sbDatastats

Documentation

Table of Contents

What is it for?

Wouldn’t it be nice to have a data analyser which shows you numerical and string outliers without much effort?

Well, this has been the design idea and the implementation approach of the Excel VBA sbDatastats application which this document describes.

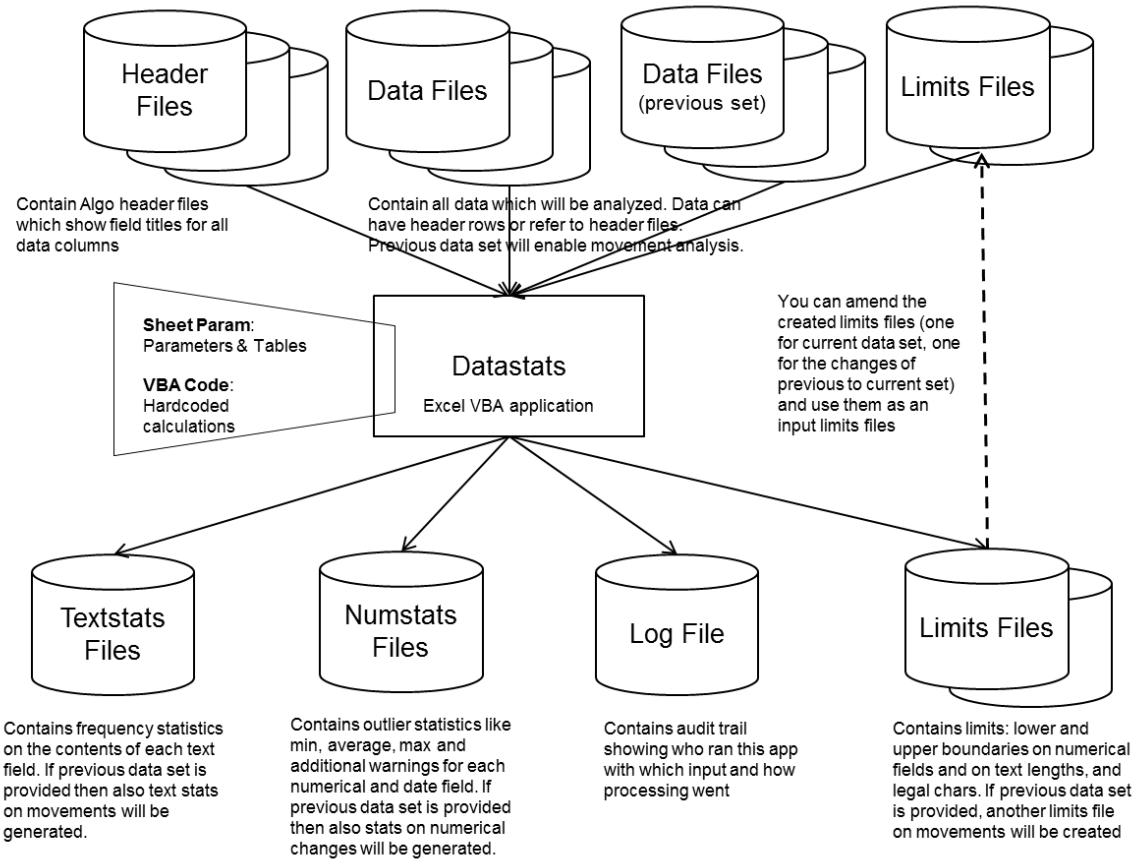
You copy all your data files into one folder, all header files into another one, and you define a log folder. Then you just press a button in the Workflow tab, and the program will deliver outlier statistics for all numerical and date fields (showing min, average, max plus some additional warnings) as well as for all text fields (showing a frequency statistic for all contents).

You, that can be any team member in Data Service, in IT or Operations, in any business unit or even internal or external audit (for a thorough IT data audit).

System Handbook

This part describes how the application needs to be installed and maintained. It requires basic MS Windows and MS Excel knowledge only.

Overview



The installation is pretty straight forward. First create a new folder and copy the Excel VBA sbDatastats application into it, then generate subfolders for data (Data_Files and Data_Files_Prev), Algo One headers (Header_Files), logs (Logs), configuration (Config), and output (Datastats) files, finally initialize application parameters in sheet Param of Excel VBA application.

Application design is that the user should only need to copy new data files into the data subfolder(s)[1] and to amend, rename, and move (to folder Config) the limits file(s) in order to direct the application on which border values to act.

[1] If you also provide a previous data set folder with corresponding data then this application will produce a movement analysis as well.

Parameters in Sheet Param

Log_Level
Usually 1 to get all log outputs including errors, warnings and simple information. If set to 2 only errors and warnings will be logged but then you will lose your audit trail because filenames in use and checking information will no longer be printed.

Warning Threshold (# of Stdevs)
Contains the number of standard deviations from which on an extreme value will be warned about in the Warning column of Numstats output. A standard value of 3 is suggested to be used.

Keep stats in tabs
The sbDatastats spreadsheet contains four tabs (sheets) called Numstats, Textstats, LimitsIn, and Limitsout which normally show the run results which will also be stored in output files Numstats_yyyymmdd.csv, Textstats_yyyymmdd.csv, Input_Limits_File, and Output_Limits_File in subfolder Datastats. For frequent runs it is very convenient to have the output in the application spreadsheet for the ease of lookup. In this case set this value to True. If you strictly need to separate the data from the program, then set this to False, and these tabs get cleared at program end (usually necessary if run by audit).

Max number of string attributes per field
This parameter drives how many different fields are maximally accepted in the Attributes field of the limits input file.

No single character check
If this parameter is set to ‘True’ we do not check whether single characters are valid.

Input File Delimiter
This parameter is defining the field separation character for input files. This would in general be ‘,’ for comma separated files (CSV), but you can define any character, for example ‘|’ for pipe separated files ‘|’ (PSV). Config files and limit files for the Datastats application are all CSV.

User Handbook

This is a description of how to run this application.

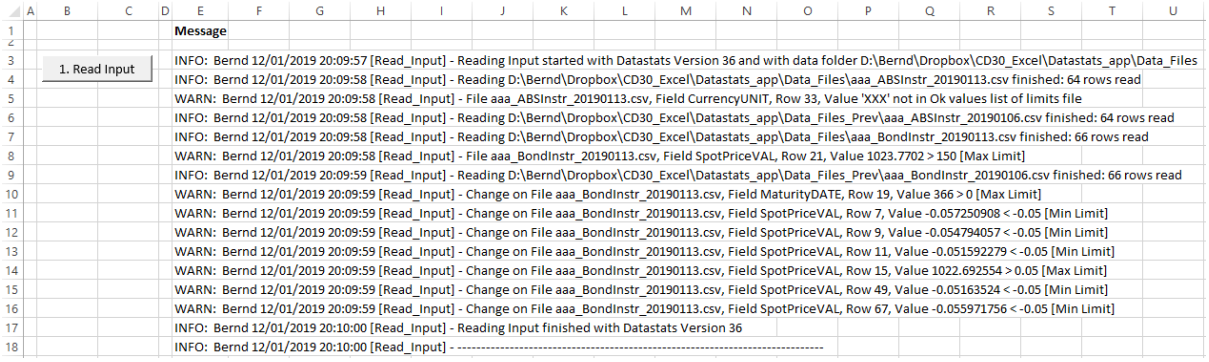
Summary

Short instruction is: Just copy your data files into subfolder Data_Files (and previous period's ones into subfolder Data_Files_Prev). For Algo data files you need to copy corresponding header files into subfolder Header_Files, for all other data files you need to specify one or more sort columns in files FileSPecs.csv in subfolder Config. Then press button '1. Read Input' in sheet Workflow.

Now wait for the final message 'INFO: …. 16/08/2015 17:06:14 [Read_Input] - Reading Input finished'.

Errors, warnings and other useful information will pop up during execution in sheet Workflow.

Example:



But they will also be written into the logfile (file sbDatastats_Logfile_YYYYMMDD.csv in subfolder Logs) for audit purposes.

The output files of this application are in subfolder Datastats.

Config File FileSpecs.csv

This file resides in subfolder Config and contains necessary information for the sbDatastats application to perform movement checks on normal (i. e. non-Algo One) input files if a previous data set is provided:

	A	B	C	D
1	Filename	SortColumn1	SortColumn2	SortColumn3
2	Input	7		
3	More_Input	7		

Up to three sort columns tell the application on which column(s) to sort the data sets so that the program can identify deleted, new or changed (can even be identical) data records. Please note that column A just needs to contain the 'core' filename. The program will automatically detect whether this core FILENAME is a prefix or a suffix or anywhere in the middle of the actual filenames which can read 'FILENAME_20171004.csv' or '20171004_FILENAME.psv', for example.

Numstats Output

With the outlier statistic on each numerical and date field you can easily spot potential errors:

	A	B	C	D	E	F	G	H	I	J	K
1	Filename	RowType	Fieldname	Warning	Min 1	Min 2	Min 3	Average	Max 3	Max 2	Max 1
2	aaa_BondInstr_20190113.csv	BAS:DM Bond FutureSPEC	MaturityDATE		26-Nov-2030	11-Jan-2031	30-May-2031	27-Feb-2035	19-Jun-2039	19-Feb-2040	04-Sep-2040
3	aaa_BondInstr_20190113.csv	BAS:DM Bond FutureSPEC	SpotPriceVAL	Max is off by more than 3 stdev.	0.97	0.99	0.99	32.13	1.25	1.29	1,023.77

The red circle highlights a coupon rate which is off by a factor of 100.

Numstats Move Output

With the outlier statistic on ***changes*** of numerical and date fields you also spot potential errors:

	A	B	C	D	E	F	G	H	I	J	K
1	Filename	RowType	Fieldname	Warning	Min 1	Min 2	Min 3	Average	Max 3	Max 2	Max 1
2	aaa_BondInstr_20190113.csv	BAS:DM Bond FutureSPEC	MaturityDATE	Max is off by more than 3 stdev.	-	-	-	11.09	-	-	366.00
3	aaa_BondInstr_20190113.csv	BAS:DM Bond FutureSPEC	SpotPriceVAL	Max is off by more than 3 stdev.	(0.06)	(0.06)	(0.05)	30.98	0.04	0.04	1,022.69

In the case above you might want to check the change of the maturity date (which normally should not change) and obviously the price change is in error.

Textstats Output

With the frequency statistic on each text field you can easily verify whether there are suspicious string values:

	A	B	C	D	E	F	G	H	I	J
1	Filename	RowType	Fieldname	Warning	Char	Char Frequency	Text Length	Frequency	Text	Count
2	aaa_ABSInstr_20190113.csv	BAS:DM Bond FutureSPEC	ContractSizeVAL		CHAR(067) = 'C'	32	15	32	'ContractSizeVAL'	32
3	aaa_ABSInstr_20190113.csv	BAS:DM Bond FutureSPEC	CurrencyUNIT		CHAR(088) = 'X'	3	3	32	'GBP'	14
4									'EUR'	11
5									'USD'	6
6									'XXX'	1

This example show the erroneous currency ‘XXX’. Infrequent string values should generally be reviewed.

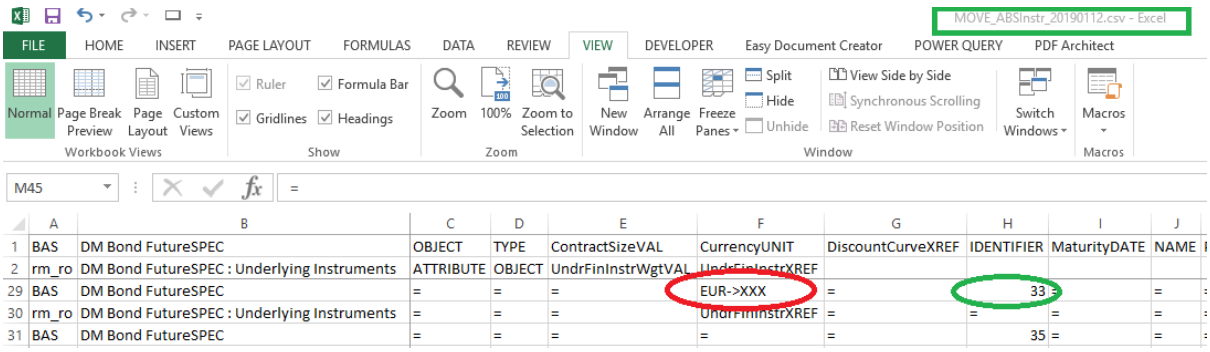
Textstats Move Output

With the frequency statistic on the ***change*** of each text field in case you provided a previous data set you can also verify whether there are strange changes:

	A	B	C	D	E	F	G	H	I	J
1	Filename	RowType	Fieldname	Warning	Char	Char Frequency	Text Length	Frequency	Text	Count
2	aaa_ABSInstr_20190113.csv	BAS:DM Bond FutureSPEC	ContractSizeVAL				1	32	'='	32
3	aaa_ABSInstr_20190113.csv	BAS:DM Bond FutureSPEC	CurrencyUNIT				1	31	'='	31
4							8	1	'EUR->XXX'	1
5	aaa_ABSInstr_20190113.csv	BAS:DM Bond FutureSPEC	DiscountCurveXREF				1	32	'='	32

This example shows the erroneous change in the currency field from ‘EUR’ to ‘XXX’.

If you want to identify the offending data record you need to look into the corresponding MOVE file for the file given in column A:



	A	B	C	D	E	F	G	H	I	J
1	BAS	DM Bond FutureSPEC	OBJECT	TYPE	ContractSizeVAL	CurrencyUNIT	DiscountCurveXREF	IDENTIFIER	MaturityDATE	NAME
2	rm_ro	DM Bond FutureSPEC : Underlying Instruments	ATTRIBUTE	OBJECT	UndrFinInstrWgtVAL	UndrFinInstrXREF				
29	BAS	DM Bond FutureSPEC	=	=	=	EUR->XXX	=	33	=	=
30	rm_ro	DM Bond FutureSPEC : Underlying Instruments	=	=	=	UndrFinInstrXREF	=	=	=	=
31	BAS	DM Bond FutureSPEC	=	=	=	=	=	35	=	=

In the above case the instrument with the identifier 33 was changed in error.

Output Limits File

A sample extract of an output limits file in subfolder Datastats after you have run the application:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Limit File Version 1.00												
2	Filename	RowType	Fieldname	Allow Empty	Min	Max	Min Date	Max Date	Length Min	Length Max	Formula	OkFormat	OkValues
3	ABSInstr	BAS:DM Bond FutureSPEC	ContractSizeVAL						15	15			ContractSizeVAL|
4	ABSInstr	BAS:DM Bond FutureSPEC	CurrencyUNIT						3	3			GBP|EUR|USD|XXX|
5	ABSInstr	BAS:DM Bond FutureSPEC	DiscountCurveXREF						17	17			DiscountCurveXREF|
6	ABSInstr	BAS:DM Bond FutureSPEC	IDENTIFIER		1	63							
41	BondInstr	BAS:DM Bond FutureSPEC	ProductTypeENUM						15	15			ProductTypeENUM|
42	BondInstr	BAS:DM Bond FutureSPEC	SettlementTYPE						14	14			SettlementTYPE|
43	BondInstr	BAS:DM Bond FutureSPEC	SpotPriceVAL		0.97169473	1023.7702							
44	BondInstr	BAS:DM Bond FutureSPEC	StlmntDayRuleBUSD						17	17			StlmntDayRuleBUSD|
45	BondInstr	BAS:DM Bond FutureSPEC	StlmntDayRuleCONV						17	17			StlmntDayRuleCONV|

As you can see, this file lists min and max numerical and date values as well as text lengths and (if field NoSingleCharCheck in tab Param is set to ‘False’) the characters used. A ‘True’ in column Allow Empty indicates that this field is empty for at least one input record.

The output limits file is intended to help you defining an input limits file. The output limits file shows limits and values which the data analyser has encountered when reading previous input.

An example of a limits input file (which you need to specify as file Limits_Input.csv in subfolder Config):

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Limit File Version 1.00												
2	Filename	RowType	Fieldname	Allow Empty	Min	Max	Min Date	Max Date	Length Min	Length Max	Formula	OkFormat	OkValues
3	ABSInstr	BAS:DM Bond FutureSPEC	CurrencyUNIT						3	3			GBP|EUR|USD|
4	BondInstr	BAS:DM Bond FutureSPEC	SpotPriceVAL		0	150							

The columns of limits files contain:

Filename, RowType, and Fieldname

These columns are identifying the fields in the input files for which the limits apply. RowType is only used or necessary for Algo(rithmics) input files.

AllowEmpty

The only valid values in here are ‘True’ or ‘False’ (or empty which means ‘False’). If set to ‘True’, empty cells are allowed for the corresponding field in the input file. If set to ‘False’, the data analyser will warn on empty input data cells.

Min, Max

These numerical values define the minimal resp. maximal allowed input values. Leave these empty if you do not want to apply them.

Min Date, Max Date

These date values define the minimal resp. maximal allowed date input values. Leave these empty if you do not want to apply them.

Length Min, Length Max

These numerical values define the minimal resp. maximal allowed input values. Leave these empty if you do not want to apply them.

Formula

You can define a worksheet formula here which is required to result either in 'True' or 'False'. The data analyser will ignore 'True' results warn on 'False' ones. You can use worksheet functions here, or to field names of the current input record: [*field*] refers to field *field*, [ThisCell] will refer to current input cell.

Beware: Excel / VBA's internal date representation is mm/dd/yyyy (US format). You will need to perform a string conversion to this format before you can apply date functions like DATEVALUE.

OkFormat

You can define a regular expression here. If the input cell adheres to this expression, the data analyser will not complain.

OkValues

Define your valid input values separated by '|', and the data analyser will only accept those.

CHAR(000) .. CHAR(255)

You need to set these columns to 'True' for string fields.

Output Limits Move File

A sample extract of an output limits move(ment) file (which you will find with name Limits_Move_Output.csv in subfolder Datastats if you have run the application with previous period's input files in subfolder Data_Files_Prev so that the program could analyse the moves / changes):

	A	B	C	D	E	F	G	H	I	J
1	Limit Moves File Version 1.00									
2	Filename	RowType	Fieldname	Allow Empty	Min	Max	Min Date	Max Date	Length Min	Length Max
33	BondInstr	BAS:DM Bond FutureSPEC	DiscountCurveXREF						1	1
34	BondInstr	BAS:DM Bond FutureSPEC	MaturityDATE		0	366				
35	BondInstr	BAS:DM Bond FutureSPEC	NAME						1	1
36	BondInstr	BAS:DM Bond FutureSPEC	PositionUnitsVAL						1	1
37	BondInstr	BAS:DM Bond FutureSPEC	PortfolioXREF						1	1
38	BondInstr	BAS:DM Bond FutureSPEC	ProductTypeENUM						1	1
39	BondInstr	BAS:DM Bond FutureSPEC	SettlementTYPE						1	1
40	BondInstr	BAS:DM Bond FutureSPEC	SpotPriceVAL		-0.057250908	1022.692554				

As you can see, this file lists min and max ***changes*** of numerical and date values as well as text lengths of changes (a '1' clearly indicates no change at all because it's '=').

For the example given above you might like to define a Limits_Move_Input.csv like:

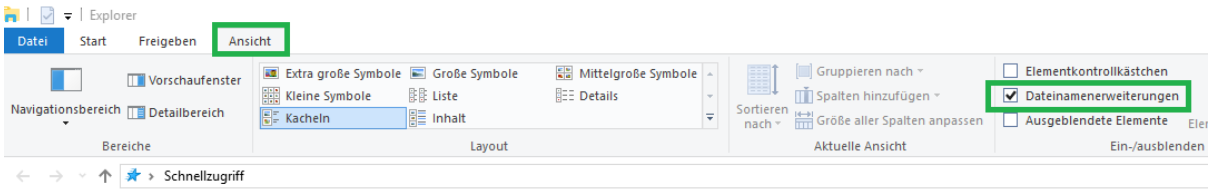
	A	B	C	D	E	F	G	H	I	J
1	Limit Moves File Version 1.00									
2	Filename	RowType	Fieldname	Allow Empty	Min	Max	Min Date	Max Date	Length Min	Length Max
3	BondInstr	BAS:DM Bond FutureSPEC	MaturityDATE		0	0				
4	BondInstr	BAS:DM Bond FutureSPEC	SpotPriceVAL		-0.05	0.05				

Best Practice and Known Issues or Errors

Set Windows File Explorer to Show File Extensions

Please note: in order for the data analyser to correctly see all input files you need to parametrize the windows file explorer to show file extensions:

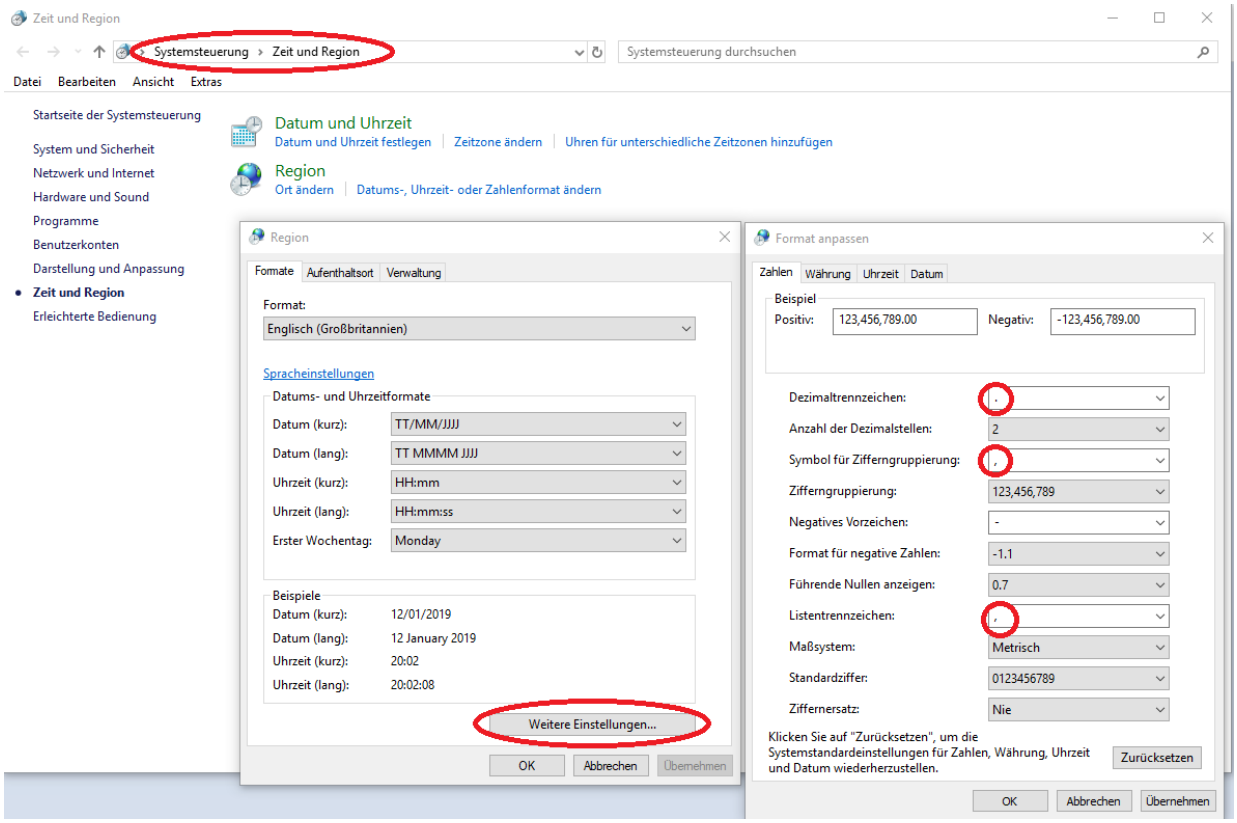
For example German Windows 10:



Specify Correct Windows Number Format for Your Data

Please specify the correct number formats for your data.

If you have a decimal separator '.', a thousands separator ',' and a list separator ',', set your Windows system settings accordingly (example for German environment given):



Application Runtime is around 7-10 times faster if run locally or on Virtual Machine (AWAVE)

It is advised to run this application locally on the C: drive or via the virtual machine AWAVE. If you run it on a network drive or in parallel to other applications – esp. Outlook, RiskWatch, Internet Explorer, etc. – then the runtime can easily become 7-10 times higher.

Release Notes

Version 37 on 19/01/2019

In module Workflow the row counter for Algo input files for tab TempMove is initialized to lRowTypes + 2 in order to create correct movement stats without inserted header information. This header information was inserted in version 36 to make lookups in MOVE_... files easier – you can now filter on header fields.

The output value for AllowEmpty fields in limit files and in limit move files was changed from True to 1. In international versions of Excel the translated value of True (for example, the German WAHR) was not treated correctly.

This version checks whether subfolder Datastats exists and terminates with an error message if not.