Evaluation and fine-tuning of a procedure for statistical analysis of beach litter data

Icastat

amo

Modellering en Optimalisatie

Evaluation and fine-tuning of a procedure for statistical analysis of beach litter data

Authors:
drs. Paul K. Baggelaar (*Icastat*)
ir. Eit C.J. van der Meulen (*AMO*)

October 30 2014

Client: Rijkswaterstaat (the executive body of the Dutch Ministry of Infrastructure and the Environment)
Contact: dr. Willem van Loon

Modellering en Optimalisatie

Niagara 18
1186 JP AMSTELVEEN
Tel.: 020 641 52 11
paul.baggelaar@planet.nl

CONTENTS

EXECUTIVE SUMMARY 3

TECHNICAL SUMMARY AND RECOMMENDATIONS 5

1 INTRODUCTION 9

1.1 BACKGROUND 9
1.2 ABOUT THIS REPORT 9
1.3 SOFTWARE PROGRAM LITTER ANALYST 10

2 BEACH LITTER DATA FOR THIS STUDY 11

2.1 SELECTION OF DATA 11
2.1.1 Data from the OSPAR database 11
2.1.2 Data from the SDN spreadsheet 11
2.2 ADMINISTRATIVE CHANGES OF THE OSPAR BEACH LITTER CLASSIFICATION IN 2010 12
2.3 ENCOUNTERED DATA PROBLEMS AND OUR PROPOSED SOLUTIONS 13
2.3.1 Survey dates that occur more than once 13
2.3.2 Incomplete application of the administrative changes 13
2.3.3 Incorrect uses of zeros and blanks 14
2.3.4 Truncated time series caused by the administrative changes in 2010 14
2.4 SUMMARY OF PROCEDURE FOR BEACH LITTER DATA CLEANUP AND PREPARATION 17

3 TREND ANALYSIS OF BEACH LITTER TIME SERIES 19

3.1 BACKGROUNDS 19
3.2 TREND ANALYSIS METHODS APPLICABLE FOR BEACH LITTER TIME SERIES 20
3.2.1 Seasonal effects? 21
3.2.2 Residuals normal? 22
3.2.3 Residuals autocorrelation? 22
3.2.4 Series autocorrelation? 22
3.2.5 Selected trend tests 22
3.3 CONCLUSIONS REGARDING THE PREFERRED TREND ANALYSIS METHOD 23
3.4 PROBLEMS WITH TIES IN TIME SERIES OF BEACH LITTER DATA 24
3.5 CHOOSE FOR ONE-SIDED OR TWO-SIDED STATISTICAL TESTING ON TREND? 25

4 CHARACTERISING THE GROUP OF ESTIMATED TRENDS 26

4.1 CHARACTERISING THE GROUP OF TRENDS WITH ITEM TREND INDICES 26
4.2 CHARACTERISING THE GROUP OF TRENDS WITH PERCENTILES 27

5 DETECTING A CHANGE OF BEACH TOTAL COUNT 28

5.1 TESTING ON A MONOTONIC TREND OF BEACH TOTAL COUNT 28
5.2 TESTING ON A STEP TREND OF BEACH TOTAL COUNT 29

6 PRESENTING THE RESULTS: THE EVALUATION MATRIX 31

6.1 METADATA 31
6.2 GROUP 1: SUMMARY OF RESULTS FROM TOP-X ITEMS 31
6.3 GROUP 2: SUMMARY OF STATISTICALLY SIGNIFICANT TRENDS 32
6.4 GROUP 3: SUMMARY OF RESULTS FROM BEACH TOTAL COUNT 33
6.5 GROUP 4: SUMMARY OF TRENDS (SLOPES) OF TOP-X ITEMS 33
6.6 EVALUATION MATRIX OF ITEMS VERSUS THOSE OF SOURCES AND MATERIALS 34

7 SOURCES AND MATERIALS ANALYSIS 35

7.1 SOURCES ANALYSIS 35
7.2 MATERIALS ANALYSIS 35

8 AGGREGATION OF RESULTS TO LARGER SPATIAL SCALES? 36

8.1 DISCUSSION 36

9 SUMMARY OF PROPOSED PROCEDURE 37

REFERENCES 38

APPENDIX 1: INITIAL SPECIFICATIONS OF THIS EVALUATION STUDY 40

APPENDIX 2 (DIGITAL): EVALUATION MATRICES OF SELECTED BEACHES 43

Executive summary

This study evaluates a procedure for the statistical analysis of beach litter data, that was developed by Germany and The Netherlands, as members of the OSPAR Intersessional Correspondence Group Marine Litter.

We encountered various problems with the beach litter data, that should be taken care of before applying a statistical analysis of the data. To handle these problems, we developed a procedure for data cleanup.

By exploring the statistical characteristics of a large number of beach litter time series, we found that there is no trend test that is optimal for all beach litter time series. Therefore, the trend analysis should preferably be tailor-made, applying to each individual time series the trend test and accompanying slope estimator that best fit the characteristics of that series. A sub-optimal solution is to test all time series on trend using the Mann-Kendall test, because this will be the optimal / nearly optimal test for the vast majority of the series (> 80%) that according to the procedure should be analyzed on trend.

The procedure for the statistical analysis of beach litter data uses a trend index to summarize the trend signs of all items with statistically significant trends. We propose a slight modification of this index and also two additional trend indices, because the original index is not sensitive to the magnitudes of the estimated slopes. These three indices are only based on the slopes that are statistically significant, which will diminish the sensitivity for a general change – affecting many items - in one direction (improvement or deterioration). Therefore, we advise to also characterise the group of all estimated trends (slopes), regardless of their statistical significances. This can be done by presenting their statistics, such as the minimum, the 25-percentile, the 50-percentile (this is the median), the 75-percentile and the maximum.

To detect a change in the average of beach total count between two assessment periods of each six years, the trend analysis should preferably be tailor made, but sub-optimal solutions are available too, such as the Mann-Kendall test on a monotonic trend (with a linear, concave, or convex form) or the Wilcoxon-ranksum test (the distribution-free equivalent of the t-test) to test on a step trend. We propose to define the average of beach total count over an assessment period as the median of the beach total count per survey. This definition is robust against missing data and it takes into account that most time series of beach total count do not come from a normal distribution.

To enable a concise presentation of the results of the various statistical analyses, we developed an evaluation matrix. It can present for each separate beach four groups of results. Examples are provided in the digital appendix of this report.

The applied selection strategy of beaches (or the lack of it) does not allow for a meaningful extrapolation of beach specific results to larger spatial scales. Such an extrapolation is only meaningful if the beaches were selected using some form of probabilistic sampling, but that was not the case. Some possibilities to aggregate beach results to larger spatial scales are described in the proposal for the statistical procedure. We propose to apply those methods to aggregate results from various beaches, but to present them as being strictly only representative for that specific group of sampled beaches. They should not be presented as representa-

tive for some regional or national population of beaches. Care should be taken in using these results for the selection of measures or targets.

As part of our study we also developed the software program Litter Analyst. It can apply the data cleanup procedure of the beach litter data and the procedure for statistical analysis, using data from the two assessment periods 2002 – 2007 and 2008 – 2013. This program will be made available as a beta version to the OSPAR members until March 1 2015.

Technical summary and recommendations

In this study we evaluated a procedure for the statistical analysis of beach litter data, that was developed by Germany and The Netherlands, as members of the OSPAR Intersessional Correspondence Group Marine Litter.

Cleanup the data before the statistical analysis

We encountered the following problems with the beach litter data: 1) survey dates that occur more than once for the same beach, 2) incomplete application of the administrative changes that came into operation in 2010, 3) incorrect uses of zeros and blanks and 4) truncated time series, caused by the administrative changes that came into operation in 2010. These problems should be taken care of before applying a statistical analysis of the data. To handle the data problems, we developed a procedure for data cleanup.

Trend analysis

By exploring the statistical characteristics of a large number of beach litter time series, we determined which trend analysis methods are applicable for beach litter time series. We considered only methods for the analysis of a monotonic trend, as these have more general applicability than methods for the analysis of a step trend. A time series has a monotonic trend if the average of the series (this is the arithmetic average in case of a symmetrical distribution and the median in case of a skewed distribution) generally changes in the same direction. The change can be linear, convex or concave, or a combination of these forms.

We found that there is no trend test that is optimal for all beach litter time series. Therefore, trend analysis of such time series should preferably be tailor-made, using a procedure that selects for each individual time series the trend test (and accompanying slope estimator) that best fits the characteristics of that series. If however, the complex software and/or knowledge to apply the intricate selection procedure is not available, a sub-optimal solution is to test all time series on trend using the Mann-Kendall test (and to estimate the trends of all time series with the Theil-Sen slope estimator). This will be the optimal / nearly optimal test for the vast majority of the series (> 80%) that according to the procedure should be analyzed on trend.

For time series with high percentages of zeros, trend analysis methods based on the Poisson distribution will be more appropriate than the combination of the Mann-Kendall test and the Theil-Sen slope estimator. This problem will hardly occur for the series of the top-X list (according to the procedure only these series should be analyzed on trend), because they mostly contain low to moderate percentages of zeros (the top-X list is a list of the minimal number of top items – but at least 10 - that covers at least 80% of the total count of all items). Problems can occur if trend analysis is applied to series of categories of sources or materials, because some of these have very high percentages of zeros.

One-sided or two-sided testing on trend?

The choice for one-sided or two-sided statistical testing should not be based on a visual inspection of the data, because that will invalidate the confidence level of the test. The choice must be made beforehand, solely based on theoretical considerations. For the statistical analysis of beach litter time series one will be interested in both positive and negative trends, therefore we advise to apply two-sided trend testing.

Characterising the group of trends

In the procedure for the statistical analysis of beach litter data the trends are summarized with a trend index that is the weighted average of the reversed trend signs of all items with statistically significant trends. We propose to remove the reversal of the trend sign, because it creates confusion. We also propose to present two additional trend indices, because the original index is not sensitive to the magnitudes of the estimated slopes. The first additional index is the sum of slopes and the second is the weighted average of slopes, where each weight is the relative contribution of the item count to the beach total count. Each of these indices integrates the information about developments of individual items, but they differ in information content.

The three indices are only based on the slopes that are statistically significant, because in calculating them all slopes that are not statistically significant or are not estimated are set to 0. This censored approach will diminish the sensitivity for a general change – affecting many items - in one direction (improvement or deterioration). Therefore, we advise to also characterise the group of all estimated trends (slopes), regardless of their statistical significances. This can be done by presenting their statistics. We propose to present the minimum, the 25-percentile, the 50-percentile (this is the median), the 75-percentile and the maximum of this group.

Detecting a change of beach total count

We propose two statistical tests to detect a change in the average of beach total count between two assessment periods of each six years. The first one tests on a monotonic trend and the second one on a step trend.

Our study indicates that non-normality is fairly common in time series of beach total count. This should be taken into account when choosing a trend test. Preferably, the choice should be tailor-made for each time series of total count, selecting a statistical test that best fits the characteristics of that series. If however, the complex software and/or knowledge to apply the intricate selection procedure is not available, we advise the same sub-optimal solution as proposed for the trend analysis of the top-X items, that is to apply only the Mann-Kendall test. This will be the optimal / nearly optimal test for the vast majority of the explored beach total series (75 - 80%). It is implemented in the software program Litter Analyst, as a part of this study.

A test on a step trend will be more appropriate than a test on a monotonic trend if it can be assumed that a change in the average of beach total count occurs in the form of a more or less instantaneous step. In the case of normality the t-test will be the best choice of step trend test, whereas in the case of non-normality the Wilcoxon-ranksum test (the distribution-free equivalent of the t-test) will be the best choice. In the cases of seasonality, autocorrelation, or both, the best choices will be adapted versions of the t-test and the Wilcoxon-ranksum test, that account for these characteristics. It is not easy however, to determine the characteristics of each series. A feasible sub-optimal solution is to test all time series of total count on a difference in the average between two assessment periods using the Wilcoxon-ranksum test.

The arithmetic average of the yearly beach total count is not a good choice to define the average of beach total count over an assessment period, because it is not robust against missing data. Possible solutions to this problem are: i) estimate the missing values, or ii) ignore data, such that each yearly total is derived from the same number of surveys. However, both solutions require complex algorithms. But perhaps an even greater disadvantage of the definition is that it implicitly assumes that the total count comes from a normal distribution, which will probably be invalid in about 65% of the cases. Therefore we propose to define the average of beach total count over an assessment period as the median of the beach total count per survey. This definition is robust against missing data and it takes into account that most time series of beach total count do not come from a normal distribution.

The relative change in the average of beach total count from assessment period A to assessment period B should preferably be quantified using the Hodges-Lehmann estimator. This is the median of all possible pairwise differences between data of period B and data of period A. It is a more precise estimator of the difference between the population medians of the two periods than the difference between the sample medians of the two periods.

Presenting the results: the evaluation matrix

To enable a concise presentation of the results of the various statistical analyses, we developed an evaluation matrix. It can present for each separate beach four groups of results. Examples are provided in the digital appendix of this report.

Aggregation of results to larger spatial scales?

We think that the applied selection strategy of beaches (or the lack of it) does not allow for a meaningful extrapolation of beach specific results to larger spatial scales. Such an extrapolation is only meaningful if the beaches were selected using some form of probabilistic sampling, but that was not the case.

Of course, it is very tempting to extrapolate beach results to larger spatial scales, because that can help to underpin important decisions. But if that is the main purpose of this monitoring system, the selection strategy of the beaches should be adapted to this information need. We should not stretch beyond the possibilities of the present selection strategy to fulfill our needs. Some possibilities to aggregate beach results to larger spatial scales are described in the proposal for the statistical procedure. We propose to apply those methods to aggregate results from various beaches, but to present them as being strictly only representative for that specific group of sampled beaches. They should not be presented as representative for some regional or national population of beaches. Care should be taken in using these results for the selection of measures or targets.

Recommendations

- Apply our procedure for beach litter data cleanup before applying the procedure for statistical analysis. Otherwise the many data problems that occur can make the results of the analysis meaningless.
- Register the real survey dates in the database and not only the first day of each quarter. This information can be important for various purposes and enables an evaluation of the survey interval.
- Try to keep the survey interval as close as possible to three months, because if the survey interval is not constant, it will be difficult to determine important process characteristics such as seasonality and autocorrelation.
- Do not change the list and definitions of registered beach litter items any further, because that will seriously damage the monitoring goal of detecting long term changes in the counts of beach litter data.
- Examine if a combination of trend test and slope estimator based on the Poisson distribution is available for count series with a high percentage of zeros. This is because some count

series of categories of sources and materials can have high percentages of zeros, that are better dealt with using statistical methods based on the Poisson distribution.

1 Introduction

1.1 Background

To monitor the amounts and sources of marine litter in the North East Atlantic region by the OSPAR countries, standardized protocols for litter surveys of beach stretches of 100 metres and 1 km were developed in 2000. In each of the participating countries these surveys are carried out on a quarterly basis since about 2002 (and sometimes later) at various beaches. Recently Germany and The Netherlands developed a common procedure for the statistical analysis of beach litter data [Schulz et al., 2014], on request of the OSPAR Intersessional Correspondence Group Marine Litter. It was agreed upon that both countries will test and evaluate the procedure in 2014, using national data and OSPAR data. Rijkswaterstaat has asked Icastat to do this evaluation study on behalf of the Netherlands and to follow the initial study guidelines described in [Van Loon, 2014] as closely as possible. In this report we present our findings, leading to some recommendations for fine-tuning of the procedure.

About OSPAR

The Convention for the Protection of the Marine Environment of the North-East Atlantic (known as 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.

Scope and limitations of this study

The initial guidelines of this study – shown in the appendix of this report - were set up by Willem van Loon of Rijkswaterstaat. We tried to follow these guidelines as closely as possible. This implied that we did not evaluate or develop other approaches for the statistical analysis of beach litter data than described in these guidelines.

In the course of our study various limitations of the data and the data analysis methods became apparent and so it was agreed upon with Willem van Loon to take a different path for certain study elements than described in the study guidelines. Furthermore, because of the many data problems that we encountered, much effort was needed to develop and test a procedure for data cleanup. This reduced the amount of time that was available for other tasks of this evaluation study.

Assessment periods used for the testing

On request of Rijkswaterstaat, in this study the procedure for the statistical analysis of beach litter data is evaluated for the assessment periods 2002 – 2007 and 2008 – 2013 and also for the combined period 2002 - 2013. The use of assessment periods of six years is as specified in the MSFD (Marine Strategy Framework Directive) and it is also advised by [Van Loon, 2014].

1.2 About this report

After this introduction, Chapter 2 describes how we selected beach litter data for this study and how we handled the many serious problems of these data. In Chapter 3 the best strategy for trend analysis of beach litter time series is derived, based upon the general characteristics of these series. Chapter 4 describes two methods to characterise a group of estimated trends.

Chapter 5 proposes two approaches to detect a change of beach total count, the first one using a test on monotonic trend and the second one using a test on step trend. Chapter 6 describes and explains the evaluation matrix that we developed to present the results of the various statistical analyses in a concise way. Chapter 7 discusses some specifics of source and materials analysis. In Chapter 8 we explain why the aggregation of results of various beaches to regional or national scales is not recommendable. Chapter 9 summarizes the steps of the proposed procedure for statistical analysis of beach litter data, after taking into account the various recommendations for fine-tuning of the procedure. The main part of this report ends with the references in alphabetical order.

This report has two appendices. Appendix 1 contains the initial guidelines of this evaluation study, as set up by Willem van Loon of Rijkswaterstaat. Appendix 2 is a separate digital appendix that contains the three evaluation matrices of the beaches that were selected for this study, respectively the items evaluation matrix, the sources evaluation matrix and the materials evaluation matrix. They are examples of the output of the proposed procedure for statistical analysis of beach litter analysis.

1.3 Software program Litter Analyst

As part of our study we also developed the software program Litter Analyst. It can apply the data cleanup procedure of the beach litter data and the procedure for statistical analysis, using data from the two assessment periods 2002 – 2007 and 2008 – 2013. This program will be made available as a beta version to the OSPAR members until March 1 2015. The digital appendix 2 of this report shows the output of this program for some selected beaches.

2 Beach litter data for this study

This chapter describes various aspects of the beach litter data that we used for this study. § 2.1 gives the details of our data selection. In § 2.2 we describe the administrative changes of the OSPAR classification of beach litter that came into operation in 2010, because they caused a great deal of the data problems that we encountered. In § 2.3 we describe these data problems and also how we handled them in the data cleaning procedure. Our proposed procedure for beach litter data cleanup is summarized in § 2.4.

2.1 Selection of data

For this study we downloaded beach litter data of the 100 meter surveys from the OSPAR database. And we received an Excel file with Dutch beach litter data from SDN[1], a non-governmental environmental organisation, that is responsible for the beach litter surveys in the Netherlands. According to Willem van Loon, the supervisor of this study, the SDN data are presumably more reliable than the Dutch data in the OSPAR database.

2.1.1 Data from the OSPAR database

The data from the OSPAR database were downloaded for each separate country as a text file in csv-format. Each of these files contains the counts of 126 litter categories, for each combination of beach and survey date.

In these csv-files most survey dates are exactly equal to the start date of the quarter (Januari 1, April 1, July 1, or October 1). Probably these are not the real survey dates. Only for the most recent years (2012 or 2013) other survey dates occur, that probably are the real survey dates. The csv-files also give sequence numbers (1, 2, 3, or 4) to the surveys in a year, in column with the heading 'Period'. However, these sequence numbers are not always equal to the quarter in which the survey was done.

We recommend to always register the real survey dates, as this information can be important for various purposes and enables an evaluation of the survey interval. Furthermore, we recommend to try to keep the survey interval as close as possible to three months, because if the survey interval is not constant, it will be difficult to determine process characteristics such as seasonality and autocorrelation.

2.1.2 Data from the SDN spreadsheet

The SDN spreadsheet contains the counts of 121 litter categories and yes/no for the presence of plastic pellets, from four Dutch beaches. Its classification of beach litter categories is not fully compatible with that of the OSPAR database, because SDN only partially applied the administrative changes of the OSPAR classification, that came into operation in 2010 (see § 2.2 below).

[1] SDN is an abbreviation of Stichting De Noordzee (North Sea Foundation).

2.2 Administrative changes of the OSPAR beach litter classification in 2010

In 2010 various administrative changes of the OSPAR classification of beach litter came into operation. For the 100 m survey this meant that:

- 5 item-codes got a different definition, starting in 2010 (31, 32, 46, 62 and 84) and - to avoid confusion - their time series from before 2010 got new item-codes (200, 201, 202, 204 and 205), as illustrated below:

Initial code	Period 2002 - 2009 new code: old definition	Period 2010 and later old code: new definition
31	200: Plastic rope/cord/nets < 50 cm	🟧
	🟧	31: Plastic rope (diameter > 1 cm)
32	201: Plastic rope/cord/nets > 50 cm	🟧
	🟧	32: Plastic string/cord (diameter < 1 cm)
46	202: Plastic/polystyrene pieces < 50 cm	🟧
	🟧	46: Plastic/polystyrene pieces 2.5 - 50 cm
62	204: Cartons/tetrapaks	🟧
	🟧	62: Non-milk cartons/tetrapaks
84	205: Metal oil drums (new, not rusty)	🟧
	🟧	84: Metal oil drums (new and old)

Explanation

🟧 beach litter item not on survey form in that period (and therefore not registered in survey)

- 5 item-codes were removed (51, 58, 85, 106 and 107) and their time series from before 2010 got new item-codes (203, 210, 206, 207 and 208), as illustrated below:

Initial code	Period 2002 - 2009 new code: old definition	Period 2010 and later item not on survey form
51	203: Rubber gloves	🟧
58	210: Textile rope/strings	🟧
85	206: Metal oil drums (old, rusty)	🟧
106	207: Human faeces	🟧
107	208: Animal faeces	🟧

- 10 other new items were introduced (112 - 121), as illustrated below:

Initial code	Period 2002 - 2009 item not on survey form	Period 2010 and later code: definition
112	🟧	112: Plastic bag ends
113	🟧	113: Rubber gloves (industr./profess.)
114	🟧	114: Plastic lobster and fish tags
115	🟧	115: Plastc nets and pieces of net < 50 cm
116	🟧	116: Plastic nets and pieces of net > 50 cm
117	🟧	117: Plastic/polystyrene pieces < 2.5 cm
118	🟧	118: Cartons/tetrapaks (milk)
119	🟧	119: Wooden fish boxes
120	🟧	120: Disposable metal BBQ’s
121	🟧	121: Bagged dog faeces

2.3 Encountered data problems and our proposed solutions

We encountered the following problems with the beach litter data that we downloaded and received:

1. survey dates that occur more than once for the same beach;
2. incomplete application of the administrative changes that came into operation in 2010;
3. incorrect uses of zeros and blanks;
4. truncated time series, caused by the administrative changes that came into operation in 2010.

The following four subparagraphs describe the elements of the data cleanup and preparation procedure that we developed to handle these problems.

2.3.1 Survey dates that occur more than once

The downloaded OSPAR data sometimes contain survey dates that occur two or even three times for the same beach. In the data from England there are two of these errors, in those from Ireland four and in those from Sweden six. Dependent upon the specific situation, we used the following solutions:

- the records are identical: remove one or two of these records, such that only one remains;
- only one of the records has a total item count > 0: remove the record(s) with total item count = 0;
- the records are different and each has a total item count > 0: contact the person responsible for the beach litter data of that country, to find a solution.

2.3.2 Incomplete application of the administrative changes

For various item-codes affected by the administrative changes that came into operation in 2010, the csv-output from the OSPAR database sometimes contains counts in periods that the item-code was not yet/anymore defined. We corrected for these inconsistencies with the steps 3 and 4 that are described in § 2.3.3.

The spreadsheet of Dutch beach litter data from SDN suffers from an incomplete application of the administrative changes that came into operation in 2010. This was also noted before [Van Franeker, 2013], but until now it was not remediated. The following OSPAR changes were incompletely applied by SDN:

- OSPAR change: 5 item-codes got a different definition, starting in 2010 (31, 32, 46, 62 and 84) and - to avoid confusion - their time series from before 2010 got new item-codes (200, 201, 202, 204 and 205). However, SDN did not create the new time series 200, 201, 202, 204 and 205, so the time series of 31, 32, 46, 62 and 84 cover the whole period 2002 – 2013. Because their definitions change in 2010, each of these series is inconsistent;
- OSPAR-change: 5 item-codes were removed (51, 58, 85, 106 and 107) and their time series from before 2010 got new item-codes (203, 210, 206, 207 and 208). However, SDN did not create the new time series 203, 210, 206, 207 and 208 and did not remove the item-codes 51, 58, 85, 106 and 107.

We corrected for these incomplete applications with the following two steps:

1. Create the item-codes 200 – 208 and 210 and fill them with the data from before 2010 of the item-codes 31, 32, 46, 51, 62, 84, 85, 106, 107and 58 respectivily (also see table 2.1 below for the combinations of item-codes to use).
2. Remove the item-codes 51, 58, 85, 106 and 107.

Table 2.1: Consequences of the administrative changes in 2010, where new item-codes were introduced and some existing item item-codes changed definition or were removed.

	Initial item-code																			
Year	31	32	46	51	58	62	84	85	106	107	112	113	114	115	116	117	118	119	120	121
2002	200	201	202	203	210	204	205	206	207	208	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
2003																				
2004																				
2005																				
2006																				
2007																				
2008																				
2009																				
2010	31	32	46	🟧	🟧	62	84	🟧	🟧	🟧	112	113	114	115	116	117	118	119	120	121
2011																				
2012																				
2013																				

Explanation	
🟧	beach litter item not on survey form in that period (and therefore not registered in survey)

2.3.3 *Incorrect uses of zeros and blanks*

A zero should only be used for a zero count and a blank should only be used for a missing count. But in the beach litter data zeros and blanks are often used wrong. These are serious errors, because they can severely distort the results of statistical analyses.

The csv-files that we downloaded from the OSPAR database mostly use zeros (and sometimes counts, see § 2.3.2) for periods that the item was not (yet / anymore) defined in the OSPAR classification system. However, these zeros (and counts) should be blanks, because counting was not possible in those periods.

The same error with zeros is made in the SDN spreadsheet with Dutch beach litter data. But also surveys that were not held (for example because of bad weather) are filled with zeros (or sometimes a text comment), instead of blanks. Furthermore, blanks are sometimes used for items that were not found, instead of zeros (count = 0).

We corrected for these errors with the following four steps.

1. In case only zeros and/or blanks and/or text comments occur for a certain survey date, set them all to blanks.
2. In case at least one item count is > 0 for a certain survey date, set all blanks of that date to zero.
3. Set all data from before 2010 of the items 31, 32, 46, 62, 84 and 112 - 121 to blanks.
4. Set all data from 2010 and later of items 200 – 208 and 210 to blanks.

2.3.4 *Truncated time series caused by the administrative changes in 2010*

The administrative changes that came into operation in 2010 led to 25 truncated time series, of which 10 stop after 2009 (truncation at the end) and 15 start in 2010 (truncation at the start). If these truncated series are included in statistical analyses over time periods that start before and end after 2010, they can distort the overall picture. Therefore, we propose to combine them into clusters for the statistical analysis, such that each cluster meets the following three conditions:

1. the cluster contains at least one time series that is truncated by the administrative changes in 2010;
2. the time series of the cluster is consistent, meaning that it represents the same subgroup of beach litter over its entire length;
3. the cluster contains the minimum number of time series to fulfill condition 2. Therefore, it should not contain time series that were unaffected by the administrative changes in 2010.

After discussing this matter with Willem van Loon, the study supervisor, it was proposed to create the six clusters that are shown below (with proposed codes 300 – 305). They contain 18 of the 25 truncated time series.

Proposed code and definition: 300 - Nets and ropes		
Code	2002 - 2009	2010 and later
31	🟧	Plastic rope (diameter > 1 cm)
32	🟧	Plastic string and cord (diameter < 1 cm)
115	🟧	Plastic nets and pieces of net < 50 cm
116	🟧	Plastic nets and pieces of net > 50 cm
200	Plastic rope/cord/nets < 50 cm	🟧
201	Plastic rope/cord/nets > 50 cm	🟧

Explanation	
🟧	beach litter item not on survey form in that period (and therefore not registered in survey)

Proposed code and definition: 301 - Plastic polystyrene pieces < 50 cm		
Code	2002 - 2009	2010 and later
46	🟧	Plastic/polystyrene pieces 2.5 - 50 cm
117	🟧	Plastic/polystyrene pieces < 2.5 cm
202	Plastic/polystyrene pieces < 50 cm	🟧

Proposed code and definition: 302 - All cartons/tetrapaks		
Code	2002 - 2009	2010 and later
62	🟧	Cartons/tetrapaks (not milk)
118	🟧	Cartons/tetrapaks (milk)
204	Cartons/tetrapaks	🟧

Proposed code and definition: 303 - Other Textiles		
Code	2002 - 2009	2010 and later
59	Other textiles	
210	Textile rope/strings	🟧

Proposed code and definition: 304 - All gloves		
Code	2002 - 2009	2010 and later
25	Gloves	
113	🟧	Rubber gloves (industrial/professional)
203	Rubber gloves	🟧

Proposed code and definition: 305 - All metal oildrums		
Code	2002 - 2009	2010 and later
84		Metal oil drums
205	Metal oil drums (new, not rusty)	
206	Metal oil drums (old, rusty)	

The statistical analysis should not include the 20 individual time series that are used to form this six clusters. The corresponding item numbers are shown below.

Item used in cluster

25	31	32	46	59	62	84	113	115	116
117	118	200	201	202	203	204	205	206	210

As proposed earlier by [Van Franeker, 2013], we also exluded two other groups of time series from the statistical analysis, for the reasons given by Van Franeker.

1. The first group is the Pollutants items, for paraffin like or other pollutants (e.g. oily, or palmoil) wastes on the beach. These are not always easily and consistently identified and generally not considered as ‘litter’ or ‘debris’ but as chemical pollution. Furthermore, they will be addressed through other channels. The corresponding item numbers are shown below.

Pollutants

108	109	110	111

2. The second group is the Faeces items. This is because the changes made in 2010 prevent any comparability over time. A ‘faeces’ cluster before and after 2009 would thus be totally different, and data cannot be used in any higher clustering. The corresponding item numbers are shown below.

Faeces items

121	207	208

The remaining items with truncated time series

Only 4 of the 25 truncated time series remain after the clustering (that ‘repairs’ 18 of the truncated time series) and the removal of the 3 faeces items (these also have truncated time series). The item numbers and definitions of the four 4 truncated time series that remain are shown below.

Explanation

🟧 beach litter item not on survey form in that period (and therefore not registered in survey)

Code	2002 - 2009	2010 and later
112	🟧	Plastic bag ends
114	🟧	Plastic lobster and fish tags
119	🟧	Wooden fish boxes
120	🟧	Disposable metal BBQ's

These items are mostly not found and when found generally occur only in small numbers, so it is not likely that their inclusion in the statistical analysis will lead to serious distortions.

2.4 Summary of procedure for beach litter data cleanup and preparation

The various steps of our proposed procedure for data cleanup and preparation should be performed in the following sequence.

1. Make sure the data are listed in chronological order.
2. Check if all survey dates are unique. If they are not, use one of the following solutions, dependent upon the specific situation:
   - the records are identical: remove one or two of these records, such that only one remains;
   - only one of the records has a total item count > 0: remove the record(s) with total item count = 0;
   - the records are different and each has a total item count > 0: contact the person responsible for the beach litter data of that country, to find a solution.
3. If for a certain survey date the data are only zeros and/or blanks, set them all to blanks.
4. If for a certain survey date at least one item count is > 0, set all blanks of that date to zero.
5. If the item numbers 200 - 208 and 210 are not present, create them and fill them with the data from before 2010 of the item numbers 31, 32, 46, 51, 58, 62, 84, 85, 106 and 107.
6. Set all data from before 2010 of the item numbers 31, 32, 46, 62, 84 and 112 - 121 to blanks.
7. Set all data from 2010 and later of the item numbers 51, 58, 85, 106, 107, 200 - 208 and 210 to blanks.
8. Construct the time series of the six clusters (see § 2.3.4 for their composition) and give them the cluster numbers 300 – 305.
9. Remove the time series of item numbers 51, 58, 85, 106 and 107 and of all the 20 item numbers in clustering (see § 2.3.4 for these 20 numbers).
10. Remove the four time series of the pollutants items, with item numbers 108, 109, 110 and 111.
11. Remove the three time series of the faeces items, with item numbers 121, 207 and 208.
12. Construct the time series of the beach total count[2] and give it the number 400.

[2] Some use the term 'total abundance' for this quantity. We think that this can lead to confusion, because the term 'total abundance' is an ecological concept, that refers to the (relative) representation of a species in a particular ecosystem [Wikipedia]. Furthermore, in ecology a higher total abundance

13. Construct the five time series of the categories of sources and give them the numbers 401 – 405 (see § 7.1).[3] These categories are constructed using only the individual items (also the 20 items used in clustering), because the clusters consist of items of different categories.
14. Construct the ten time series of the categories of materials and give them the numbers 406 – 415 (see § 7.2). These categories are constructed using only the individual items (also the 20 items used in clustering), because the clusters consist of items of different categories.
15. The data cleanup rules are partly based on the following premises:
    - If a beach is surveyed, at least one item count will be > 0.
    - At each survey only the items are registered that are defined in the OSPAR Marine Litter Monitoring Survey Form that is valid for that moment.

generally has a positive connotation, but more beach litter certainly is not positive. Therefore we prefer to use the term 'beach total count' in this report.

[3] It deserves notion however, that a general attribution of sources to items for the complete OSPAR region is doubtful. The only source assignments which are clear, are those of fishing.

3 Trend analysis of beach litter time series

In this chapter we examine which trend analysis methods are applicable for beach litter time series and then advise on the method to use. We consider only methods for the analysis of a monotonic trend[4], as these have more general applicability than methods for the analysis of a step trend.

3.1 Backgrounds

The procedure for the statistical analysis of beach litter data [Schulz et al., 2014] uses trend analyses for each combination of beach and assessment period of: i) the counts of each individual item and ii) the total count over all items. [Schulz et al., 2014] propose to perform the trend analyis using the Mann-Kendall trend test [Mann, 1945 and Kendall, 1938 and 1975], combined with the Theil-Sen slope estimator [Theil, 1950 and Sen, 1968].
To test this part of the procedure, [Van Loon, 2014] advises to also apply the seasonal Mann-Kendall trend test [Hirsch et al., 1982][5] and compare the p-values of the two tests, to see which results in lower p-values (higher statistical significances). And [Van Loon, 2014] also advises to examine the option to choose for one-sided testing, based on a visual inspection of the time series.

Discussion

We will use a somewhat different strategy to come to the choice of the optimal trend analysis method, for the reasons that are outlined below.

- A comparison of p-values of two trend tests applied to the same (real) time series data is insufficient to enable an objective choice of the best test. This is because the preferred trend test for a particular process[6] should meet the following two criteria:
  1. if the process is without trend, the empirical significance level of the test must not exceed the nominal significance level (in most studies this latter is set beforehand to 5%). Otherwise the trend test has a higher risk of detecting a trend when in fact there is no trend, than accepted beforehand as the testing risk (also known as the type-I risk);
  2. if the process contains a real trend, the power of the test is higher than that of other tests that comply with criterion 1.

  Thus, if we apply two trend tests to real time series, their differences in p-values do not give us enough information to choose the best test, unless we know with certainty if the time series contain trends or not. This latter is only possible if artificially simulated time series with known trends are used.

- The Mann-Kendall test and the seasonal Mann-Kendall test are not the best trend tests for time series that have autocorrelation. If, for example, autocorrelation occurs together with

[4] A time series has a monotonic trend if the average of the series (this is the arithmetic average in case of a symmetrical distribution and the median in case of a skewed distribution) generally changes in the same direction. The change can be linear, convex or concave, or a combination of these forms.
[5] It is important to specify that if a time series shows seasonal effects, the Theil-Sen slope estimator should be replaced with Kendall's seasonal slope estimator (see also § 3.2.1).
[6] A process is the generating mechanism of an ensemble of different time series that all have the same statistical characteristics (such as average, variance, normality, seasonal effects and/or autocorrelation).

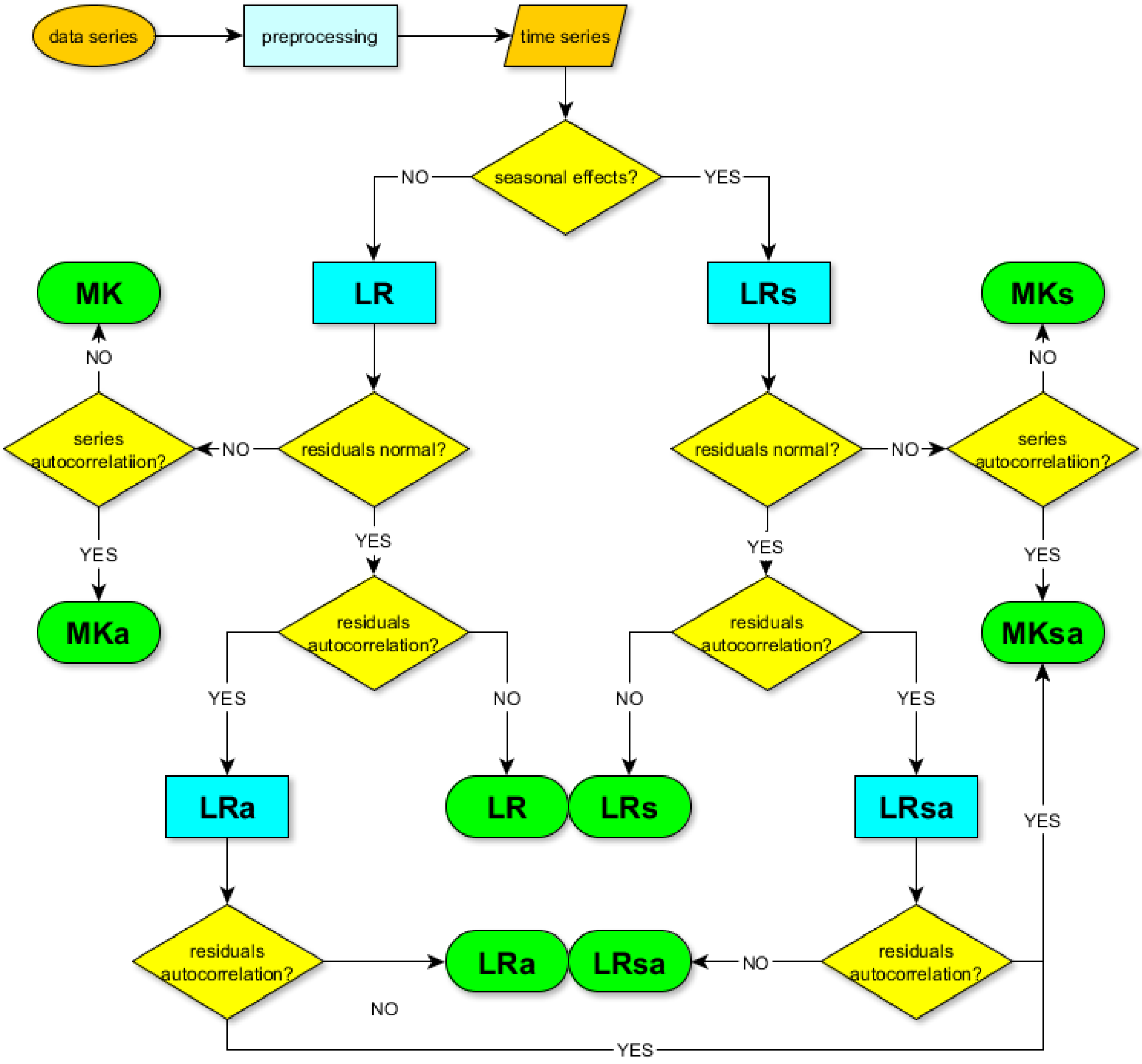
seasonal effects, than the seasonal Mann-Kendall test with correction for autocorrelation is a better choice [Hirsch and Slack, 1984]. And we do not yet know if the occurrence of autocorrelation is frequent in time series of beach litter items count.

To come to the choice of the optimal trend analysis method, in the following we will first examine which trend analysis methods are applicable for beach litter time series (§ 3.2) and then advise on the method to use (§ 3.3).

3.2 Trend analysis methods applicable for beach litter time series

By exploring a large number of beach litter time series, we determined which trend analysis methods are applicable for beach litter time series. For each series we applied a procedure to select the trend test that best fits the characteristics of that series, using various formal statistical tests to determine if its values come from a normal distribution and if it has seasonal effects and/or autocorrelation. The flowchart of this procedure is shown in figure 3.1.

Figure 3.1: Flowchart of our procedure to select the optimal trend test for a time series.



Yellow diamond: decision node | Blue rectangle: trend test | Green tube: selected trend test | LR: Linear regression test | MK: Mann-Kendall test | s: with correction for seasonal effects | a: with correction for autocorrelation

Our selection procedure can choose between eight trend tests, coming from the following two groups:

1. linear regression trend tests;
2. Mann-Kendall trend tests.

Both groups contain four trend tests, namely the original test and three adaptations, that make it possible to deal with: i) seasonality, ii) autocorrelation and iii) seasonality and autocorrelation. The procedure is also applied in our program *Trendanalist* [Baggelaar en Van der Meulen, 2012a].

The selection procedure aims at selecting the trend test that fulfills the following criteria:

1. if the process is without trend, the empirical significance level of the test does not exceed the nominal significance level;
2. if the process contains a real trend, the power of the test is higher than that of other tests that comply with criterion 1.

In the flow chart of figure 3.1 we see that the four linear regression tests are also used as intermediate steps in the selection procedure. This is because the choice for a linear regression test can only be verified by examining the residuals of the linear regression model on normality and lack of autocorrelation.

Furthermore we see in figure 3.1 that, if after applying a linear regression test that accounts for autocorrelation (LRa or LRsa) the model residuals still have autocorrelation, the procedure selects the Mann-Kendall test that corrects for seasonality and autocorrelation (MKsa). This is because it can be assumed that if LRa or LRsa do not succeed in removing the autocorrelation, there probably is a non-linear trend and MKsa is also robust in case of a non-linear trend.

3.2.1 Seasonal effects?

To check if a time series has seasonal effects, the selection procedure applies the Kruskal-Wallis test. This is the distribution-free equivalent of one-way analysis of variance. The test will be more effective if the time series has no trend, therefore the selection procedure applies it after removing a possible trend, by subtracting the value of the estimated trend line (see below) from each corresponding value of the time series.

Estimated trend line that is used for removing a possible trend

The trend line is estimated as:

$$\hat{Z}_t = b_0 + b_1 \cdot T_t$$

where b_0 (unit) is the intercept, b_{Ks} Kendall's seasonal slope (unit/year), T time (year) and t the time index (t=1, 2, ..., n), where n is the number of values in the series.

The following distribution-free estimator of the intercept is used [Conover, 1980]:

$$b_0 = \text{median}[Z_t \text{ for } 1 \le t \le n] - b_1 \cdot \text{median}[T_t \text{ for } 1 \le t \le n]$$

where Z represents the values of the variable.

And Kendall's seasonal slope [Sen, 1968; Hirsch et al., 1982] is estimated as:

$$b_{Ks} = \text{median}[\frac{Z_{tj} - Z_{kj}}{T_t - T_k} \quad \text{for all } 1 \le k < t \le n_j \quad \text{and for } j = 1,\ldots,s]$$

where t and k are values of the year index, j is the seasonal index and n_j is the number of years with a time series value in season j. First the slopes are determined for all $1 \le k < t \le n_j$ and for j=1,...s, where s is the number of time series values in a year (if the time unit is a quarter, then s

is 4). Then Kendall's seasonal slope is calculated as the median of all slopes between values that are exactly one or more years apart.
This distribution-free estimator is fairly resistant against outliers, because the median of all individual slopes is taken. Futhermore, it is not influenced by seasonal effects, because only slopes are determined between values in the same season, that are one or more years apart. Finally the estimator is unbiased – which means that it has no systematic error – and more precise than the linear regression slope for time series with seasonal effects, autocorrelation and a skewed distribution [Hirsch et al., 1982].

3.2.2 Residuals normal?

After estimating the linear regression model, the selection procedure checks if the model residuals come from a normal distribution, using the Lilliefors test on normality [Lilliefors, 1967 and 1969]. This is an adaptation of the Kolmogorov-Smirnov test on normality, for cases in which the mean and the standard deviation of the population must be estimated from the sample data.

3.2.3 Residuals autocorrelation?

If the null hypothesis that the residuals of the linear regression model come from a normal distribution is not rejected by the Lilliefors test, the selection procedure checks if the residuals show autocorrelation. This is done with the Portmanteau test [Ljung and Box, 1978], that uses the sum of the squared autocorrelation coefficients of the model residuals.

3.2.4 Series autocorrelation?

To check if the time series has autocorrelation, the selection procedure applies the runs test. This test is distribution-free. Of relevance is the autocorrelation that remains after clearing the time series from a possible trend and - if necessary - also from seasonal effects. To check on this autocorrelation a time series is processed with the following steps:

1. a possible trend is removed, by subtracting the value of the estimated trend line (see § 3.2.1) from each corresponding value of the time series;
2. if the null hypothesis that the time series has no seasonal effects is rejected by the Kruskal-Wallis test (see § 3.2.1), the median of all the trend corrected values that occur in the corresponding season is subtracted from each value.

3.2.5 Selected trend tests

For our inventory we selected time series of beach litter items that cover the period 2002 – 2013. The series were selected from all eleven OSPAR-countries, using the OSPAR-database and the original dataset of The Netherlands, both after data cleaning according to the procedure described in § 2.4.
The dataset of The Netherlands consists of four beaches and the OSPAR-dataset consists of 91 beaches of the other ten European countries, making a total of 12,520 time series of beach litter items and clusters. It is important that this inventory is as broad as possible, because it is possible that the characteristics of time series of beach litter data differ between countries.

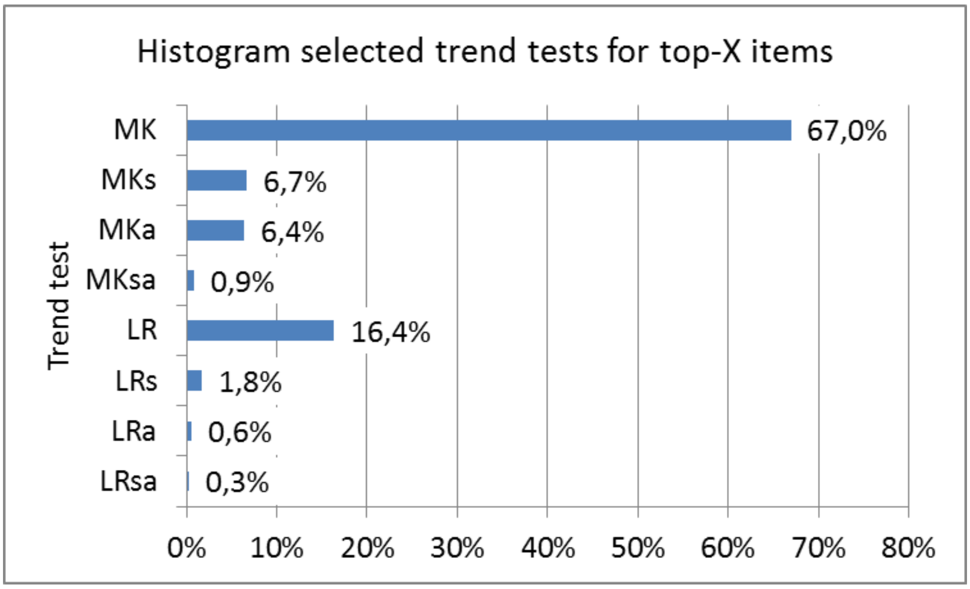
We only included time series that comply to our following criteria for trend analysis:

- the series length is at least 4.5 years (period between first and last measurement);
- the series contains at least 5 values;
- the series has at least one value in each of the three four years periods 2002 - 2005, 2006 - 2009 and 2010 – 2013 (criterion for homogeneous distribution of data in time).

The application of these selection criteria resulted in 2,835 time series of beach litter items and clusters, coming from 27 beaches of seven countries, being Belgium, Germany, Spain, The Netherlands, Sweden, Portugal and United Kingdom. The selection criteria strongly reduce the number of OSPAR beaches with data suitable for trend analysis from 95 to 27. The data from Denmark, Ireland, France and Norway do not pass these criteria. The most influential selector is the third criterion.

According to the proposed procedure for the statistical analysis of beach litter data only the top-X list of items[7] should be analyzed on trend [Schulz et al., 2014]. Therefore we made an inventory of the trend tests that were selected for the 342 top-X items of the 27 beaches that remain after applying the selection criteria. The results are shown in figure 3.2.

Figure 3.2: Histogram of trend tests that were selected for the 342 time series of top-X items of 27 beaches of seven OSPAR-countries.



In figure 3.2 we see that for the vast majority of analysed series (67,0%) the Mann-Kendall test (MK) is the best choice. For 16,4% of the series the linear regression test (LR) is the best choice. And for 16,7% of the series the best choices are adapted versions of the Mann-Kendall test or the linear regression test, that account for seasonality (s), autocorrelation (a), or both (sa).

3.3 Conclusions regarding the preferred trend analysis method

As shown in § 3.2.5, there is no trend test that is optimal for all beach litter time series. Therefore, trend analysis of such time series should preferably be tailor-made, using a procedure that selects for each individual time series the trend test (and accompanying slope estimator) that best fits the characteristics of that series, such as the selection procedure shown in figure 3.1.

[7] The top-X list is a list of the minimal number of top items – but at least 10 - that covers at least 80% of the beach total count (the total count of all items).

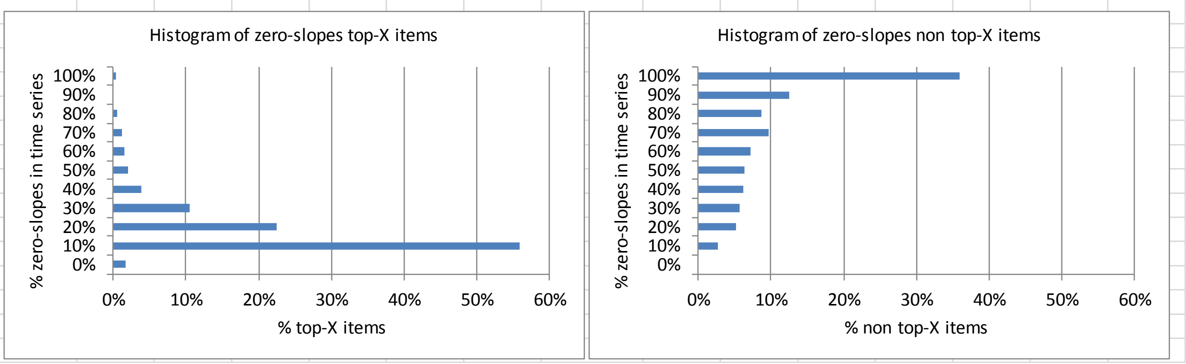
If however, the complex software and/or knowledge to apply the intricate selection procedure is not available, a sub-optimal solution is to test all time series on trend using the Mann-Kendall test (and to estimate the trends of all time series with the Theil-Sen slope estimator). In § 3.2.5 we saw that the Mann-Kendall test is the optimal trend test for 67,0% of the explored top-X items. And it will be a nearly optimal trend test for 16,4% of these series, because the use of the Mann-Kendall test instead of the linear regression test will only give a slight loss of statistical power. For example, this can be seen in the simulation results of [Baggelaar en Van der Meulen, 2012b]. Some earlier references that discuss the advantages of distribution-free tests compared to parametric tests are [Bradley, 1968; Helsel and Hirsch, 1988; Helsel and Hirsch, 1991, Hirsch et al., 1991 and Önöz and Bayazit, 2003].
Thus, probably for the vast majority of the series (> 80%) the Mann-Kendall test will be the optimal / nearly optimal test. But the other series may show seasonal effects and/or autocorrelation, which make the Mann-Kendall test a sub-optimal choice for these series. There are optimal trend tests available for such situations, but it takes complex algorithms to select such a test, based on an exploration of the statistical characteristics of the time series.

3.4 Problems with ties in time series of beach litter data

Each time series of beach litter data consists only of integer values greater than or equal to zero. In such series ties can occur. A tie is a group of equal values - you may also call them 'tied values' -, such as zeros, or ones, or tens, etc. For example, the time series 1, 0, 3, 1, 0, 1, 3, 5, 2 contains the ties (0, 0), (1, 1, 1) and (3, 3). When applying the Mann-Kendall test to a time series with ties, a correction to the test should be applied, because it is based on the signs of the slopes between pairs of values and ties cause zero slopes, that contain no information. In our exercises we used a correction for the Mann-Kendall test that accounts for ties. It can be assumed however, that this correction is less effective at high percentages, but we did not find literature on the limit of its application region.
There is also a problem that can occur when applying the Mann-Kendall test in combination with the Theil-Sen slope estimator on a series with a high percentage of ties. In certain cases the Mann-Kendall test can detect a statistical significant trend, while the Theil-Sen slope is estimated as exactly zero. This is because the test and the slope estimator are based on different algorithms. If more than 50% of the slopes between data pairs is zero (due to ties), the Theil-Sen slope estimator will be exactly zero, because it is the median of all possible slopes between data pairs. If this occurs in combination with a statistical significant trend detection of the Mann-Kendall test, it is advisable to set the estimated slope at a very small positive or negative number, deriving the sign of this number from the sign of the linear regression slope. A better solution for count series with a high percentage of zeros would be to apply a combination of trend test and slope estimator based on the Poisson distribution.

The percentage of ties is highest in the time series of items that seldomly occur, because these series contain many zeros. We see in figure 3.3 that high percentages of zero-slopes between pairs of values (indicative of ties) mainly occur in the non top-X items.

Figure 3.3: Histogram of beach litter time series with zero-slopes (indicative of ties). Left: for the 342 top-X items of the 27 explored beaches. Right: for the other 2,493 items of these beaches.



For series with high percentages of zeros (and therefore also high percentages of ties), a trend test based on the Poisson distribution will be a better choice than the Mann-Kendall test. But, as we can see in figure 3.3 left, the series of the top-X items mostly contain low to moderate percentages of ties. Thus, it is not likely that much problems will occur with the Mann-Kendall test in trend analysis of the top-X items. However, problems can occur with trend analyses of series of categories of sources or materials, because some of these have very high percentages of zeros. This can be seen in the sources evaluation matrix and the materials evaluation matrix of selected beaches in the digital appendix of this report. For these series it is advisable to apply a combination of trend test and slope estimator based on the Poisson distribution, but this goes too far for this evaluation study.

3.5 Choose for one-sided or two-sided statistical testing on trend?

In the specifications of this evaluation study, [Van Loon, 2014] advises to examine the option to choose for one-sided testing, based on a visual inspection of the time series.

Discussion

The choice for one-sided or two-sided statistical testing should not be based on a visual inspection of the data, because that will invalidate the confidence level of the test. The choice must be made beforehand, solely based on theoretical considerations. For example, in cases where the effect of a measure is evaluated, one-sided testing is appropriate. However, in most cases one will be interested in both positive and negative trends and then two-sided testing will be appropriate. Also for the statistical analysis of beach litter time series one will be interested in both positive and negative trends, therefore we advise to apply two-sided trend testing.

4 Characterising the group of estimated trends

In this chapter we describe two methods to characterise the group of trends of the top-X items of a beach, as determined for a certain assessment period. These methods use item trend indices (§ 4.1) and trend percentiles (§ 4.2).

4.1 Characterising the group of trends with item trend indices

[Schulz et al., 2014] propose to characterise the group of trends of the top-X items with an item trend index (*ITI*), that uses the trend signs of the statistically significant trends:

$$ITI = \sum_{i=1}^{m} \left(\frac{n_i}{N} \cdot \left(-1 \cdot \mathrm{sign}\left(s_i\right)\right)\right) \qquad [1]$$

where *m* is the total number of items (the top-X items plus all the other items), *i* the index of the item (i = 1, 2, .., *m*), n_i the count of item *i* in that assessment period, *N* the total count over all items in that assessment period, *s* the estimated magnitude of the trend slope and sign(*s*) the trend sign. If the trend is statistically significant, the trend sign is set to +1 for *s*>0 and to -1 for *s*<0. And if the trend is not statistically significant, or is not estimated (this is the case for all the items that are not in the top-X list), the trend sign is set to 0.

The formula [1] leads to an *ITI* that is a weighted average of the reversed trend signs of all items, where a sign is quantified as 0, +1 or -1. We propose to remove the reversal of the trend sign in formula [1], because it creates confusion if this *ITI* is presented along with the individual top-X trends, such as in the evaluation matrix (see chapter 6). The adjusted definition of this *ITI* then becomes:

$$ITI_{\text{weigthed average of trend signs}} = \sum_{i=1}^{m} \left(\frac{n_i}{N} \cdot \mathrm{sign}\left(s_i\right)\right) \qquad [2]$$

The *ITI* as defined in formula [2] is not affected by the magnitude of the estimated trend slope, which may render it insensitive in some situations. Therefore we propose to also present the following two trend indices, that make use of the magnitude of the estimated trend slope in characterising the group of trends:

$$ITI_{\text{sum of slopes}} = \sum_{i=1}^{m} s_i^{*} \qquad [3]$$

and

$$ITI_{\text{weighted average of slopes}} = \sum_{i=1}^{m} \left(\frac{n_i}{N} \cdot s_i^{*}\right) \qquad [4]$$

where *s** is the filtered magnitude of the estimated trend slope, such that $s_i^{*} = s_i$ if s_i is statistically significant and $s_i^{*} = 0$ if s_i is not statistically significant or is not estimated (this latter is the case for all the items that are not in the top-X list).

Each of these ITI's integrates the information about developments of individual items, but they differ in information content. The $ITI_{\text{weighted average of trend signs}}$ highlights the general direction of change of the items that have statistical significant trends, whereas the $ITI_{\text{sum of slopes}}$ presents the net change of these items. And the $ITI_{\text{weighted average of slopes}}$ presents the average slope, weighted by the contribution of each item to the beach total count and with all slopes set to 0 that are not statistically significant or are not estimated.

4.2 Characterising the group of trends with percentiles

In deriving the three ITI's of § 4.1, all slopes that are not statistically significant or are not estimated are set to 0. This censored approach will diminish the sensitivity for a general change – affecting many items - in one direction (improvement or deterioration). Therefore, we advise to also characterise the group of all estimated trends (slopes), regardless of their statistical significances. This can be done by presenting their statistics. We propose to present the minimum, the 25-percentile, the 50-percentile (this is the median), the 75-percentile and the maximum of this group. For example 25% of the estimated slopes is less than or equal to the 25-percentile.

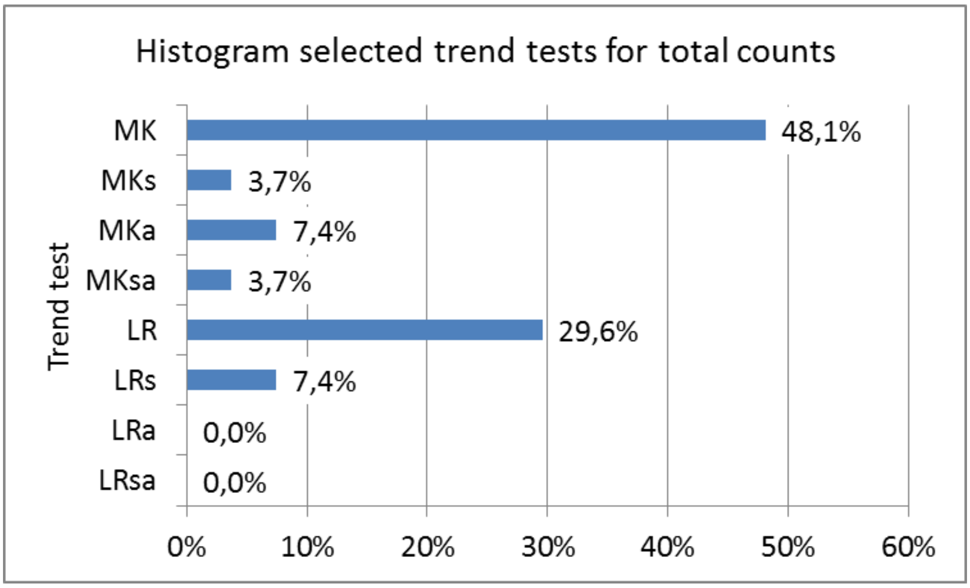
5 Detecting a change of beach total count

As requested by [Van Loon, 2014], in this chapter we propose two statistical tests to detect a change in the average of beach total count between two assessment periods, each six years in length. The first one tests on a monotonic trend (§ 5.1) and the second one on a step trend (§ 5.2).

5.1 Testing on a monotonic trend of beach total count

To enable the selection of the optimal statistical test on a monotonic trend of beach total count over a period of twelve years (two assessment periods), we first explored the statistical characteristics of time series of beach total count, covering the period 2002 – 2013. The explored series come from the 27 beaches that were selected for the exercise described in § 3.2.5. For this exploration we made an inventory of the trend tests that were selected by our procedure, that is described in § 3.2. The results are shown in the histogram of figure 5.1.

Figure 5.1: Histogram of trend tests that were selected for the time series of beach total count of 27 beaches of seven OSPAR-countries.



In figure 5.1 we see that for 48,1% of the series of beach total count the Mann-Kendall test (MK) is the best choice for testing on trend and for 29,6% of the series the linear regression test (LR) is the best choice. For the remaining 22,2% of the series the best choices are adapted versions of the Mann-Kendall test or the linear regression test, that account for seasonality (s), autocorrelation (a), or both (sa).

The results from § 5.1 indicate that non-normality is fairly common in time series of beach total count. This should be taken into account when choosing a trend test. Preferably, the choice should be tailor-made for each time series of total count, selecting a statistical test that best fits the characteristics of that series. In the case of normality the linear regression test will be the best choice, whereas in the case of non-normality the Mann-Kendall test will be the best choice. In the cases of seasonality (s), autocorrelation (a), or both (sa), the best choices will be adapted versions of these tests, that account for these characteristics.

If however, the complex software and/or knowledge to apply the intricate selection procedure is not available, we advise the same sub-optimal solution as proposed in § 3.3 for the trend analysis of the top-X items. This is to apply only the Mann-Kendall test. Figure 5.1 shows that this is the optimal statistical test for 48,1% of the explored series of beach total count. And it will be a nearly optimal trend test for 29,6% of these series, because the use of the Mann-Kendall test instead of the linear regression test will only give a slight loss of statistical power (see § 3.3). Thus, probably for the vast majority of the explored beach total series (75 - 80%) the Mann-Kendall test will be the optimal / nearly optimal test. But the other series may show seasonal effects and/or autocorrelation, which make the Mann-Kendall test a sub-optimal choice for these series. There are optimal trend tests available for such situations, but it takes complex algorithms to select such a test, based on an exploration of the statistical characteristics of the time series.

5.2 Testing on a step trend of beach total count

A test on a step trend will be more appropriate than a test on a monotonic trend, if it can be assumed that a change in the average of beach total count occurs in the form of a more or less instantaneous step. However, this test requires that it is known beforehand when the time of the change point occurs. The time of the change point should not be derived from a visual inspection of the series, because that will make the risk of falsely detecting a trend (also known as the type-I risk) substantially higher than the nominal risk (which usually is set to 5%).

In the case of normality the t-test will be the best choice of step trend test, whereas in the case of non-normality the Wilcoxon-ranksum test (the distribution-free equivalent of the t-test) will be the best choice [Helsel and Hirsch, 1991]. It is equivalent to the Mann-Whitney test, which is also a distribution-free test.
In the cases of seasonality (s), autocorrelation (a), or both (sa), the best choices will be adapted versions of the t-test and the Wilcoxon-ranksum test, that account for these characteristics. It is not easy however, to determine the characteristics of each series. A feasible sub-optimal solution is to test all time series of total count on a difference in the average between two assessment periods using the Wilcoxon-ranksum test.

How to define the average of beach total count over an assessment period

[Van Loon, personal communication] defines the average of beach total count over an assessment period of six years as the arithmetic average of the yearly beach total count. We can write this formally as:

$$\overline{T} = \frac{\sum_{j=1}^{6}\sum_{k=1}^{4} T_{jk}}{n_y} \qquad [5]$$

where $\overline{T}$ is the arithmetic average of beach total count over the assessment period, *j* is the year index (*j* = 1, 2, ..., 6), *k* is the survey index (*k* = 1, 2, 3, 4), T_{jk} is the total count in survey *k* of year *j* and n_y is the number of years in the assessment period with at least one survey.
A serious problem with formula [5] is that it is not robust against missing data. If missing data occur – a situation that is fairly common in time series of beach liter data - they will lead to a biased estimate of the average of the yearly total count using formula [5]. Possible solutions to this problem are: i) estimate the missing values, or ii) ignore data, such that each yearly total is

derived from the same number of surveys. However, both solutions require complex algorithms. But perhaps an even greater disadvantage of the definition is that it implicitly assumes that the total count comes from a normal distribution.[8] In § 5.1 we saw that this assumption will probably be invalid in about 65% of the cases. Therefore we propose to define the average of beach total count over an assessment period as the median of the beach total count per survey:

$$T_{0,5} = \text{median}\left\lfloor T_{jk}, \text{ with } j = 1,\ 2,\ \ldots,\ 6 \text{ and } k\ =\ 1,\ 2,\ 3,\ 4 \right\rfloor \qquad [6]$$

This definition is robust against missing data and it takes into account that most time series of beach total count do not come from a normal distribution.

And the relative change in the average of beach total count from assessment period *A* to assessment period *B* can be quantified as:

$$\%\text{change} = 100\% \cdot \frac{D_{BA}}{T_{0,5;A}} \qquad [7]$$

where D_{BA} is the Hodges-Lehmann estimator [Hodges and Lehmann, 1963] of the difference between the two periods. It is the median of all possible pairwise differences between data of period B and data of period A. If period B has data from *n* surveys and period A data from *m* surveys, then there are *n·m* pairwise differences. This is a more precise estimator of the difference between the population medians of the two periods than the difference between the sample medians of the two periods [Helsel and Hirsch, 1991].

[8] The use of the arithmetic average is only advisable if the data come from a normal distribution. In cases of non-normality the median is a better choice to represent the central location of the population.

6 Presenting the results: the evaluation matrix

To enable a concise presentation of the results of the various statistical analyses, we developed an evaluation matrix. It can present for each separate beach four groups of results. In § 6.1 we describe the metadata of the evaluation matrix. In § 6.2 - 6.5 we describe the details of the four groups and indicate what statistical conclusions can be drawn from them. And in § 6.6 we describe the differences between the items evaluation matrix and the evaluation matrices of sources and materials.

6.1 Metadata

The metadata of the evaluation matrix are the name of the country, the name of the beach and the analyzed period. This latter is 2002 – 2007, 2008 – 2013, or 2002 – 2013.

6.2 Group 1: Summary of results from top-X items

First subgroup

Code and definition of each item of the top-X group, for the two six years periods and for the twelve years period.

Second subgroup

Descriptive statistics of the top-X items, for the two six years periods and for the twelve years period. This subgroup presents for each top-X item the median, the arithmetic average, the standard deviation (all in count/survey) and the coefficient of variation (the ratio of standard deviation and average). Also presented is the relative contribution of each item to the beach total count over that period. The top-X items are sorted from highest to lowest contribution to the beach total count.

The median is a better measure of central location than the arithmetic average in the case of a non-symmetrical distribution. Because non-symmetrical distributions are predominant for count of beach litter items, it is advisable to use the median instead of the arithmetic average for evaluations.

Third subgroup

If the number and temporal distribution of the survey data fulfill the criteria for trend analysis (see below), this subgroup shows the results of the trend analysis of the top-X items for a six years period or for a twelve years period, otherwise the corresponding cells are left empty. The following criteria for trend analysis are used:

- the series length is at least 4.5 years (period between first and last measurement);
- the series contains at least 5 values;
- there is a more or less homogeneous distribution of the data in time:
    - for a six years period: the series has at least one value in each of the three two years blocks of the six years period;
    - for a twelve years period: the series has at least one value in each of the three four years blocks (2002 - 2005, 2006 - 2009 and 2010 – 2013).

If these criteria are fulfilled the magnitudes (slopes) and statistical significances (p-values) of the estimated trends of the top-X items are shown, otherwise the corresponding cells are left empty.
The trend magnitude (slope, expressed in count/year) is estimated with the Theil-Sen estimator and the p-value is the result of testing on a monotonic trend, using the Mann-Kendall test.
The p-value is the two sided probability of observing this trend or a larger trend, if the null hypothesis of no trend is true. We consider it acceptable to use a confidence level of 95% for the testing. Thus, if the p-value is less than 0.05, than we can say with 95% confidence that there is a statistically significant trend.
If an estimated trend magnitude (slope) is negative, that cell is green and if it is positive, the cell is orange. If the p-value is less than 0.05, indicating a statistically significant trend, that matrix cell is grey.

In evaluating the trend results one should bear in mind that the overall risk of detecting one or more statistically significant trends when in reality no trends exist, increases with the number of time series that is analyzed on trend. If only one time series is analyzed on trend this risk is 5%, because we test with 95% confidence. However, if two time series are analyzed on trend this risk increases to 9,8%, according to the following formula:

$$\text{Overall risc} = 1 - (95\%)^{n}$$

where n is the number of time series that is analyzed on trend. For 10 time series the overall risk is 40,1%, for 15 time series it is 53,7% and for 20 time series it is 64,2%. If for example the top-X list contains 15 items, then the risk of detecting one or more statistically significant trends when in reality no trends exist is 53,7%. Therefore, one should be somewhat cautious with the interpretation of the statistically significant trends.

Fourth subgroup
The fourth subgroup shows the number of items in the top-X list and the number of surveys in the analyzed period, for the two six years periods and for the twelve years period.

6.3 Group 2: Summary of statistically significant trends

First subgroup
Number of statistically significant trends (slopes).

Second subgroup
The three versions of the ITI (Item Trend Index), for the two six years periods and for the twelve years period. The ITI's are presented in the sequence $ITI_{\text{sum of slopes}}$, $ITI_{\text{weighted average of slopes}}$ and $ITI_{\text{weighted average of trend signs}}$. See § 4.1 for their definitions.
Each of these ITI's integrates the information about statistically significant developments of individual items.

It is important to realise that in deriving these ITI's all slopes that are not statistically significant or are not estimated (this latter is the case for all the items that are not in the top-X list) are set to 0. We refer to group 4 for an uncensored presentation of the characteristics of the slope estimates (regardless of their statistical significances), because that presentation can be more sensitive to a general tendency of change in one direction (improvement or deterioration).

6.4 Group 3: Summary of results from beach total count

First subgroup

Descriptive statistics of beach total count, for the two six years periods and for the twelve years period. Presented are the median, the arithmetic average, the standard deviation (all in count/survey) and the coefficient of variation (the ratio of standard deviation and average).

Second subgroup

If the number and temporal distribution of the survey data fulfill the criteria for trend analysis (see § 6.2), this subgroup shows the results of the trend analysis of the beach total count for a six years period or for a twelve years period, otherwise the corresponding cells are left empty.
If the criteria are fulfilled, the magnitude (slope) and statistical significance (p-value) of the estimated trend of the beach total count are presented, for the two six years periods and for the twelve years period. The trend magnitude (slope, expressed in count/year) is estimated with the Theil-Sen estimator and the p-value is the result of testing on a monotonic trend, using the Mann-Kendall test.
If an estimated trend magnitude (slope) is negative, that cell is green and if it is positive, the cell is orange.
If the p-value is less than 0.05, indicating a statistically significant trend, that matrix cell is grey.

Third subgroup

The third subgroup is only presented for the twelve years period. It shows the magnitude (step) and the percentage change of the estimated step trend of the beach total count, by comparing the second six years period with the first. The trend (step, expressed in count/survey) and the percentage of change is estimated with the Hodges-Lehmann estimator (see § 5.2). Only if for both six years periods the survey data fulfill the criteria for trend analysis (see § 6.2), the statistical significance (p-value) of the estimated step trend of the beach total count is shown, otherwise the corresponding cell is left empty. This p-value is the result of testing on a step trend, using the Wilcoxon-ranksum test.
If an estimated trend or change is negative, that cell is green and if it is positive, the cell is orange.
If the p-value is less than 0.05, indicating a statistically significant trend, that matrix cell is grey.

6.5 Group 4: Summary of trends (slopes) of top-X items

This group summarizes the results of all trends (slopes) of the top-X items, regardless of their statistical significances. This summary can be more sensitive to a general change in one direction (improvement or deterioration) than the ITI's, because these latter are only based on censored information (the statistically non-signifcant trends are set to zero).

First subgroup

Percentage of negative slopes (cell is green if the percentage is not zero) and the percentage of positive slopes (cell is orange if the percentage is not zero), for the two six years periods and for the twelve years period.

Second subgroup

Statistics of the estimated slopes, for the two six years periods and for the twelve years period. Presented are the minimum, the 25-, 50- and 75-percentile and the maximum. If a slope statistic is negative its cell is green and if it is positive its cell is orange.

6.6 Evaluation matrix of items versus those of sources and materials

The evaluation matrices of sources and materials are somewhat different from the items evaluation matrix. This is because no top-X list is used, but all five categories of sources or ten categories of materials are used instead. For completeness the evaluation matrices of sources and materials also present the summary of results of the beach total count (group 3). However, this summary is exactly the same for all three matrices, because the beach total count is the same at the levels of items, sources and materials.

7 Sources and materials analysis

7.1 Sources analysis

For the sources analysis, we used a slight adaptation of the item source classification list of [Van Franeker, 2013], as proposed by Willem van Loon, the supervisor of this study. However, it deserves notion that a general attribution of sources to items for the complete OSPAR region is doubtful. The only source assignments which are clear, are those of fishing.

Five categories of sources are distinguished, with the following identifcation codes:

- 401: Sanitation
- 402: Fishing
- 403: Tourism
- 404: Shipping
- 405: Other

These categories are constructed using only the individual items (also the 20 items used in clustering), because the clusters consist of items of different categories.

The results of the statistical analysis at the sources level are shown in the sources evaluation matrix (see the digital appendix 2).

7.2 Materials analysis

For the materials analysis, we used the item material classification list of the OSPAR database.

Ten categories of materials are distinguished, with the following identifcation codes:

- 406: Plastic/polystyrene
- 407: Rubber
- 408: Cloth/textile
- 409: Paper/cardboard
- 410: Wood
- 411: Metal
- 412: Glass
- 413: Ceramic/pottery
- 414: Sanitary
- 415: Medical

These categories are constructed using only the individual items (also the 20 items used in clustering), because the clusters consist of items of different categories.

The results of the statistical analysis at the materials level are shown in the materials evaluation matrix (see the digital appendix 2).

8 Aggregation of results to larger spatial scales?

In the specifications of this evaluation study, [Van Loon, 2014] advises to develop (preferably simple) methods to aggregate beach results for each country and region and to apply these methods to the data and compare and evaluate them.

8.1 Discussion

We think that the applied selection strategy of beaches (or the lack of it) does not allow for a meaningful extrapolation of beach specific results to larger spatial scales. Such an extrapolation is only meaningful if the beaches were selected using some form of probabilistic sampling, but that was not the case.

Of course, it is very tempting to extrapolate beach results to larger spatial scales, because that can help to underpin important decisions. But if that is the main purpose of this monitoring system, the selection strategy of the beaches should be adapted to this information need. We should not stretch beyond the possibilities of the present selection strategy to fulfill our needs.

Some possibilities to aggregate beach results to larger spatial scales are described in [Schulz et al, 2014]. We propose to apply those methods to aggregate results from various beaches, but to present them as being strictly only representative for that specific group of sampled beaches. They should not be presented as representative for some regional or national population of beaches. Care should be taken in using these results for the selection of measures or targets.

9 Summary of proposed procedure

Based upon our evaluation of the procedure of statistical analysis of beach litter data developed by [Schulz et al,, 2014], we indicated some possibilities for fine-tuning (see this report). After integration of these possibilities the steps of the procedure become as described below.

1. Import the required data from the OSPAR database, or (for the Netherlands) from the SDN-spreadsheet.
2. Apply the procedure for data cleanup (see § 2.4).
3. Compute for each beach the top-X list of items.
4. Compute the statistics of the counts of all top-X items, the beach total counts, the five categories of sources and the ten categories of materials. These statistics are the median, the arithmetic average, the standard deviation (all in count/survey) and the coefficient of variation (the ratio of standard deviation and average). Also presented is the relative contribution of each item or category to the beach total count over that period.
5. Analyse the time series of all top-X items, the beach total count, the five categories of sources and the ten categories of materials on a monotonic trend (applying two-sided testing, with 95% confidence), if the following criteria are fulfilled:
   - the series length is at least 4.5 years (period between first and last measurement);
   - the series contains at least 5 values;
   - there is a more or less homogeneous distribution of the data in time:
     - for a six years period: the series has at least one value in each of the three two years blocks of the six years period;
     - for a twelve years period: the series has at least one value in each of the three four years blocks (2002 - 2005, 2006 - 2009 and 2010 – 2013).

   Preferably this trend analysis is tailor-made, using a procedure that selects for each individual time series the trend test (and accompanying slope estimator) that best fits the characteristics of that series. A practical sub-optimal solution is to test all time series on trend using the Mann-Kendall test and to estimate the trends of all time series with the Theil-Sen slope estimator.
6. Compute the three trendindices (see § 4.1), using only the slopes that are statistically significant, for each of the following groups: 1) the top-X items, 2) the five categories of sources and 3) the ten categories of materials.
7. Compute the minimum, the 25-, 50- and 75-percentile and the maximum of all the estimated slopes, regardless of their statistical significances, for each of the following groups: 1) the top-X items, 2) the five categories of sources and 3) the ten categories of materials.
8. Estimate the difference in beach total count between the two six years periods (and the percentage change) with the Hodges-Lehmann estimator (see § 5.2).
9. Analyse the time series of beach total count over the twelve years period on a step trend between the two six years periods. Preferably this trend analysis is tailor-made, using a procedure that selects for each individual time series the trend test that best fits the characteristics of that series. A practical sub-optimal solution is to test on a step trend using the Wilcoxon-ranksum test.
10. Present all the results in an item evaluation matrix, a sources evaluation matrix and a materials evaluation matrix (see Chapter 7).

References

- Baggelaar, P.K. en Van der Meulen, E.C.J. (2012a): *Handleiding Trendanalist*. Icastat-AMO, februari 2012, 43 pp.

- Baggelaar, P.K. en Van der Meulen, E.C.J. (2012b): *Trendanalyse op maat voor een meetnet waterkwaliteit*. Stromingen 18 (2012), nummer 2, page 77 - 96.

- Bradley, J.V. (1968): *Distribution-Free Statistical Tests*. Prentice Hall, Englewood Cliffs.

- Conover, W.J. (1980): *Practical nonparametric statistics*. John Wiley, New York.

- Guidance (2013). *Guidance on Monitoring of Marine Litter in European Seas*. EU JRC .

- Helsel, D.R. and Hirsch, R.M. (1988): *Discussion of paper by R.H. Montgomery and J.C. Loftis: Applicability of the t-test for detecting trends in water quality variables*. Water Resources Bulletin, vol 24, page 201-204.

- Helsel, D.R. and Hirsch, R.M. (1991): *Statistical Methods in Water Resources*. Studies in Environmental Science 49. Elsevier, Amsterdam, 510 pp.

- Hirsch, R.M., Slack, J.R. and Smith, R.A. (1982): *Techniques of trend analysis for monthly water quality data*. Water Resources Research, vol. 18, no. 1, February 1982, page 107 – 121.

- Hirsch, R.M. and Slack, J.R. (1984): *A nonparametric trend test for seasonal data with serial dependence*. Water Resources Research, vol. 20, no. 6, page 727 - 732.

- Hirsch, R.M., Alexander, R.B. and Smith, R.A. (1991): *Selection of methods for the detection and estimation of trends in water quality*. Water Resources Research, vol. 27, no. 5, May 1991, page 803 – 813.

- Hodges, J.L. (Jr.) and Lehmann, E.L. (1963): *Estimates of location based on rank tests*. Annals Mathematical Statistics 34, page 598-611.

- Kendall, M.G. (1938): *A new measure of rank correlation*. Biometrika, 30, 1938, page 81 - 93.

- Kendall, M.G. (1975): *Rank Correlation Methods*. Charles Griffin, London, 1975.

- Lilliefors, H.W. (1967): On the Kolmogorov-Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association, 62, page 399-402.

- Lilliefors, H.W. (1969): *Correction to the paper: On the Kolmogorov-Smirnov test for normality with mean and variance unknown.* Journal of the American Statistical Association, 64, page1702.

- Ljung, G.M and Box, G.E.P (1978): On a measure of lack of fit in time series models. Biometrika, 65, page 297-303.

- Mann, H.B. (1945): *Non-parametric tests against trend*. Econometrica 13, page 245 - 259, 1945.

- Önöz, B. and Bayazit, M. (2003): *The power of statistical tests for trend detection*. Turkish J. Eng. Env. Sci., 27, page 247 – 251.

- Schulz, M., Fleet, D., Van Loon, W. and Oosterbaan, L. (2014). *Joint proposal for a harmonized OSPAR beach litter assessment method*. Version 6 March 2014, 20 pp.

- Sen, P.K. (1968): *Estimates of the regression coefficient based on Kendall's tau*. Journ. Am. Statist. Assoc., 63, page 1379 - 1389, 1968.

- Theil, H. (1950): *A rank-invariant method of linear and polynomial regression analysis, 1,2 and 3*. Ned. Akad. Wetensch. Proc., 53, page 386 - 392, 521 - 525 and 1397 - 1412.

- Van Franeker, J.A. (2013). *Survey of methods and data analyses in the Netherlands OSPAR Beach Litter Monitoring program*. IMARES, unpublished report, Texel, June 2013, 35 pp.

- Van Loon, W. (2014): *Specifications beach litter analysis*. Version 2 April 2014, 3 pp.

Appendix 1: Initial specifications of this evaluation study

Specifications Beach Litter Analysis

Willem van Loon
Checked with Marcus Schulz
2 april 2014

These specifications describe the project description for the Dutch beach litter analysis project in 2014 and are harmonized with the German project.

Nr	Description step
	General specifications
G1	The beach litter data analysis method to be tested is based on the document:
G2	The working plans for the German and Dutch testing projects will be harmonized. Additional agreements and actions will be added into a additional joint document.
G3	It is expected that when the BLA method has been finalized (in view of the substantial number of analysis steps) a user friendly analysis tool, e.g. as an R package, could become very useful for OSPAR countries.
Gx	References and data sources used: • M. Schulz, D. Fleet, Germany, W. van Loon, L. Oosterbaan, 2014. Joint proposal for a harmonized OSPAR beach litter assessment method., version 6 march 2014. • J.A. van Franeker, 2013. Survey of methods and data analyses in the Netherlands OSPAR Beach Litter Monitoring program. IMARES, unpublished report, June 2013. • OSPAR litter database # • EU JRC, 2013. Guidance on Monitoring of Marine Litter in European Seas.
	Data collection and preparation
D1	**Beach litter data** The Dutch beach litter dataset, period 2002-2013, which will be extracted from the OSPAR database, will be used. **Action:** Prior to this, the consistency of the Dutch source data and the OSPAR database will be checked. The OSPAR dataset must be exported into Excel2007 or earlier format for analysis in Mystat. In addition to the Dutch data, it is proposed to select in addition two UK beaches, two Swedish beaches and two Spanish beaches. This will give a sufficiently broad data basis for the Dutch testing project. The choice of beaches will be based on the best quality data sets available, and based upon advice of the UK and Spanish collegues. These datasets will be extracted from the OSPAR database.
D3	**Item clustering** • Due to the change of some item codes and names around 2010, it is necessary for trend analysis to cluster these corresponding items into a single item group. This item clustering is a point of attention for Willem and Marcus. For example, should the sources of the items clustering be the same, or is this not necessary ? This item clustering has been specified in Van Franeker (2013) but will be reviewed with respect to source attribution etc.. • The clustering of plastic fragments must be well defined and will be discussed with Marcus

	Data analysis
A1	**Assessment periods** For the MSFD, assessment period of 6 years will be used. **Action Willem/Marcus/Lex/David:** define the assessment periods. E.g.: baseline period: 2001-2006; ; assessment period 1: 2007-2011; assessment period 2: 2012-2016.For NL, 2017 is the first interim MSFD reporting year.
A2	**Check the choice of beach specific analysis** A major choice in the method which has in principle been made is to use beach specific analysis, in view of the observed heterogeneity of beaches. It is proposed to check this choice by comparing the trend analysis results (primarily p-values) for the top-X items, and total abundance, for the 4 Dutch beaches separately.
A3	**Calculate the top-X list per beach per assessment period** For every assessment period of 6 years, the average top-X list must be calculated per beach; and will be used as (a) information as is and (b) their fractions will be used as weight factors for the item trend index.
A4	**Constitute a selected item list per beach** Based on the top-X list, a selection of items is made which contains at least 80% of the total abundance. This 80% basis gives a sufficiently objective starting point for the item trend analysis. In addition, items of special interest (e.g. with ecological risks) may be added freely to this list.
A7	**Calculate trends per beach for the selected item list & total abundance** • Calculate for each beach the trends for the item assessment list & total abundance. • Use the Mann Kendall trend analysis method with Sen-Teil slope estimation. • Test if the use of Mann Kendall *seasonal* gives higher trend significanties. To use this option, a grouping variable (1 to 4) has to be assigned to each data pair. • For blind data analysis, for Mann Kendall the two-sided test has to be chosen. After visual inspection a one-sided test may be possible, but this is probably not a user-friendly procedure.
A8	**Calculate a item trend index per beach (6 and 12 year periods)** * calculate this trend index with and without plastic fragments * use all the significant trend results which were found in the selected item list.
A8	**Calculate the change of the average total abundance for the MSFD periods per beach** * For each beach, and MSFD period of 6 years, the average total abundance is calculated. * The percentage decrease or increase of the total abundance per MSFD period per beach is reported. These average total abundances per period will primarily be calculated as arithmetic averages. In addition, the use of the Senn-Teil slope will be investigated for this purpose (useful for non-significant trends?) * A non-parametric test (e.g. the Kruskal-Wallis-H-test will be used to test the significance of the difference between the total abundance averages of the two periods compared.
A9	**Make an evaluation matrix and aggregate results per country and region** • make a table which reports per beach and MSFD period (and 2 periods combined) item trend index and change (percentage + significance?) of total abundance. • Design and test possible (preferably simple) methods to aggregate beach results per country and region.

A10	**Perform source and material analyses** • Per beach, perform analyses of the total abundance summed for the item sources (a) shipping, (b) fisheries, (c) tourism, (d) sanitation (see OSPAR database classification; UK comment)and (e) unreliable classification. • The use of the item source classification list of the OSPAR database is proposed by Marcus to be used. This OSPAR list will be checked and compared with the list of Van Franeker (2013) and the source classification list in the Guidance. Design and test methods to aggregate the source analysis results per beach into the country and regional level. • Per beach, perform analyses of the total abundance of plastics and other materials (more classification necessary ?) (see Guidance, Annex 8.1 - Master List of Categories of Litter Items).
	Reporting

Appendix 2 (digital): Evaluation matrices of selected beaches

See the digital appendix of this report.