 2.2339, df = 6.198, p-value = 0.0655,

=> There is no significant difference between the two means (5% level).

iii) Test of Wilcoxon-Mann-Whitney

W = 27, p-value = 0.01212

=> There is a significant difference between the two means (5% level), this gives evidence that radionuclide concentrations decreased between the baseline and the assessment period.

We are in the case where different conclusions can be drawn from the two mean comparison tests. As in 1PE, we conclude that there is some evidence for a statistical difference. However, it can be noted that, according to hypothesis on data distribution, the non-parametric test of Wilcoxon-Mann-Whitney is more suitable and the statistical significance of this difference (5% level) is definitely reliable.

Monitoring region 2, seawater, 3H

1. Graphical presentation of data

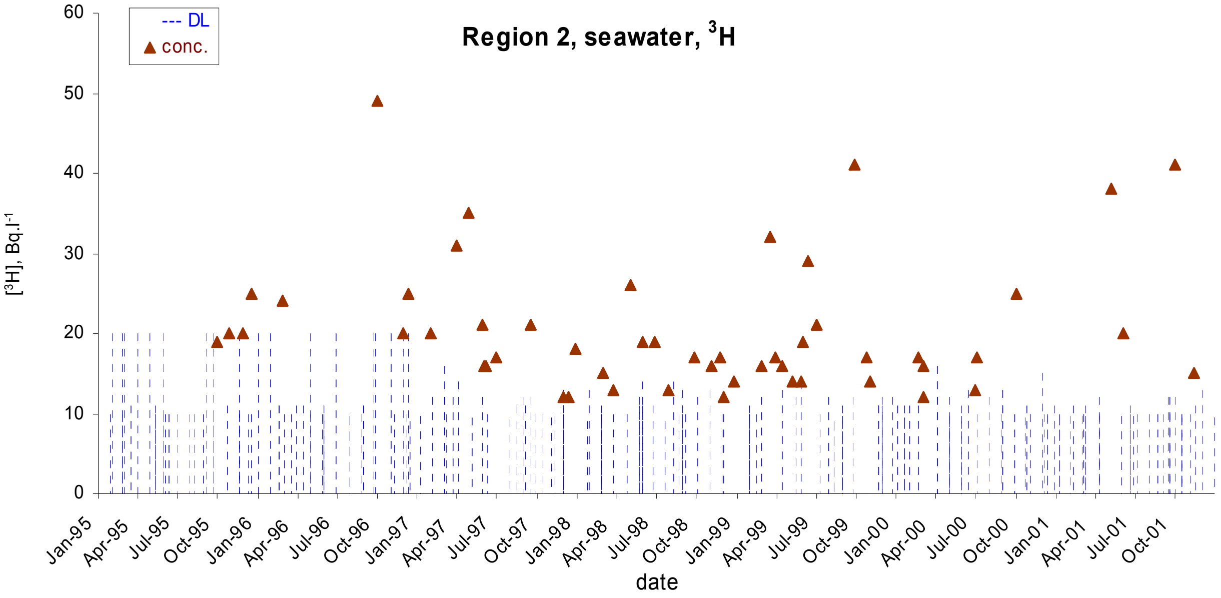


Figure A2.2. *Time series measurements of ^{3}H concentrations (Bq.L^{-1}) in seawater from Monitoring region 2 (triangle symbols). Non-detects values are represented by a vertical dotted line between zero and the DL value which means that the actual value lies within this interval.*

2. Statistical description of the dataset

a. Individual data

Table A2.3. *Statistical parameters describing dataset 3H in seawater from Monitoring region 2. (1) Total number of observations; (2) percentage of non-detects values; (3) number of different detection limit values, [min; max] lowest and highest DL values; (4) lowest and highest detected (>DL) values.*

Period	**Tot No.** (1)	**non-detects (%)** (2)	**No. DLs [min; max]** (3)	**Detects [min; max]** (4)	**median**	**mean**	**Standard deviation**
Baseline (1995 - 2001)	237	185 (78.06%)	18 [8.9;20.0]	[12.0 ; 49.0]	5.55	8.42	9.62
Assessment (2002 - 2005)	113	61 (53.98%)	11 [9.0;12.0]	[9.1 ; 56.0]	8.69	10.98	8.47

b. On a yearly basis

Table A2.4. *Statistical parameters describing dataset 3H in seawater from Monitoring region 2 on a yearly basis. (1) Total number of observations; (2) percentage of non-detects values; (3) number of different detection limit values, [min; max] lowest and highest DL values; (4) lowest and highest detected (>DL) values.*

period	**Tot No.** (1)	**non-detects (%)** (2)	**No. DLs [min; max]** (3)	**Detects [min; max]** (4)	**median**	**annual mean**	**Ann. Std. Dev.**	**Period mean**	**Period Std. Dev.**
1995	34	30 (88.24%)	5 [9.8;20.0]	[19.0;25.0]	*> 80% non-detects values: no computing*			**10.94** *(n_1=3)*	**0.51**
1996	34	30 (88.24%)	8 [9.0 ; 20.0]	[20.0;49.0]					
1997	34	23 (67.65%)	11 [9.2 ; 16.0]	[12.0;35.0]	7.62	10.69	7.72		
1998	34	23 (67.65%)	9 [9.1 ; 14.0]	[12.0;26.0]	8.91	10.61	4.81		
1999	33	21 (63.64%)	10 [8.9 ; 14.0]	[14.0;41.0]	7.67	11.52	8.87		
2000	34	28 (82.35%)	13 [9.2 ; 16.0]	[12.0;25.0]	*> 80% non-detects values: no computing*				
2001	34	30 (88.24%)	9 [9.4 ; 13.0]	[15.0;41.0]					
2002	29	17 (58.62%)	9 [9.0 ; 12.0]	[11.0;56.0]	6.00	11.27	11.65	**10.90**	**0.93**
2003	29	16 (55.17%)	5 [9.0;11.0]	[9.3;29.0]	7.39	9.51	6.04		
2004	27	15 (55.56%)	5 [9.0;11.0]	[9.6;45.0]	5.87	11.51	12.29		
2005	28	13 (46.43%)	6 [9.0 ; 11.0]	[9.1;20.0]	9.2	11.30	3.48		

It should be pointed out that statistical parameters are not estimated for years 1995, 1996, 2000 and 2001 because more than 80% of data are below detection limits.

3. *Comparison of the reference and the assessment periods*

a. Starting from individual data (Table A2.3)

Comparison of the two periods with the non-parametric generalized Wilcoxon test gives:

Chisq= 6.7 on 1 degrees of freedom, p= 0.00988,

indicating that a significant difference exists between the two periods at the 5% threshold level. So it yields to the conclusion that concentrations have increased between the reference period and the assessment period.

b. With the two "period means" derived from annual means (Table A2.4)

i) Testing the hypotheses of normal distribution of data and homogeneity of variances. The test of Shapiro-Wilk accepts the hypothesis of normality and the test of Fisher accepts the hypothesis of homogeneity of variance. So the suitable test to compare the two "period means" is the parametric Student t test. It must be outlined that the sample sizes are "indecently" small (n_1=3 and n_2=4).

ii) Student t Test
t = 0.0651, df = 5, p-value = 0.9507
=> There is no significant difference between the two means (5% level).
However, for the purpose of consistency, the same statistical tests as those used in the report 1PE can be performed:

iii) Test of Welch-Aspin
t = 0.0713, df = 4.733, p-value = 0.9460,
=> There is no significant difference between the two means (5% level).

iv) Test of Wilcoxon-Mann-Whitney
W = 6, p-value ≈ 1
=> There is no significant difference between the two means (5% level).

4. *Discussion*

There is a discrepancy between the conclusions drawn from the comparison of the two periods either using individual data or on a yearly basis. Several comments should be made about the processing of this dataset:

o The graphical presentation of the dataset does not display any clear pattern suggesting an increase or a decrease with time (Figure 7.1).

o Individual data corresponding to the baseline [1995 - 2001] include 78% of values below the detection limits, which is very close to the threshold of 80%, arbitrarily set by statisticians and where inferred conclusions are tenuous.

o As mentioned above, the comparison of means assumes that the hypotheses of homogeneity of the variances and normal distribution of data are first checked (test of Fisher and Shapiro-Wilk) in order to select the suitable test. But with so small samples (n_1 = 3 for the baseline and n_2 = 4 for the assessment period), checking those hypotheses may be not considered as reliable. With no hypothesis on data distribution, the non-parametric rank test of Wilcoxon Mann Whitney is unlikely to give evidence for any significant difference because of the very small size of samples and reliability of the conclusion would be tenuous, anyway.

It is most likely that the processing of this particular dataset illustrates the limitations of the statistical methods which consist in comparing the two periods when too many values are below the detection limits.

Monitoring region 3, seaweed, 137Cs

1. *Graphical presentation of data*

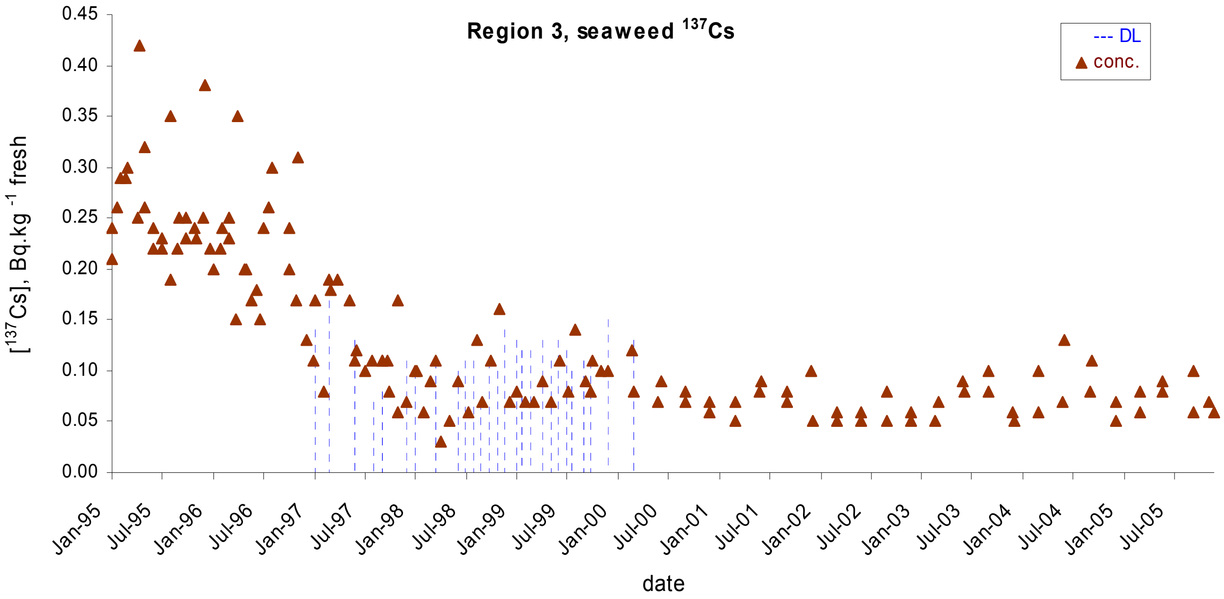


Figure A2.3. *Time series measurements of ^{137}Cs concentrations (Bq.kg^{-1} fresh) in seaweed from Monitoring region 3 (triangle symbols). Non-detects values are represented by a vertical dotted line between zero and the DL value which means that the actual value lies within this interval.*

2. *Statistical description of the dataset*

a. Individual data

Table A2.5. *Statistical parameters describing dataset 137Cs in seaweed from Monitoring region 3. (1) Total number of observations; (2) percentage of non-detects values; (3) number of different detection limit values, [min; max] lowest and highest DL values; (4) lowest and highest detected (>DL) values.*

Period	**Tot No.** (1)	**non-detects (%)** (2)	**No. DLs [min; max]** (3)	**Detects [min; max]** (4)	**median**	**mean**	**Standard deviation**
Baseline (1995 - 2001)	133	27 (20.30%)	9 [0.07;0.17]	[0.03 ; 0.42]	0.11	0.14	0.086
Assessment (2002 - 2005)	32	0 (0.00%)	0	[0.05 ; 0.13]	0.07	0.07	0.020

b. On a yearly basis

Table A2.6. *Statistical parameters describing dataset of 137Cs in seaweed from Monitoring region 3 on a yearly basis. (1) Total number of observations; (2) percentage of non-detects values; (3) number of different detection limit values, [min; max] lowest and highest DL values; (4) lowest and highest detected (>DL) values.*

Period	**Tot No.** (1)	**non-detects (%)** (2)	**No. DLs [min; max]** (3)	**Detects [min; max]** (4)	**median**	**annual mean**	**Ann. Std. Dev.**	**Period mean**	**Period Std. Dev.**
1995	24	0 (0%)	0	[0.19;0.42]	0.2	0.26	0.056	**0.13**	**0.077**
1996	22	1 (4.55%)	1 [0.14]	[0.13;0.35]	0.20	0.22	0.058		
1997	24	6 (25%)	6 [0.07;0.17]	[0.06;0.19]	0.11	0.11	0.045		
1998	23	9 (39.13%)	5 [0.07;0.14]	[0.03;0.16]	0.07	0.08	0.037		
1999	24	11 (45.83%)	5 [0.10;0.15]	[0.03;0.13]	0.08	0.09	0.022		
2000	8	0 (0%)	0	[0.06;0.12]	0.07	0.08	0.019		
2001	8	0 (0%)	0	[0.05;0.10]	0.07	0.07	0.018		
2002	8	0 (0%)	0	[0.05;0.08]	0.05	0.06	0.01	**0.07**	**0.011**
2003	8	0 (0%)	0	[0.05;0.10]	0.07	0.07	0.018		
2004	8	0 (0%)	0	[0.05;0.13]	0.07	0.08	0.027		
2005	8	0 (0%)	0	[0.06;0.10]	0.07	0.08	0.015		

3. *Comparison of the reference and the assessment periods*

a. Starting from individual data (Table A2.5)

Comparison of the two periods with the non-parametric generalized Wilcoxon test gives:

Chisq= 20.8 on 1 degrees of freedom, p= 5.23E-06,

indicating that a significant difference exists between the two periods at the 5% threshold level. So it can be concluded that concentrations have decreased between the reference period and the assessment period.

b. With the two "period means" derived from annual means (Table A2.6)

i) Testing the hypotheses of normal distribution of data and homogeneity of variances.

Whether data are log-transformed or not, the test of Shapiro-Wilk rejects the hypothesis of normality. The test of Fisher on log-transformed data accepts the hypothesis of homogeneity of variance but rejects the hypothesis on non-transformed data. So the suitable test to compare the two "period means" is the non-parametric test of Wilcoxon-Mann-Whitney. However, for the purpose of consistency, the same statistical tests as those used in the report 1PE can be performed:

ii) Test of Welch-Aspin
t = 1.9684, df = 6.414, p-value = 0.09347,
=> There is no significant difference between the two means (5% level).

iii) Test of Wilcoxon-Mann-Whitney

W = 24, p-value = 0.07273

=> The difference between the two means is not significant (5% level).

There is a discrepancy between the conclusions drawn from the comparison of the two periods either using individual data or on a yearly basis. This illustrates the lack of power of the statistical analysis performed on small samples when dealing with annual means rather than individual data. Looking at the graph suggests that ^{137}Cs level in seaweed from Monitoring region 3 decreases with time, especially during the baseline [1995 - 2001] (Figure A2.2). Aggregating individual data as annual means reduces the size of samples (n_1 = 7 for the baseline and n_2 = 4 for the assessment period) and probably yields to a type 2 error (suggesting that the difference is not significant though it actually exists).

Annex 3: Brief description of methods to estimate datasets statistical parameters

1. **Kaplan-Meier method**
2. **Robust ROS (regression on order statistics) method**
3. **Maximum likelihood estimation**
4. **The generalized Wilcoxon test**

1. *Kaplan-Meier method*

Principle

Kaplan-Meier (K-M) is the standard nonparametric method for estimating summary statistics of multiply censored data. It has seen widespread usage in the fields of medical sciences and systems-engineering where it is employed within a more general framework termed ‘‘survival analysis’’ or ‘‘reliability analysis’’; in these contexts data are right-censored data (expressed as "greater than" values). Nevertheless, Kaplan-Meier method could also be employed for estimating statistics when data are left-censored (expressed as "less than" values) as there are in the environmental sciences. Kaplan-Meier method consists in computing a survival probability function S for "greater than" values, usually defined as:

S(x) = Prob(X > x) ***(Equation 1.1)***

In order to use Kaplan-Meier with "less than" values, there are firstly to be flipped into "greater than" form, using the following equations:

Flipped Data = Constant – Original Data ***(Equation 1.2)***

or

$Flipp_i = M - c_i$ ***(Equation 1.3)***

Using flipped data and equation 3 the following result is obtained:

S(x) = Prob (Flipp > x) = Prob(M-c > x) ***(Equation 1.4)***

Equation 5 shows that survival probabilities of the flipped data are also cumulative distribution function of the original x data, usually defined by equation 6.

Prob(M-c > x) = Prob(c < M-x) ***(Equation 1.5)***

F (x) = Prob (X ≤ x) ***(Equation 1.6)***

Thus, with "less than" values Kaplan-Meier method produces empirical cumulative distributions (ECDFs) which are discrete-interval step functions.

Regarding the flipped data, K-M method computes the survival probabilities S for each detected value. "Using the flipped values, the detected observations ("failures", or "deaths" in survival analysis terminology) are ranked from small to large, accounting for the number of censored data in between each detected observations. […] K-M places each nondetect at its detection limit prior to ranking. The "number at risk" b equals the number of observations, both detected and censored, at and below each detected concentration. The number of detected observations at that concentration is d, where d is greater than 1 for tied values. The incremental survival probability is the probability of "surviving" to the next lowest detected concentration, given the number of data at and below that concentration or (b-d)/b. The survival function probability is the product of the j = 1 to k incremental probabilities to that point, going from high to low concentration for the k detected observation" (Helsel D, 2005).

$$S = \prod_{j=1}^{k} \frac{b_j - d_j}{b_j}$$ ***(Equation 1.8)***

For the case of ties, K-M assigns the smallest rank possible to each observation, rather than the average rank as is done for most nonparametric test. K-M will assign a probability of 0 to the smallest observation (largest flipped value), if there are no nondetects below this value in the data set. This represents a plotting position of i/n for the empirical distribution function of flipped values, so that the probability of exceeding the last value is 0. If the smallest concentration is a censored value, as is usually the case, the smallest detected observation will have a nonzero exceedance probability, while probabilities are indeterminate for all nondetects below the lowest detected observation"(Helsel D, 2005).

Detailed Example:

The dataset used is the Oahu one from the NADA add-on package (Lee and Helsel, 2005;2007).

Table A3.1. *The Oahu dataset*

As	***AsCen****
1.0	TRUE
1.0	TRUE
1.7	FALSE
1.0	TRUE
1.0	TRUE
2.0	TRUE
3.2	FALSE
2.0	TRUE
2.0	TRUE
2.8	FALSE
2.0	TRUE
2.0	TRUE
2.0	TRUE
2.0	TRUE
2.0	TRUE
0.7	FALSE
0.9	FALSE
0.5	FALSE
0.5	FALSE
0.9	TRUE
0.5	FALSE
0.7	FALSE
0.6	FALSE
1.5	FALSE

*AsCen is the censoring variable, when it equals TRUE that means that the data is a nondetect one.

Table A3.2. *Computation of Kaplan-Meier survival probabilities for the Oahu dataset (n=24).*

As (original data)	***AsCen***	***FlipAs (Flipped data = 5 –original data)***	***rank r***	***Number at risk b=(n-r+1)***	***Event (d)***	***incremental survival probabilities p=(b-d)/b***	***Survival probabilities (flipped data) = Cumulative probabilities (original data)***
3.2	FALSE	1.8	1	24	1	0.958	0.958
2.8	FALSE	2.2	2	23	1	0.957	0.917
1.7	FALSE	3.3	11	14	1	0.929	0.851
1.5	FALSE	3.5	12	13	1	0.923	0.786
0.9	FALSE	4.1	17	8	1	0.875	0.688
0.7	FALSE	4.3	19	6	2	0.667	0.458
0.6	FALSE	4.4	21	4	1	0.750	0.344
0.5	FALSE	4.5	22	3	3	0.000	0.000

For example, the survival function probability of 0.688 for the concentration at 0.9 equals 0.786 * (7/8).

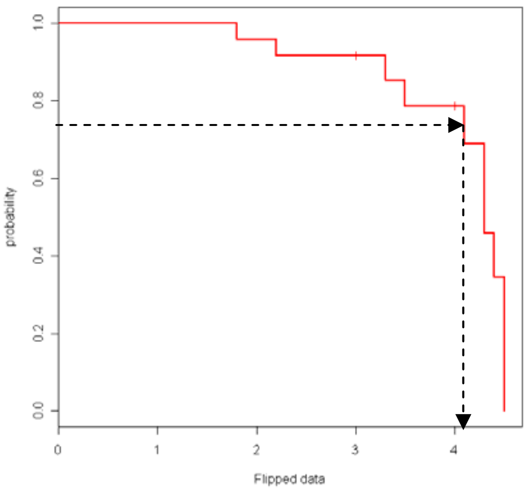


Figure A3.1. *Survival probability function S of the multiply-censored flipped Oahu data.*

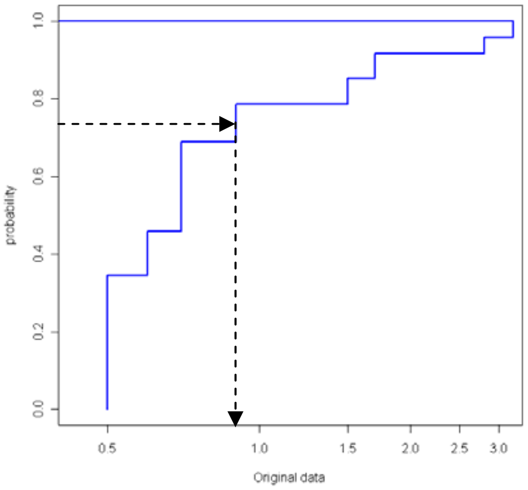


Figure A3.2. *Empirical cumulative distribution function of the multiply-censored Oahu original data.*

<u>Estimation of the summary statistics</u>

"For **percentiles**, the estimate is the minimum X value on the survival function graph that is intersected by the line drawn at probability value from the Y-axis. It is the smallest flipped observation having a survival probability equal to or less than the stated probability of the percentile. The 25th (Q1) has a survival probability of exceedance) of 0.75. A horizontal line drawn from 0.75 on the Y-axis intersects the vertical line at an X-value of 4.1 (cf. Figure A3.1). Looking at Table A3.2, the flipped observation at 4.1 is the smallest flipped value for which the survival probability is 0.75 or less. Subtracting this from the flipping constant of 5, the 75th percentile of the original data is 0.9. The process is similar for others percentiles"(Helsel D, 2005).

"The ***mean*** is computed by integrating the area under the K-M survival curve. To see why this is so, consider the usual equation for the mean of n observations

$$\mu = \square \frac{x}{n} \qquad \textit{(Equation 1.9)}$$

Where there are severaı observations at the same value, the equation can be stated as

$$\mu = \square \frac{f_i}{n} x_i \qquad \textit{(Equation 1.10)}$$

Where f_i is the number of observations at each of the i unique values of x and f_i/n is the proportion of the data set at that value. The mean is the sum of the products of the proportion of data for each value times the magnitude of the observation's value. This is just what is accomplished when integrating under the K-M survival curve. The curve is divided by drawing horizontal lines at the value of each detected observation. The resulting set of rectangles has as their height the estimated proportion of data at that value, with the proportions summing to 1. The width of rectangle is the magnitude of the observation, x. The mean is estimated by multiplying the width of each rectangle by it's height to get the area, and then summing over all rectangles"(Helsel D, 2005).

"Location estimates for flipped data (mean, median, other percentiles) must be re transformed back into the original scale by subtraction from the constant M used to flip the data". (Helsel D, 2005).

The estimate for the **standard error** of the survival function (S) is known as Greewood's formula:

$$\text{Std Error of S} = \text{s.e [S]} = \text{S} * \sqrt{\sum_{j=1}^{k} \frac{d_j}{b_j(b_j - d_j)}}$$

(Equation 1.11)

The standard deviation (sd) could be estimated multiplying standard error of the mean by the square root of the sample size n.

$$\text{sd} = \text{std.error} * \sqrt{(n)}$$

(Equation 1.12)

Estimates of variability (variance, standard deviation, standard error, IQR) are the same for both flipped and original units; no retransformation is needed.

Table A3.3. *Summary statistics using kaplan-meier for the multiply censored Oahu data*

Mean	***sd***	***Q1***	***median***	***Q3***
0.949	0.807	0.5	0.7	0.9

Remarks

When more than 50% of data are censored, and the smallest observation (largest flipped value) is censored, the median cannot be estimated using K-M. A method which assumes some sort of model for the data distribution must be employed if an estimate for the median is required.

2. Robust ROS (regression on order statistics) method

The robust ROS is a semiparametric method developed by Helsel and Cohn (1988).

"It is a probability plotting and regression procedure that models censored distributions using a linear regression model of observed concentrations vs. their normal quantiles (or "order statistics"). The method has been evaluated as one of the most reliable procedures for developing summary statistics of multiplycensored data (Shumway et al., 2002)" *from Lee and Helsel, 2007].*

The robust ROS method can be summarize in four steps. They will be described using the dataset Oahu of the NADA add-on package (Lee and Helsel, 2005; 2007).

Step 1: Computation of plotting position for both censored and uncensored data

"Plotting positions of both censored and uncensored data are computed using the exceedance probability, Ej, of each censoring limit. Ej is the probability of exceeding the jth censoring limit. It is defined as

$$\mathbf{E_j = E_{j+1} + (A_j / [A_j + B_j]) (1 - E_{j+1})}$$

(Equation 2.1)

where A_j is the total number of uncensored observations in the range[j, j+1) and B_j is the total number of observations, censored and uncensored, less than or equal to the j^{th} censoring limit. For a given uncensored observation, a Weibull-type plotting position p can be calculated by considering the exceedance probability of the censoring limit below the observation E_j , the exceedance probability of the censoring limit above the observation E_{j+1}, and the observation's rank among all the values within the j and j + 1 censoring limit. In general, the Weibull-type plotting positions for uncensored observations are

$$\mathbf{p(i) = (1 - E_j) + (Ej - E_{j+1})r_i / (Aj + 1),}$$

(Equation 2.2)

where r_i is the rank of the i^{th} observation among the observations in the range (j, j+1] (Hirsch and Stedinger, 1987).

Similarly, the Weibull-type plotting positions for censored observations are given by

$p(i) = (1 – E_j) r_i / (C_j +1)$ ***(Equation 2.3)***

where C_j is the total number of censored values in the range (j; j + 1]" (Lee L and Helsel D , 2005).

" When j = the highest limit, $E_{j+1} = 0$ and A_j and $B_j = n$. The numbers of nondetects below the j^{th} detection limit is defined as $C_j = B_j – Bj-1 – A_{j-1}$" (Helsel D, 2005).

<u>*Step 2: Forming the linear regression model*</u>

"A linear regression of the uncensored observations vs. the normal quantiles of the uncensored plotting positions is formed. The normal quantiles of the plotting positions are the ‘‘order statistics’’ of the ROS method"(Lee and Helsel, 2005).

"The intercept, the y value associated with a normal score of 0 at the centre of the plot, estimates the mean of the distribution. The slope of the line equals the standard deviation, as normal scores are scaled to units of standard deviation"(Helsel D, 2005).

<u>*Step 3: Estimation of the censored concentrations*</u>

"The censored concentrations are modelled using the parameters of the linear regression and normal quantiles of the censored data. These modelled censored observations are only used corporately, along with the uncensored observations, to model the distribution of the sample population. Individually, they are not considered the values that would have existed in the absence of censoring"(Lee and Helsel, 2005).

<u>*Step 4: Computation of summary statistics:*</u>

"The observed uncensored values are combined with modelled censored values to corporately estimate summary statistics of the entire population. By combining the uncensored values with modelled censored values, this method avoids transformation bias (Helsel and Cohn, 1988)" (Lee and Helsel, 2005).

<u>*Remark*</u>

"The ROS method assumes that all censoring thresholds are ‘‘left censored’’, i.e., all censored values are ‘‘less thans’’. It is applicable to any dataset containing 0 to80% of its values censored. As noted by Helsel and Cohn (1988) and Helsel (2005), statistics derived from ROS models of populations having 80% or more censored values are very tenuous. For data whose highest detection limit is below the 50th percentile, the median will equal the sample median computed by standard software without special consideration for censored values. The primary advantages of using ROS are realized when 50% to 80% of data are below the highest detection limit, or when estimates of the mean and standard deviation are required. Unlike the median, the mean and standard deviation cannot be estimated without some accommodation for censoring.

Additional assumptions are those inherent to linear regression. This includes the assumptions that the response variable (concentration) is a linear function of the explanatory variable (the normal quantiles) and that the error variance of the model is constant. Since the statistical distribution of water-quality data is typically skewed, these assumptions are usually addressed by transforming the data prior to analysis. Since most water-quality data with multiple censoring limits are lognormally distributed, the default behaviour of our routines is to perform a log-normal transformation to input data prior to computation. However, this feature can be entirely suppressed or the user may provide an alternative set of transformation functions" (Lee and Helsel, 2005).

Detailed example:

Table A3.4. *Oahu dataset*

As	AsCen*
1.0	TRUE
1.0	TRUE
1.7	FALSE
1.0	TRUE
1.0	TRUE
2.0	TRUE
3.2	FALSE
2.0	TRUE
2.0	TRUE
2.8	FALSE
2.0	TRUE
2.0	TRUE
2.0	TRUE
2.0	TRUE
2.0	TRUE
0.7	FALSE
0.9	FALSE
0.5	FALSE
0.5	FALSE
0.9	TRUE
0.5	FALSE
0.7	FALSE
0.6	FALSE
1.5	FALSE

* When AsCen variable =TRUE that means that the data is a nondetected one.

Step 1: Computation of plotting position for both censored and uncensored data

First data have to be ranked from highest value to lowest value. When a nondetect and a detect data have the same value, the detect one has to be placed before the nondetect. As there are 3 censoring limits in the Oahu dataset, j varies from 1 to 3, from the lowest detection limit to the highest one. The exceedance probability of j=3 has to be computed first, then E(j=2) and E(j=1). E(3) = 0 + (2 /[2 +22]) (1 – 0) = 0.083, then E(2) = 0.083+ (2 /[2 +12) (1 – 0.083) = 0.214, and E(1) = 0.214+ (1 /[1 +7) (1 – 0.214) = 0.313.

Then the plotting position for the detects observations can be computed. The plotting position corresponding to the 3.2 detect value is estimated by

(1-0.083) + (0.083 -0)*2 /(2+1) = 0.972; the plotting position corresponding to the 2.8 detect value is estimated by (1-0.083)+(0.083 -0)*1 /(2+1) = 0.944.

In the same way, the others plotting position are estimated. For the detects data which are under the lowest detection limit Ej=1. Thus the plotting position for the 0.7 value corresponding to r_i=6 is : (1-1)+(1-0.313)*6/(6+1) = 0.589.

For the censored observation computation of plotting position involves no difficulties.

***Table A3.2.** Computation of plotting position for both detects and nondetects data*

j	r_i	As	AsCen	A_j	B_j	C_j	E_j	p(i) detected	p(i) censored
	2	3.2	FALSE					0.972	
	1	2.8	FALSE					0.944	
3	*8*	2.0	TRUE	2	22	8	0.083		0.815
	7	2.0	TRUE						0.713
	6	2.0	TRUE						0.611
	5	2.0	TRUE						0.509
	4	2.0	TRUE						0.407
	3	2.0	TRUE						0.306
	2	2.0	TRUE						0.204
	1	2.0	TRUE						0.102
	2	1.7	FALSE					0.873	
	1	1.5	FALSE					0.829	
2	*4*	1.0	TRUE	2	12	4	0.214		0.629
	3	1.0	TRUE						0.471
	2	1.0	TRUE						0.314
	1	1.0	TRUE						0.157
		0.9	FALSE					0.737	
1	*1*	0.9	TRUE	1	7	1	0.313		0.344
	6	0.7	FALSE					0.589	
	5	0.7	FALSE					0.491	
	4	0.6	FALSE					0.393	
	3	0.5	FALSE					0.295	
	2	0.5	FALSE					0.196	
	1	0.5	FALSE					0.098	

Step 2: Forming the linear regression model

First, the standard normal quantiles of the detect plotting positions have to be computed from a table of the standard normal distribution.

Table A3.3. *Computation of the normal quantiles of the detect plotting positions*

As	ln As	detect plotting positions	Standard normal quantiles of the detect plotting positions
3.2	1.163	0.972	1.915
2.8	1.030	0.944	1.593
1.7	0.531	0.873	1.141
1.5	0.405	0.829	0.952
0.9	-0.105	0.737	0.633
0.7	-0.357	0.589	0.226
0.7	-0.357	0.491	-0.022
0.6	-0.511	0.393	-0.272
0.5	-0.693	0.295	-0.540
0.5	-0.693	0.196	-0.854
0.5	-0.693	0.098	-1.292

Then a linear regression of logarithms of detect observations (because log normal distribution is usually assumed with environmental data) vs. the normal quantiles of the detect plotting position is computed in order to estimate the mean and the standard of the distribution.

In the example, mean equals - 0.23, and the standard deviation equals 0.6468.

Step 3: Estimation of the censored concentrations

First standard normal quantiles of the nondetects have to be computed from a table of the standard normal distribution. Then, using the mean and the standard deviation estimated in step 2, and the plotting position of the nondetects data, log values for individual nondetect data are predicted.

Table A3.4. *Estimation if the nondetect concentrations*

As	Non-detect plotting positions	Standard normal quantiles of the nondetect plotting positions	predicted log values = - 0.24 + 0.65 normal quantiles
2.0	0.815	0.896	0.351
2.0	0.713	0.562	0.135
2.0	0.611	0.282	-0.047
2.0	0.509	0.023	-0.215
2.0	0.407	-0.234	-0.382
2.0	0.306	-0.508	-0.560
2.0	0.204	-0.828	-0.768
2.0	0.102	-1.271	-1.055
1.0	0.629	0.328	-0.017
1.0	0.471	-0.072	-0.277
1.0	0.314	-0.484	-0.544
1.0	0.157	-1.006	-0.883
0.9	0.344	-0.402	-0.491

Step 4: Computation of summary statistics:

Using the detect values and the retransforming individual predicted log values summary statistics are computed

Table A3.5: *Computation of summary statistics*

As	AsCen	Predicted log values	Retransforming predicted log values (exp)	Data use to compute summary statistics
3.2	FALSE			3.20
2.8	FALSE			2.80
2.0	TRUE	0.35	1.42	1.42
2.0	TRUE	0.13	1.14	1.14
2.0	TRUE	-0.05	0.95	0.95
2.0	TRUE	-0.21	0.81	0.81
2.0	TRUE	-0.38	0.68	0.68
2.0	TRUE	-0.56	0.57	0.57
2.0	TRUE	-0.77	0.46	0.46
2.0	TRUE	-1.05	0.35	0.35
1.7	FALSE			1.70
1.5	FALSE			1.50
1.0	TRUE	-0.02	0.98	0.98
1.0	TRUE	-0.28	0.76	0.76
1.0	TRUE	-0.54	0.58	0.58
1.0	TRUE	-0.88	0.41	0.41
0.9	FALSE			0.90
0.9	TRUE	-0.49	0.61	0.61
0.7	FALSE			0.70
0.7	FALSE			0.70
0.6	FALSE			0.60
0.5	FALSE			0.50
0.5	FALSE			0.50
0.5	FALSE			0.50
			Mean	**0.97**
			sd	**0.72**

Hirsch, R., Stedinger, J., 1987. Plotting positions for historical floods and their precision. *Water Resources Research* 23 (4), 715–727.

Helsel, D.R., Cohn, T.A., 1988. Estimation of descriptive statistics for multiply-censored water quality data. *Water Resources Research* 24 (12), 1997–2004.

3. Maximum likelihood estimation

When nondetect data are present, in the most general case, L (the likelihood function) can be considered to be the product of three pieces, where the censored data component is split into two, one for left-censored and one for right-censored data :

$$L = \prod p[\mathrm{x}] \prod (F[\mathrm{x}]) \prod S[\mathrm{x}] \qquad \textit{(Equation 3.1)}$$

where p[x] is the pdf (probability density function) as estimated from detected observations, (F[x]) is the cdf (cumulative density distribution) as determined by left-censored observation, and S[x] is the survival function as determined by right-censored observations ("greater thans"). Greater-thans are not typically found among environmental data, and so likelihood function is environmental studies typically deal with only the first two pieces.

For censored data, two variables x and δ are required to represent each observation. The value for measurement, or for the detection limit, is given by x. The indicator variable δ is a 0/1 variable that designates whether an observation is censored (0) or detected(1). As one of the simpler likelihood functions, the equation for L when estimating the mean and standard deviation of a normal distribution using MLE is:

$$L = \prod p[\mathrm{xi}]^{\delta_i} \bullet F[\mathrm{x}_i]^{1-\delta_i} \qquad \textit{(Equation 3.2)}$$

where δ is as defined above, and for the a normal distribution pdf is

$$p[\mathrm{x}] = \frac{\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)}{\sigma\sqrt{(2\pi)}} \qquad \textit{(Equation 3.3)}$$

For detected observation δ = 1 and the second term in equation 2 becomes 1 and so drops out. For censored observations, δ = 0 and the first term becomes 1 and so drops out. The cumulative distribution function for a normal distribution is

$$F[\mathrm{x}] = \Phi\left(\frac{x-\mu}{\sigma}\right) \qquad \textit{(Equation 3.4)}$$

where Φ is the cdf of the standard normal distribution

$$\Phi[\mathrm{y}] = \frac{1}{\sqrt{(2\pi)}} \int_0^y \exp(-u^2/2)\, du \qquad \textit{(Equation 3.5)}$$

After substituting in the above and setting the partial derivatives on ln(L) to 0 […], the nonlinear equations are solved by iterative approximation using the Newton-Raphson method. The solution provides the parameters mean and standard deviation for the distribution that best matches both the pdf and cumulative distribution function (or 1-survival function) estimated from the data. In other words, the estimates of mean and standard deviation will be the parameters for the assumed distributional shape that had the highest likelihood of producing the observed values for the detected observations and the observed proportion of data below each of detection limits" (Helsel D, 2005).

"Environmental data are more often similar to a lognormal than a normal distribution, so the mean and variance of the logarithms are more typically estimated by MLE, whether with the table adjustment or by direct solution, and subsequently reconverted to estimates in original units. The traditional formulae for reconversion are:

$$\hat{\mu} = \exp(\hat{\mu}_{\ln} + \hat{\sigma}^2_{\ln}/2) \quad \textit{(Equation 3.6)}$$

$$\hat{\sigma} = \hat{\mu}^2 \bullet [\exp(\hat{\sigma}^2_{\ln}) - 1] \quad \textit{(Equation 3.7)}$$

$$C.V = [\exp(\hat{\sigma}^2_{\ln}) - 1]^{1/2} \quad \textit{(Equation 3.8)}$$

Where $\hat{\mu}_{\ln}$ and $\hat{\sigma}^2_{\ln}$ are estimates of the mean and variance, respectively, of the natural logarithms of the data. These equations will work reasonably well if the data are close to their values. However, for small samples the estimates are typically poor enough to bias estimates in original units (Cohn, 1988) leading to overestimation of the mean and variance.

Estimates for percentiles are obtained by computing the percentiles in log units, assuming that the logarithms follow a normal distribution, and then retransforming. The kth percentile value is therefore computed as:

$$p_k = \exp(\mu_{\ln} + z_k\sigma_{ln})$$

where p_k is the kth percentile value in original units, and z_k is the kth percentile of a standard normal distribution. For the median, k=0.5 and z_k=0, so that $p_{0.5}$ = exp($\mu_{\ln}$). The exponentiated mean of tht logarithms is sometimes given a special name, the geometric mean. When the logarithms of data follow a normal distribution, the geometric mean estimates the median of the data's original units (and not the mean)" (Helsel D, 2005).

Remarks

"The most crucial consideration for MLE is how well data fit the assumed distribution. A major problem with MLE is that for small data sets there is often insufficient information to determine whether the assumed distribution is correct or not, or to estimate parameters reliability. MLE has been shown to perform poorly for data sets with less than 25 to 50 observations (Gleit, 1985; Shumway et al., 2002). For larger data sets, MLE is an efficient way to estimate parameters, given that the chosen distribution is correct. The term "efficient" means that the fitted parameters have relatively small variability, so that their confidence limits are as small as possible. For data sets of at least 50 observations, and where either the percent censoring is small (so that the distributional shape can be evaluated) or the distribution can be assumed from knowledge outside the data set, MLE methods are the method of choice" (Helsel D, 2005).

Cohn, T.A., 1988, Adjusted maximum likelihood estimation of the moments of lognormal populations from type I censored samples: U.S Geological Survey Open-File Report 88-350,34pp.

Gleit, A., 1985, Estimation for small normal data sets with detection limits. *Environmental Science and technology* 19, 1201-1206.

Shumway, R.H., Azari, R.S., Kayhanian, M., 2002. Statistical approaches to estimating mean water quality concentrations with detection limits. *Environmental Science and Technology* 36 (15), 3345–3353.

4. The generalized Wilcoxon test

The generalized Wilcoxon test is a nonparametric test which permits to compare two groups having multiple detection limits.

"Peto and Peto (1972) proposed a modification to the Gehan test called the "generalised Wilcoxon test". Prentice (1978) and Prentice and Marek (1979) elaborated on its properties, so the test is also called the Peto-prentice test"(Helsel D, 2005).

The null and alternative hypothesis of the generalized Wilcoxon test are:

H_0 : distributions of data in the two groups are identical (their ECDFs are the same);

H_1 : distributions of data in the two groups are different (their ECDFs are differents).

"Scores for the generalised Wilcoxon test are a weighted version of the Gehan test, adjusting the U scores of +1 or -1 by the survival function (edf) at that observation to create a new score. The U score for the generalized Wilcoxon test is:

U_{ij} = $S(t_i) + S(t_{i-1}) - 1$ for all censored observations t ***(Equation 4.1)***

$S(t_{i-1}) - 1$ for all censored observation t*

Where $S(t_{i-1})$ is the value of the survival function for the previous uncensored observation. For the first observation in the dataset i=1, and the value of $S(t_0)$ equals 1. There is a 100 percent probability of exceeding a value smaller than the smallest observation in the data set. j grade concerns the groups.

The scores for one group are summed to obtain the test statistic W:

$$W = \sum_{i=1}^{n} U_i$$ ***(Equation 4.2)***

Dividing W by the square root of the variance for this statistic produces a Z statistic that can be compared to a table of the standard normal distribution. The permutation variance of W is " (Helsel D, 2005):

$$\mathrm{Var}\,[W] = \frac{mn \sum U^2}{(m+n)(m+n-1)}$$ ***(Equation 4.3)***

where m and n are the sample sizes that means the numbers of observations in the two groups.

$$Z = \frac{W}{\sqrt{\mathrm{Var}[W]}}$$ ***(Equation 4.4)***

<u>Detailed example:</u>

The dataset used is the Cadmium one from the NADA add-on package (Lee and Helsel, 2005;2007).

Table A3.6. *Cadmium dataset*

Cd	Monitoring region	CdCen*
81.3	SRKYMT	FALSE
3.5	SRKYMT	FALSE
4.6	SRKYMT	FALSE
0.6	SRKYMT	FALSE
2.9	SRKYMT	FALSE
3	SRKYMT	FALSE
4.9	SRKYMT	FALSE
0.6	SRKYMT	FALSE
3.4	SRKYMT	FALSE
0.4	COLOPLT	FALSE
0.8	COLOPLT	FALSE
0.3	COLOPLT	TRUE
0.4	COLOPLT	FALSE
0.4	COLOPLT	FALSE
0.4	COLOPLT	TRUE
1.4	COLOPLT	FALSE
0.6	COLOPLT	TRUE
0.7	COLOPLT	FALSE
0.2	SRKYMT	TRUE

*CdCen is the censoring variable, when it equals TRUE that means that the data is a nondetect one.

"Flipping the Cd data into a right-censored variable ("The flippedCd" column) produces values look like t or "time to censoring" of traditional survival analysis. The "Number Beyond" column lists the number of observation known to exceed the value of t. This is the same as the number of observation below the original cadmium concentration, and so equals the ranks of the cadmium observation minus one. The survival function S(t) is the probability of survival beyond each observation of FlipCd. This survival function is identical to the empirical distribution function of the original data, and equals i/n, where i is the rank of the original observation from low to high. Here tied observations were assigned tied rank as is standard hypothesis testing. For example, the three detected observations cadmium concentration of 0.4 would have had the ranks of 3, 4 and 5 had there been enough precision in the measurement to tell the observations apart. Without that precision, any of the three observations could be the highest, or lowest. All three are given a rank of 4, the median of the three possible ranks. In the survival analysis literature, tied values often follow another convention, assigning the minimum value for S, rather than the median value used here. Using the median assures that the sum of ranks for data with ties is the same as it would have been without ties, an important property for hypothesis tests".

If the null hypothesis is true, observations for each group will be randomly scattered the list in Table with about half of the scores positives and half negative. So W, the sum of scores will be near zero. If the null hypothesis is not true, the data from one group will be predominately near the top, or the bottom, of the list in Table 3.7. Consequently the absolute value of W will be larger than zero. From Table 3.7, the test statistic Z equals 2.637, and from a table of the standard normal distribution the associated one-sided p-value is 0.0042. The null hypothesis is soundly rejected, and it is conclude that cadmium concentration in fish livers in the Southern Rocky Mountains are higher than those in fish from streams in the Colorado Plateau" (Helsel D, 2005).

Table A3.7. *Computation of the generalized Wilcoxon test for the Cadmium data.*

Cd	Monitoring region	CdCen	FlipCd	No. Beyond	S(t)	U
81.3	SRKYMT	FALSE	18.7	18	0.947	0.947
4.9	SRKYMT	FALSE	95.1	17	0.895	0.842
4.6	SRKYMT	FALSE	95.4	16	0.842	0.737
3.5	SRKYMT	FALSE	96.5	15	0.789	0.632
3.4	SRKYMT	FALSE	96.6	14	0.737	0.526
3	SRKYMT	FALSE	97	13	0.684	0.421
2.9	SRKYMT	FALSE	97.1	12	0.632	0.316
1.4	COLOPLT	FALSE	98.6	11	0.579	0.211
0.8	COLOPLT	FALSE	99.2	10	0.526	0.105
0.7	COLOPLT	FALSE	99.3	9	0.474	0.000
0.6	SRKYMT	FALSE	99.4	7	0.368	-0.158
0.6	SRKYMT	FALSE	99.4	7	0.368	-0.158
0.6	COLOPLT	TRUE	99.4	6	0.000	-0.632
0.4	COLOPLT	FALSE	99.6	4	0.211	-0.421
0.4	COLOPLT	FALSE	99.6	4	0.211	-0.421
0.4	COLOPLT	FALSE	99.6	4	0.211	-0.421
0.4	COLOPLT	TRUE	99.6	2	0.000	-0.789
0.3	COLOPLT	TRUE	99.7	1	0.000	-0.789
0.2	SRKYMT	TRUE	99.8	0	0.000	-0.789
				W(SRKYMT) = Var (SRKYMT) = Z= one sided p-value		3.316 1.581 2.637 0.0042

Remarks

An alternate form of the test statistic exists which follows a chi-square with one degree of freedom distribution rather than the normal distribution." The value of the chi-square test statistics will approximately equal the square of the test statistics using the normal approximation" (Helsel D, 2005).

"The generalized Wilcoxon test can also be used to compare three or more distributions, analogous to the Kruskal-Wallis test"(Helsel D, 2005).

"[…] the test (Peto-Peto tests, or generalized wilcoxon test) is more powerful than the log-rank test, and is therefore more likely to detect true differences when data come from a lognormal distribution (Lee, 1992). The Peto–Peto test ''gives more weight to early failures'', meaning that it is sensitive to differences in the higher values of left-censored data sets (Lee, 1992). Because many environmental data sets are approximately lognormal, and the upper portions of groups are where detected differences often occur, the Peto–Peto test is judged to be the most appropriate […]"(Lee and Helsel, 2007).

Annex 4: General Description of Trend Detection Techniques

The trend detection techniques used in this report are as follows:

1. Kendall's Tau Correlation
2. Mann-Kendall test
3. Theil Slope test
4. Pearson's Correlation
5. Model Utility Test for Simple Linear Regression Model
6. Spearman Correlation
7. Independent two sample heteroscedastic "t" test
8. Wilcoxon Rank Sum test
9. Mann-Whitney test
10. Fryer and Nicholson Lowess test as implemented by Trend-Y-Tector software
11. Lag 1 autocorrelation test

Tests 4 and 5 above are in fact equivalent and so only test 4 has been used. Tests 8 and 9 are also equivalent.

It should be noted that testing the Theil Slope and conducting the Mann-Kendall test are actually equivalent to testing Kendall's Tau correlation. Consequently discussions are restricted to the Kendall's Tau test.

Kendall's Tau

This is a measure of the strength of association between a set of observations X and another rset Y. Let (X_i, Y_i) and (X_j,Y_j) be a pair of (bivariate) observations. If X_j -X_i and Y_j - Y_i have the same sign, we shall say that the pair is *concordant*, if they have opposite signs, we shall say that the pair is *discordant.* In the (x,y)-plane points with a positive slope ${}_{+}{}^{+}$ form a concordant pair, while the points with a negative slope ${}^{+}{}_{+}$ form a discordant pair.

Given *n* pairs of observations (X_i,Y_i) we can form n(n-1)/2 pairs corresponding to choices 1 <= i <j<= *n.* Let *C* stand for the number of concordant pairs and *D* stand for the number of discordant pairs. Kendall's S may be computed as S = *C* - *D* and this clearly measures the association between X and Y.

S may be standardized by computing tau =2S/n(n-1) which will always have values between -1 and 1. Tau is called Kendall's correlation coefficient.

The maximum value +1 is achieved if all *n(n* -l)/2 pairs are concordant which corresponds to a monotonic increasing trend for all points in the X-Y plane and the minimum value -1 is achieved if all pairs are discordant (monotonic decreasing trend)

Pearson's Correlation Coefficient

Pearson's correlation coefficient differs from Kendall's tau in that it measures the strength of the linear association between X and Y. A non-linear association may not necessarily be detected by Pearson's correlation coefficient "r". It is computed for a sample of n (x,y) observations as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Pearson's correlation like Kendall's tau takes values between -1 and +1 with values close to +1 indicating a strong positive linear association between X and Y and values close to -1 indicating a strong negative linear association between X and Y. Values close to 0 indicate no linear association between X and Y there may however be a nonlinear association.

Spearman's Correlation Coefficient

To compute Spearman's correlation coefficient, first consider the N X observations. A rank is assigned to each observation which determines its position in the set of X observations. So for example the smallest X observation will receive rank 1 and the largest rank N. Separately assign ranks to the Y observations. Finally compute Spearman's correlation coefficient by using the same formula as for Pearson's correlation but replacing each (x,y) pair by the corresponding pair of x and y ranks. Spearman's correlation has two potential advantages over Pearson's correlation: firstly it does not require the normality assumption; secondly it does not test only for a linear relationship but rather for any monotonic relationship.

Independent two sample heteroscedastic "t" test also called Welch-Aspin approximate test

The two sample t test may be used to test for the presence of a trend as follows. Consider a series of observations from time T_{1995} to time T_{Now} with a reference time set at 2001. Split the observations into two subsets, those before the reference year in what is called the baseline period and those after 2001. Compute the mean and variance of the observations from the baseline period as $\bar{x}_b$ and s_b^2 and the mean and variance of the observations from the post 2001 period as $\bar{x}_a$ and s_a^2 . Let n_a and n_b be the number of observations in samples A and B respectively. Then assume that the baseline observations are chosen independently from a Normal distribution and that the post 2001 observations are chosen independently from a possibly different Normal distribution then we may test the hypotheses that there is a significant difference between the true mean in the baseline period and the mean in the post 2001:

$$H_o: \mu_b - \mu_a = 0 \quad \text{vs} \quad H_a: \mu_b - \mu_a < 0$$

using the test statistic:

$$t = \frac{\bar{x}_b - \bar{x}_a}{\sqrt{\frac{s_b^2}{n_b} + \frac{s_a^2}{n_a}}}$$

This statistic follows under the H_0 hypothesis a Student distribution with $\dfrac{\left(s_a^2/n_a + s_b^2/n_b\right)^2}{\dfrac{\left(s_a^2/n_a\right)^2}{n_a - 1} + \dfrac{\left(s_b^2/n_b\right)^2}{n_b - 1}}$ degrees of freedom[13].

Wilcoxon Rank Sum Test

This test is a non parametric version of the two-sample t-test. Test the hypotheses that there is a significant difference between the true mean in the baseline period and the mean in the post 2001:

$$H_O: \mu_b-\mu_a=0 \quad \text{vs} \quad H_a: \mu_b-\mu_a <0$$

The data are split into two groups as before:

Baseline Group: $B_1, \ldots, Bn_1$

Post 2001 Group: $A_1, \ldots, An_2$

1. Combine the samples into one sample of Wi's. Order data in the combined sample W(1) ,W(2) , . . . , $W(n_1+n_2)$
2. Assign rank i to the ith smallest observation (in the case of ties, assign the average rank to each observation)
3. Let R_1 = sum of ranks attached to observations in sample 1
4. $K_1 = R_1 - n_1(n_1+1)/2$
5. The test statistic is Uobs = $\max(K_1,\ n_1n_2 - K_1)$ and it follows a special Wilcoxon Rank Sum distribution.

Trend-Y-Tector Lowess Test

The essential idea is to fit a smooth curve f(t) to the time series data. This smooth curve is supposed to be a better representation of the underlying process with random variation removed. The principle is then to assess which of three hypotheses is more appropriate:

H0: f(t)= Smoother is constant

H1: f(t)= Smoother is a linear function of time

H2: f(t)= Smoother is an unspecified smooth function of time i.e. the level is not changing, there is a linear trend, or there is a more complex change taking place.

Three formal tests are then conducted:

Loess Level Test:

H0 vs. H2: Do contaminant levels vary with time?

If so then one attempts to establish if the trend is linear or more complex

Loess Linear Test:

H0 vs. H1: Do contaminant levels vary linearly with time?

[13] http://en.wikipedia.org/wiki/Student's_t-test and Annex 5 of PE1.

Loess Non-linear effect:

H1 vs. H2: Do contaminant levels vary non-linearly with time?

This test is described in the attached paper by Fryer and Nicholson. Implementing this test is a difficult task. To date I have identified two programmes both developed under the auspices of OSPAR which have incorporated the test:

Trend-Y-Tector	**http://www.trendytector.nl/**
R-Trend	**http://www.quodata.de/**

As it is freely available, the Trend-Y-Tector has been used for this work.

Lag 1 Autocorrelation Test

The Lag-1 autocorrelation is defined as

$$r_1 = \frac{\sum_{i=1}^{n-1}(x_t - \overline{x})(x_{t+1} - \overline{x})}{\sqrt{\sum_{t=1}^{n}(x_t - \overline{x})^2}}$$

This correlation measures the association between observations at time “t” and at time “t-1”. Knoke (1975) suggested that this autocorrelation could be used as a test of non-randomness in data and consequently as a trend detection tool. The autocorrelation follows an approximate Normal distribution with mean and variance given by:

$$mean = \frac{-1}{n}$$

$$\mathrm{var} = \frac{(n-2)^2}{n^2(n-1)}$$

This can be used to test the hypotheses:

H_o: r_1=0 vs H_a: r_1 >0

using the standard normal test statistic

$$z = \frac{r_1 + \frac{1}{n}}{\sqrt{\frac{(n-2)^2}{n^2(n-1)}}}.$$

OSPAR
COMMISSION

New Court
48 Carey Street
London WC2A 2JQ
United Kingdom

t: +44 (0)20 7430 5200
f: +44 (0)20 7430 5225
e: secretariat@ospar.org
www.ospar.org

OSPAR's vision is of a healthy and diverse North-East Atlantic ecosystem

ISBN 978-1-906840-94-5
Publication Number: 454/2009