Australian Government
Civil Aviation Safety Authority

www.casa.gov.au

PETA – Training Materials

Publishing information

Published by:	Civil Aviation Safety Authority
Postal address:	GPO Box 2005, Canberra ACT 2601
Office:	16 Furzer Street, Phillip ACT 2606
Telephone:	131 757 +61 2 6217 1111 (from outside Australia)
Facsimile:	+61 7 3144 7575
Email:	safetysystems@casa.gov.au
Internet:	https://www.casa.gov.au/

Ownership of intellectual property rights in this publication

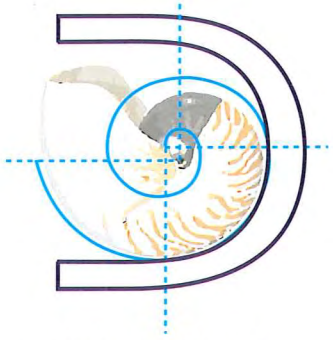
Disclaimer

The material contained in this publication is made available on the understanding that CASA is not providing professional advice, and that users exercise their own skill and care with respect to its use, and seek independent advice if necessary.

CASA makes no representations or warranties as to the contents or accuracy of the information contained in this publication. To the extent permitted by law, CASA disclaims liability to any person or organisation in respect of anything done, or omitted to be done, in reliance upon information contained in this publication.

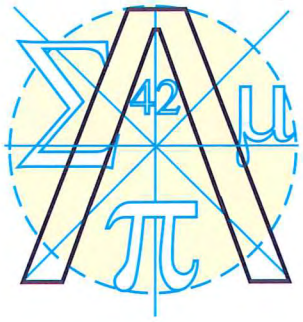
Creative Commons licence

DATA



STATISTICS

ANALYSIS



MATHEMATICS

AUSTRALIA

INFORMATION

S T R A T E G I C
I N F O R M A T I O N
C O N S U L T A N T S

Trend Analysis Training

June 2014

Project:CASA/2

Module One

Overview and Running the Trend Analysis

Prepared for the

Civil Aviation Safety Authority

June 2014

Introduction

The Safety Performance Analysis (SPA) Branch of the Civil Aviation Safety Authority (CASA) is responsible for conducting analysis and reporting of safety-related trends and risk factors to CASA's Executive Management. CASA has developed a trend analysis tool that is currently used for monitoring and reporting relevant developments in the safety-related incident data. This training covers the latest iteration of the tool built by Data Analysis Australia.

The six modules in this training are:

- Module One: Overview and Running the Trend Analysis
- Module Two: Data Sourcing
- Module Three: Interpretation of the Outputs
- Module Four: Poisson Regression and the Base Models
- Module Five: Model Extensions and Other Flags
- Module Six: Excel Template

Remember that this is a *screening* tool. We have aimed to include a large number of flags that *might*, in some cases, be associated with changes in the pattern of safety incidents so that they can be investigated further. We hope that we have included enough characteristics to detect any issues of concern, but there is no expectation that every flag raised will relate to a safety issue.

Additional Notes

Data Analysis Australia recommends that a whole number of years be analysed, and suggests either five years (20 quarters) or six years (24 quarters). The suggested models fit up to three parameters, or up to five with seasonality included, and $n = 20$ is a suggested minimum for this; $n = 24$ would be preferable, but must be tempered with validity of the data, consistency of definitions and reliability of reporting, and the likelihood of consistency of patterns or trends over that period. However four years ($n = 16$) or fewer observations is unlikely to give sufficient power to detect changes in patterns.

The tool expected the data to be quarterly and while it will run with alternative input some of the models (particularly those investigating seasonality) will return invalid output.

The data processing does not allow for combining data where the definition has changed over time. Where the definitions or method of recording the data has changed, this will need to be considered separately.

Module Two

Data Sourcing

Prepared for the

Civil Aviation Safety Authority

June 2014

CASA/2 – Data Preparation Guide

Overview of the data sourcing and preparation

The data sourcing script currently conducts reading in and preparation of four input files:

Filename	Source
OCCURRENCE.csv	Air Safety Incident Report (ASIR)
OCCURRENCE_TYPES.csv	Air Safety Incident Report (ASIR)
OCCURRENCE_ARICRAFT.csv	Air Safety Incident Report (ASIR)
SDR.csv	Service Difficulty Reports (SDR)

Table 1. List of input files incorporated in the trend analysis tool.

The four data files in Table 1 have a similar structure, where descriptive information of safety incidents is provided in a standard table format. The data preparation routine converts these tables into table of counts by variable/variables of interest and year – quarter periods. Table 2 and Table 3 below show the structures of the tables before and after the data preparation.

ID	Type.ID	Is.Primary.Flag	Year	Quarter	yr_Q	Occurence.Type.Description
221171	406506	N	2006	1	2006_1	TECHNICAL : POWERPLANT / PROPULSION : ENGINE FAILURE OR MALFUNCTION
221171	278136	Y	2006	1	2006_1	OPERATIONAL : TERRAIN COLLISIONS : COLLISION WITH TERRAIN
221186	278138	Y	2006	1	2006_1	OPERATIONAL : CREW AND CABIN SAFETY : FLIGHT CREW INCAPACITATION
221289	391922	Y	2006	1	2006_1	AIRSPACE : AIRSPACE INFRINGEMENT
...	...	...	...	...	...	...
304971	411105	Y	2014	1	2014_1	ENVIRONMENT : WILDLIFE : BIRDSTRIKE
304974	411109	Y	2014	1	2014_1	ENVIRONMENT : WILDLIFE : BIRDSTRIKE
304976	411115	Y	2014	1	2014_1	ENVIRONMENT : WILDLIFE : BIRDSTRIKE
304983	411135	Y	2014	1	2014_1	ENVIRONMENT : WILDLIFE : BIRDSTRIKE
304984	411138	Y	2014	1	2014_1	ENVIRONMENT : WILDLIFE : BIRDSTRIKE

Table 2. OCCURRENCE_TYPES.csv data as standard table format – Please note that a preliminary data preparation (addition of year, quarter, yr_Q, and collated description fields) have been done to the data shown above.

year	quarter	AIRSPACE : AIRCRAFT SEPARATION : AIRBORNE COLLISION ALERT SYSTEM WARNING	AIRSPACE : AIRCRAFT SEPARATION : COLLISION	...	TECHNICAL : SYSTEMS : OIL SYSTEM	TECHNICAL : SYSTEMS : OTHER
2006	1	8	0	...	0	0
2006	2	4	0	...	0	1
2006	3	7	0	...	0	1
2006	4	2	0	...	0	0
2007	1	6	1	...	0	5
2007	2	4	1	...	0	2
...	...	...	...	...	...	...
2012	4	12	0	...	0	12
2013	1	8	0	...	0	11
2013	2	14	1	...	0	7
2013	3	17	0	...	0	14
2013	4	21	2	...	0	19
2014	1	1	0	...	0	3

Table 3. OCCURRENCE_TYPES.csv data as table of counts by occurrence description and year – quarter periods.

A conversion of the data from the standard table format to the count table format is done via predefined function (*fnDFBuild*). It takes the input data and collates counts of each values in user specified variable, or combination of variables. As with any predefined functions, it expects input data to have a specific format and requires specific inputs from the users. It currently does not include any automated data checking algorithms as meaningful checks generally require knowledge of the data.

The following section provides demonstrations of data diagnostics available in R to aid this processes. The section subsequent to that will discuss how the input data for the *fnDFBuild* function must be formatted.

Data Diagnostic Guide

The first check that should be conducted on any data after reading them in to the workspace is confirming that the input file was read in successfully and the data contains all the expected rows and fields. A useful R command for checking this is *str* function.

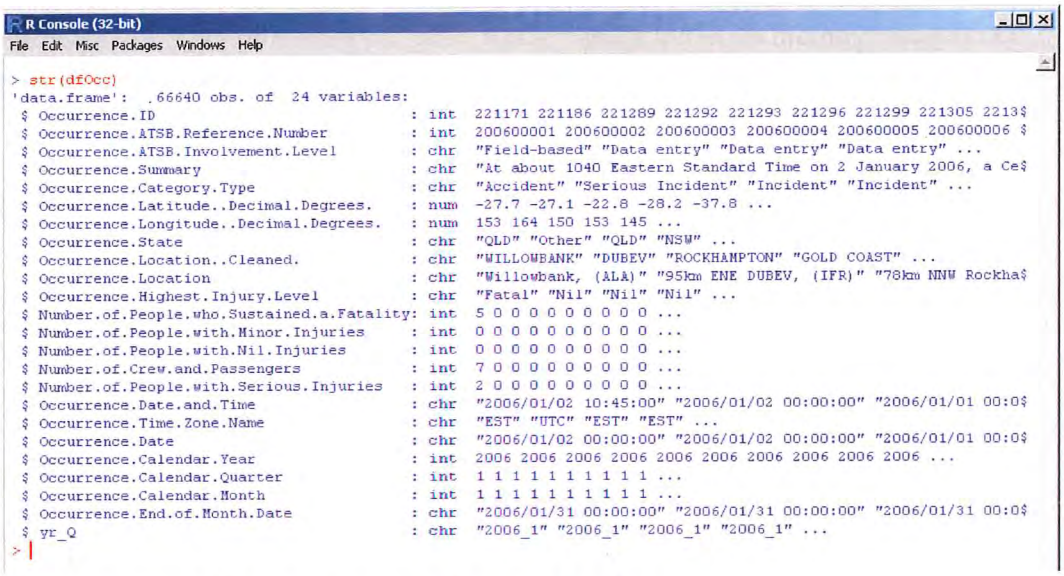


Figure 1. Output information returned from *str* function.

The first line of the str output contains the following information of the object:

1. Data class – such as data.frame, matrix, list, etc;
2. Number of observations, i.e. number of rows; and
3. Number of variables, i.e. number of columns.

Then a list of all variables in the data object is printed – indicating the variable name, variable class and first few values of each variable.

Key questions about the data object that *str* output can or help address are:

- Do the data contain a field or fields indicating the date of the incident occurrence? This is crucial information required in all input data. The date information is used to assign year-quarter period to the incidents, therefore information in any format can be accepted as long as they provide year and month of the occurrence. R have a set of built-in date formats you can use in *as.Date* to convert date information in text format to dates.

Format components	Definition
%d	Day of the month (01 – 31)
%m	Month (01 – 12)
%Y	Year with century
%H / %I	Hours (00 – 23) / Hours (01 – 12)
%M	Minute (00 – 59)
%S	Second as decimal number (00 -61)
%p	AM/PM indicater

? as.Date in R will assist.

Table 4. List of date formats available in R.

For example, a text string of "2014/05/30" can be converted to a date value with *as.Date("2014/05/30", format = "%Y/%m%d").*

- Do the data contain expected fields? If not, why are there missing or unexpected variables in the data? You may need to review the input file to ensure there are no unexpected breaks or gaps in the data. If the input file is not CSV, then consider using *read.table*.
- Do the variables have appropriate names? R automatically converts some special characters while reading in the data – do you need to rename any variables?
- Are there any variables with unexpected class assignment, such as numerical variable read in as character? A numerical variable can be read in as character if there are non-numeric items in the field such as hyphons, periods or "NA" to indicate blank. Classes of the variables can be converted with *as.numeric* and *as.character* commands in R, however inconsistent values, such as "NA" in numeric fields, must be addressed prior to such conversion.
    - Should there be any missing, and if so, how should they be indicated? – as blanks or as actual value such as "NA", "NULL", etc. Consistent expression for the same value of a variable should be used, that is, there shouldn't be a mix of "NA" and blank in the same variable to indicate missing, unless there are meaningful difference between "NA" and blank.

Once the preliminary checks on the data are done and the data has been confirmed to be loaded successfully, they can be processed in to a format accepted by *fnDFBuild*.

Generation of Table of Counts with *fnDFBuild*

- *fnDFBuild*, henceforth referred to as the function, only accepts standard table structure – also expressed as wide format as opposed to long. A screenshot of an R console showing examples of wide and long format data is given in Figure 2. Given the nature of the data, we expect that most, if not all, the possible input data would be in the wide format. However, if there is a need to change the data structure from long to wide, *reshape* command in R can be useful.

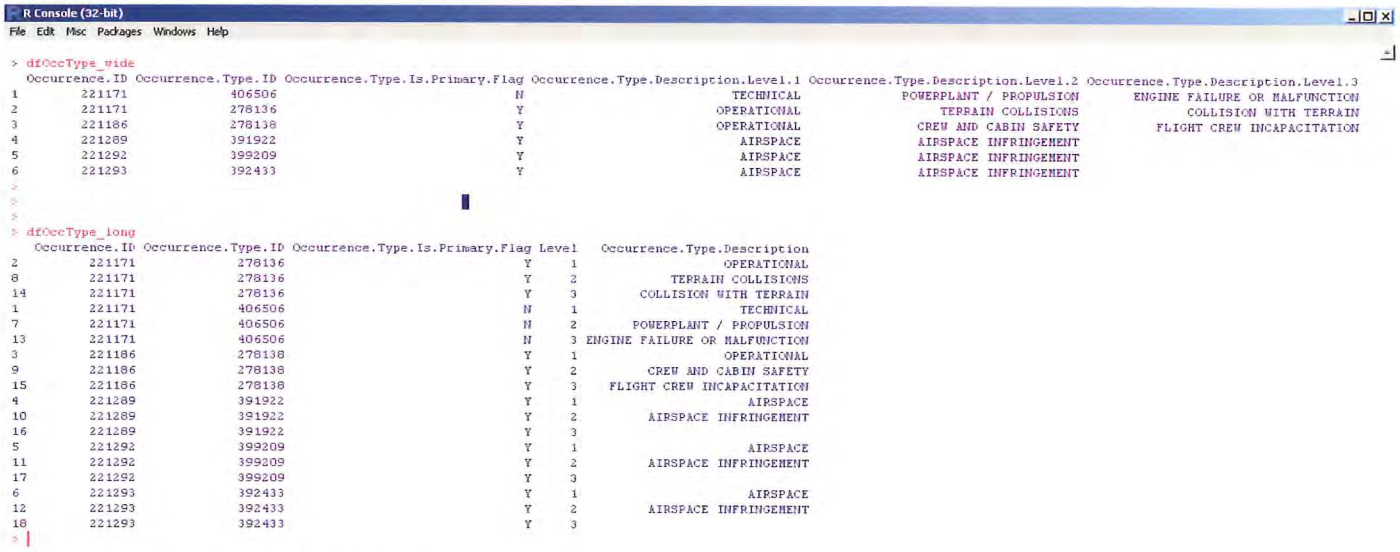


Figure 2. Examples of wide and long format data.

- The input data must have a collated year and quarter field – "yr_Q" (e.g. 2014_1 to indicate first quarter in 2014). If the data does not have occurrence date field, it must be sourced from appropriate data – for example, "OCCURRENCE_TYPES.csv" does not have occurrence date, so it had to be sourced from "OCCURRENCE.csv".
- You may need to consider recoding variables with long string or boolean values as the models and analysis results are referred by their values or combination of values. For example, if we were to conduct trend analysis on counts of incidents due to a combination of "Cause.Corrosion" and "Cause.Inadequatte.Maint" (in the SDR data), the analysis output will show TRUE_._TRUE, FALSE_._TRUE, TRUE_._FALSE and FALSE_._FALSE. It will be feasible to identify which model corresponds to which combination, but will make interpretation of the final output difficult. For the case of the above example, the boolean values for "Cause.Corrosion" can be converted to "Corrosion" and "NotCorrosion".
- The function will not accept a combination of variables from both the occurrence and SDR data. This is because the two data sets do not share consistent reference ID, therefore cannot be merged to generate single table to collate counts of incidents. The function will print an error message *"fVarlist cannot contain variables from SDR and Occurrence data."*

As shown in Table 3, the headers of the table of counts are values or combination of values of the variable/s being analysed. These headers are used in the modelling process to define the model formula. There are number of restriction on the characters acceptable in the model formula, so we have placed a process where the headers of the table of counts are formatted to consist only the acceptable characters.

Trend Analysis – Troubleshooting

The trend analysis routine proceeds to the model fitting, model selection and tabulation of relevant flags following the generation of table of counts. The function, *fnRunTAnalysis*, conducts these procedures. There are various conditional statements in this function that prevents critical errors from incapacitating the analysis. However, an unforeseen error could pass such cautionary measures and affect the tool, so given below are examples of errors and issues we encountered during the tool's testing phase that may help future users should any problems arise.

- Duplication of values – In the process of formatting the header of the table of counts, some values, or combination of values may become duplicates of the other. For example, "MURRIN-MURRIN" and "MURRIN MURRIN" in "Occurrence.Location..Cleaned." will become duplicated columns of "MURRIN_MURRIN" in the table of counts after header formatting. This will stop the analysis tool from working.

- Be wary of any special characters – so far we have noted the following to cause issues:

  Space (" "), forward slash ("/"), hyphen ("-"), brackets ("(" and ")"), and ampersand ("&").

  A typical error message displayed when a special character was found in the formula is *"Error in eval(expr, envir, enclos) : object 'AIRSPACE' not found"*, which in this example, there was a special character just after *"AIRSPACE"* in the value being analysed.

- Blanks in the data:
  - A variable cannot have all its values as blank. This will cause the fnDFBuild function to fail.
  - If a combination of variables is being analysed – there cannot be a combination where the first value is blank. This is also due to the limitation of the characters acceptable in the model formula. The analysis function skips over such values, or combination of values.

Module Three

Interpretation of the Outputs

Prepared for the

Civil Aviation Safety Authority

June 2014

Module Four

Poisson Regression and the Base Models

Prepared for the

Civil Aviation Safety Authority

June 2014

Some Guiding Statistical Principles

Generalised Linear Modelling for Count Data

Safety incident data is count data (that is discrete and numeric). The Poisson distribution is often used to model count data. The Poisson distribution gives the probability of k events occurring in a specified time interval (and assumes that the times between events are independent), it also assumes the events are independent. It is appropriate even if the expected number of events is small.

The consequence of assuming that aviation safety incidents have a Poisson distribution is that generalised linear models must be used in the place of ordinary linear regression. The generalised linear modelling framework allows for the response variable to have a distribution that is not normal. Each distribution has a natural link function (a transformation applied to the expected value of the response variable to ensure fitting a linear model is reasonable) and the natural link function for the Poisson distribution is the logarithmic link. A linear model is fitted to the logarithm of the expected count at a given time, with the observed counts assumed to follow a Poisson distribution whose mean is equal to that expected mean.

Identifying Different Sources of Variation

Many different aspects can contribute to variation in observed data. A common aim is to identify and allow for different sources, and to flag for further attention sources of variation that might be of concern or interest. Such explanations for variation could be linear trend, the last observation is different or special (that is, it departs from the earlier pattern), clusters, seasonality, or different exposure to risk of the incident.

Recording and Detecting Clusters

There are two forms of clusters that may occur in aviation safety incidents.

First, one event might generate more than one incident (for example, if there was insufficient separation between two aircraft then at least two incidents would be reported). This is correctly accounted for during the data setup, where multiple records may be associated with a single occurrence id.

Second, behaviours like systematic failings in maintenance might lead to a cluster of incidents. This should be detected in the trend analysis as proposed by Data Analysis Australia as an unusual spike in the relevant category; however, it does highlight the need to perform analyses both on individual occurrence types and on aggregates by meaningful categories such as geographical locations.

Generalised Linear Modelling with an Offset

A key consideration of the trend analysis is understanding the expected behaviour. The assumption of a linear trend over time may be appropriate in some circumstances. If, for example, there was a linear increase in the number of flights

departing from a particular location then it would potentially be acceptable to see a corresponding linear increase in incidents at that location over time. However, if the number of flights departing remained constant but the number of incidents increased in a linear fashion this may be considered cause for concern. Therefore, the trend analysis should compare the observed trend relative to the expected trend.

One key consideration in the analysis of safety incidents is not just the incident itself, but the number of incidents relative to opportunity. Opportunity for incidents, or exposure, might be difficult to measure. If it was possible to develop an exposure measure the generalised linear model framework outlined above can be extended to allow for exposure by including it as an offset. Then we require a value for exposure corresponding to each timepoint (that is each quarter).

An apparent increase in the number of incidents might be simply explained by increasing exposure – the number of occurrences per unit exposure might not have changed. However, if, after allowing for exposure, there is still an increasing trend in the number of incidents, this might give additional cause for concern. Then what might seem like an unchanged trend (and therefore of no concern) in a particular incident class might become a concern if the potential for that type of incident is declining.

Trend Analysis Tool

Building on the trend analysis approach of CASA with the primary focus of the quarterly reporting focusing on concerns about recent events, Data Analysis Australia has incorporated the following elements into the tool:

- A generalised linear modelling framework;
- Assuming an underlying Poisson model;
- A logarithmic link, modelling the logarithm of the expected counts as a linear function of time and thereby precluding negative expected counts;
- An evaluation of whether the most recent observation is consistent with expected behaviour based on previous observations;
- An examination of other patterns or sources of variation in the data;
- Detecting and flagging overdispersion, and modifying the analysis to allow for it;
- Detecting and flagging underdispersion; and
- The ability to include an exposure measure, if available, to account for additional variation and estimate the rate of incidents per unit of exposure. (If the exposure is changing linearly over time, the time variable can act as a proxy for exposure, but an independent measure of exposure is much preferred.) To come in the next version.

Generalised Linear Modelling Framework

When data arises from counts of random rare events it typically has a Poisson distribution and is therefore usually modelled using a generalised linear modelling framework, with a logarithmic link function.

The primary model fitted by Data Analysis Australia has two terms, a linear trend term and a `final observation' effect. The final observation effect is an indicator variable which is one for the final observation and zero elsewhere and will be significant if the final observation is `special' in some way and doesn't fit in with the pattern established over the proceeding time periods. Once this model is fitted, model-selection techniques are used to determine whether the model can be simplified. The final (best-fitting) model is selected from among special cases of these models is the simplest model that adequately describes the data. The final model would be one of four possibilities:

M1. A model with both a linear trend and final observation effect;

M2. A model with a final observation effect;

M3. A model with a linear trend; or

M4. A constant model.

The relationships between these models are displayed in Figure 1. An arrow connecting a model A to a model B indicates that model B is a special case of model A – we say model B is nested within model A, and a likelihood ratio test can be performed to compare the models, or, equivalently, to test whether the term that has been dropped is statistically significant.

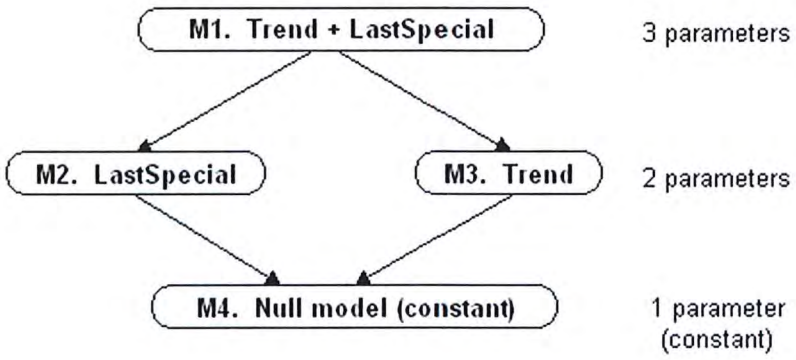


Figure 1. Lattice diagram of nested models for model selection. An arrow from a model leads to a simpler model that is a special case of the first model. Likelihood ratio tests can be used to compare two such models.

Starting with the most general model, M1, paths are followed through the lattice, applying likelihood ratio tests until a statistically significant difference is encountered, implying the model cannot be further simplified. The best-fitting model is the simplest model that can be reached in this way.

If either M1 or M2 is found to be the best-fitting model (we cannot proceed to M3 or M4), this means that the LastSpecial term is statistically significant. This suggests that the final observation is inconsistent with the pattern established by the prior observations, and therefore warrants attention. The level of statistical significance is used to highlight the level of concern. Both higher and lower counts than expected can be identified, and a plot of the data and fitted models is provided to quickly identify the direction of effects. P-values for a given effect are obtained from the likelihood ratio test between models including and excluding the relevant term.

Note that the statistical significance of any given term depends on which other terms are in the model. For example a linear Trend term might be not significant if a Last special term is present (comparing models M1 and M2), but significant if the Last special term is absent (comparing models M3 and M4).

Typically, we follow a path down through the lattice, testing at each step, dropping terms that are not statistically significant, until we can simplify the model no further. The aim is to find the simplest model (most parsimonious, with fewest parameters) that adequately describes the data.

A single unambiguous "best" model is not guaranteed. For example we may find that M1 and and M2 are not significantly different, but we cannot proceed to M4, and also that M1 and M3 are not significantly different, but we cannot proceed from M3 to M4. Either model M2 or model M3 might be acceptable – we need *either* Trend or Last special to explain the variation in the data, but *not both* – we don't need the generality of model M1. In this case we choose between models M2 and M3 on the basis of a measure of goodness of fit for each model.

The Poisson Generalised Linear Model

The counts (Y) are assumed to follow a Poisson distribution with a mean (μ) which may depend on a time (t). This is denoted

$$Y \sim \mathrm{Po}\big(\mu(t)\big).$$

We model the mean as a function of time t with a logarithmic link function, which ensures expected counts are never negative.

$$\ln\big(\mu(t)\big) = \alpha + \beta t$$

And we estimate the intercept (a) and slope (b) respectively.

We can test the strength of the linear relationship (whether the slope, β, could be zero). Furthermore, we can test whether the most recent observation is inconsistent with the pattern of the rest of the data by including an indicator variable for the last timepoint. This indicator variable takes the value zero for all timepoints except the last, where it is one.

If a separate measure of exposure within each time period (e.g. quarter) is available, such as the number of aircraft flying hours for a particular type of aircraft, or the number of takeoffs or landings, which might vary over time, this can be incorporated into the analysis, effectively modelling a rate, the number of occurrences per unit of exposure as a linear function of time. The term $\mathrm{Exposure}_t$ is called an offset. Note, it

has no coefficient to be estimates. This is the model we fit, but the equation can be rearranged showing that α and β define a linear trend of time for (the logarithm of) the average number of occurrences per unit exposure.

$$\ln(\mu_t) = \alpha + \beta t + \ln(\text{Exposure}_t)$$

$$\ln(\mu_t) - \ln(\text{Exposure}_t) = \alpha + \beta t$$

$$\ln(\mu_t \,/\, \text{Exposure}_t) = \alpha + \beta t$$

$$\text{Average number of occurrences per unit of exposure} = \frac{\mu_t}{\text{Exposure}_t} = \exp(\alpha + \beta t)$$

Modelling with the offset in the explanatory model permits the observed counts to still be modelled as a Poisson-distributed variable with average μ_t, whereas if we divided the observed counts by the exposure measure to create a rate before analysing, we would lose the Poisson distribution, and the relationship of the variance to the mean, whereby larger counts tend to have larger variability – whether those larger counts are due to increased exposure or increased rate of occurrences. Then the assumed statistical model would be inappropriate and it would be difficult to properly assess whether an individual value or pattern is unusual.

If in fact exposure in each time period did vary over the periods included in an analysis, incorporation of a relevant measure of exposure would facilitate a superior analysis, by explaining more of the observed variability and thereby allowing greater sensitivity to detect departures from the expected trends or patterns.

Dispersion

Under the Poisson model, the variance of the observed count Y at any given time t is equal to its variance at that time. Sometimes the variance in the data is larger than this. This is called overdispersion. Occasionally the variance is smaller than expected, called underdispersion.

By preference, these models for count data are fitted using the Poisson distribution and tests comparing nested models (one model is a special case of the other, connected by a line in Figure 1) are based on the chi-square distribution for the change in deviance. This is a likelihood ratio test. However, if overdispersion is identified in model M1, this is flagged, and the quasi-likelihood method is used.

Dispersion can be estimated by the Pearson chi-squared goodness of fit statistic, sum of $(\text{observed} - \text{expected})^2/\text{expected}$, divided by the residual degrees of freedom. It should be approximately one. The Pearson chi-squared goodness of fit test can be used to test for overdispersion.

If dispersion is found to be statistically significantly greater than one in model M1, this is flagged as overdispersion and the quasi-likelihood method is used, specifying the distribution family as quasi-Poisson.

Then the likelihood ratio chi-square tests are replaced by F tests, with the F statistic calculated as the change in deviance divided by the change in degrees of freedom, divided by the estimated dispersion, and compared with an F distribution with numerator degrees of freedom equal to the change in degrees of freedom (which is

the same as the change in the number of parameters estimated) and denominator degrees of freedom equal to the residual degrees of freedom in the more general model (which is the same as the number of observations minus the number of parameters estimated in this model).

If the dispersion is statistically significantly less than 1, this is flagged as underdispersion, and suggests unusually small variation, possibly indicating improper recording of incidents and worthy of investigation. However the Poisson model is fitted in this case.

Summary of Flags Relating to the Base Models and Dispersion

The following six flags relate to the base models, and investigations into dispersion:

- **Last special**. The last observation deviates from the pattern (linear trend or constant) for the prior observations. The direction and strength of evidence is indicated by bold (🡅, 🡇) or finer (↑, ↓) arrows, and colour-coded as in the CASA START tool to strongly highlight this most important indicator of something different happening in the last quarter.
- **Trend**. A linear trend is evident in the data. A sloping arrow (↗ or ↘) indicates the direction.
- **Second-last special**. When the last observation is omitted, the second-last observation deviates from the prior pattern. Indicated similarly to the Last special flag (🡅, ↑, ↓ or 🡇). Outcome might differ from outcome of the previous run due to using slightly different data.
- **Last quarter changes pattern**. This flag will be triggered if, for example, there is no linear trend prior to the last quarter, but when the last quarter is included a linear trend becomes evident, or vice versa.
- **Underdispersion**. Displayed as ^^^, this flag indicates that, in the model including linear Trend and Last special, there was less variability in the data than expected from a Poisson distribution – the counts were unusually consistent. A *possible* cause of this is inadequate reporting, and therefore this might warrant checks on the adequacy of reporting. Poisson models are still fitted for all the screening tests, despite the underdispersion.
- **Overdispersion**. Displayed as /\/\/\, this flag indicates that, in the model including linear Trend and Last special, there was greater variability in the data than expected from a Poisson distribution – there is unexplained extra variation. This is analogous to the high variability flag reported in in the CASA START spreadsheet. In the models fitted and statistical tests used, a quasi-Poisson model is used to adjust for this.

Module Five

Module Extensions and other Flags

Prepared for the

Civil Aviation Safety Authority

June 2014

Model Extensions and Other Flags

In addition to fitting the four basic models M1 to M4 and flagging linear trend, last special, and unusual dispersion (variability), the Excel tool fits several other models to highlight whether the series has any of these features:

1. Evidence of a change in level;
2. Evidence of non-linearity;
3. Evidence of seasonality;
4. Dispersion is reduced by at least one of a change in level, non-linearity, or seasonality;
5. An excess of zeros, more than expected from the best-fitting Poisson distribution; and
6. One or more outlier.

Model Extensions

In the procedure implemented by Data Analysis Australia for the CASA safety incident screening, we focus initially on selecting the best model from among models M1 to M4. Various extensions of the models described in the previous module can be investigated to see if alternative explanations of the data are feasible. Some of these might raise flags for further investigation, or they might alleviate concern by providing a feasible alternative explanation for the patterns observed.

We consider some generalisations of models M1 to M4 to flag possible causes of variation including a change in the level (M5), non-linearity (M6), and finally we consider the possible effects of annual seasonal cycles (models M7 to M10).

The full modelling process implemented by Data Analysis Australia is described here, containing more statistical modelling details. A sequence of models is fitted and nested models (where one model is a special case of another) are compared to determine statistically significant trends and patterns.

Testing for Change in Level or Non-linear Trend

A general principle of statistical model-fitting is to choose the simplest or most parsimonious model (fewest parameters fitted) that adequately describes the variation in the data. We call this the best-fitting model.

Initially we choose the best-fitting model out of M1 to M4 and the flags for LastSpecial, Trend, Underdispersion and Overdispersion are based on these alone. This was covered in the previous module.

We then fit two models that generalise one of these (see Figure 1):

M5. Change in level – begins at one level, and at an estimated changepoint jumps (up or down) to another level. Three parameters must be estimated. Model M2 is a special case of this, with the changepoint set

between the last two observations, and thus, in the lattice diagram, M5 sits above M2 with an arrow leading to M2.

M6. Quadratic trend model – enables a crude test for evidence of nonlinearity, by adding a squared term to the linear trend model M3. The M6 is a generalisation of M3 and sits above it in the figure below.

If model M5 is found to be statistically significantly better than M2, a flag for Change in level is raised. If model M6 is statistically significantly better than M3 (the quadratic term is significant), then a flag for nonlinearity is raised.

Of course many other models are possible, but we chose to limit the investigation to models with up to three parameters to be estimated. With only 20 or 24 quarterly observations (from 5 or 6 years) estimating more parameters could be misleading.

These six models cover a range of possible shapes with the opportunity to identify different ways to describe the trends or patterns and to detect possible safety issues.

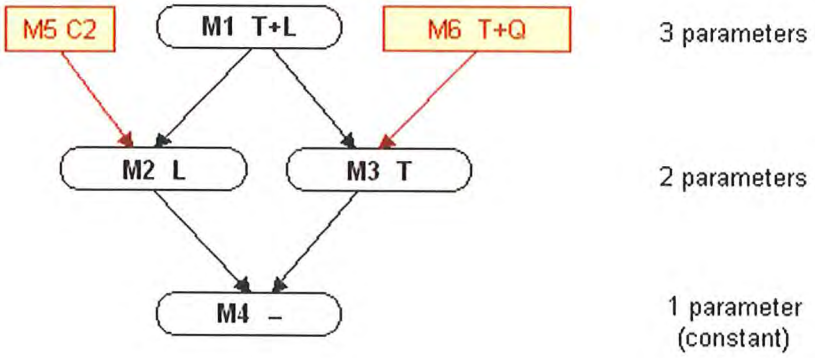


Figure 1. The base models with the addition of models M5, Change in level at an estimated changepoint, and M6, Quadratic trend. T indicates linear Trend, L indicates Last observation special, C2 indicates Change in level at an estimated Changepoint (two parameters to be estimated), and Q indicates a Quadratic trend term.

Models M5 and M6 cannot be directly compared by a likelihood ratio test because they are not nested, but because they have the same number of parameters, we can compare them on the basis of their residual deviance. Deviance is the generalised linear modelling analogue of residual sum of squares, and smaller deviance indicates a closer fit to the observed data. If the change in level model has smaller deviance it is said to have the better fit. The comparison of models M5 and M6 is not reported on in the current version of the tool.

Investigating Seasonality

Notwithstanding the comments about the number of parameters fitted, we examine possible annual seasonal cycles in a limited set of models, M1 to M4. A simple sinusoidal cyclic seasonal term requires two parameters, which can be thought of as an amplitude (how large the oscillations are) and a phase shift (where the peak occurs along the cycle). (In fact we fit them as the sum of a sine and a cosine curve –

a mathematically equivalent parameterisation but simpler to fit.) This yields models M7 to M10 in Figure 2.

The Seasonality flag is triggered if an acceptable model with seasonality (one of M7 to M10) is significantly different from the corresponding model without seasonality (among M1 to M4) – this is a two degree of freedom test because it involves dropping two parameters simultaneously – it is meaningless to consider the amplitude or the phase in isolation – unlike the other comparisons which are all tests with one degree of freedom.

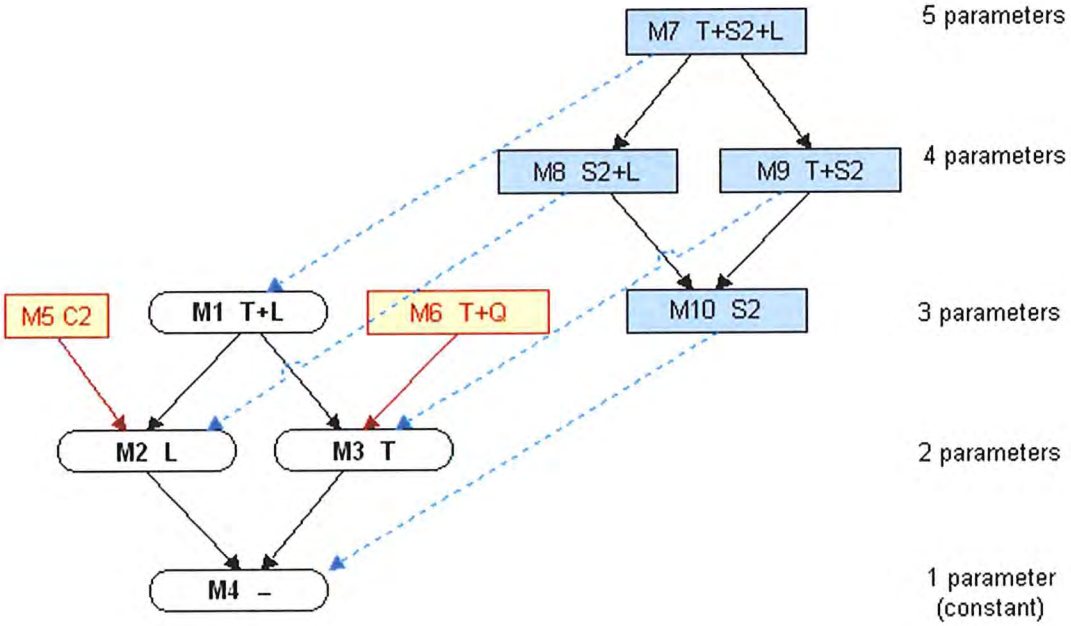


Figure 2. Figure 1 with the addition of an annual seasonal cycle. C2 indicates a change in level at an estimated Changepoint (two parameters), T indicates (linear) Trend, Q indicates Quadratic trend, S2 indicates Seasonality (two parameters), and L indicates Last observation special.

Figure 2 displays all the models considered in the suggested approach. The number of parameters in each model (the same within a row) is shown on the right hand side.

Every model includes a constant or intercept parameter. Each additional code letter indicates one added parameter, except for C2 and S2 which each require two parameters.

Following an arrow from any model to a model lower in the lattice corresponds to dropping a term from the model, and the second model is a special case of the first. We say the second model is *nested within* the first. A statistical hypothesis test comparing these two models is a test of whether that dropped term is needed. If the test is statistically significant, that term is important and cannot be dropped.

When a Poisson distribution is assumed, the tests are chi-square tests with degrees of freedom corresponding to the difference in the number of parameters estimated – thus one degrees of freedom for most tests, but two degrees of freedom for the tests of seasonality, corresponding to the blue dashed lines in Figure 2.

Summary of Model Extension Flags

The following four flags relate to the model extensions and possible reduction in dispersion.

- **Change in level**. Generalises the Last special model to allow a change in level at any time point. The flag triggers if the Change in level model is statistically significantly better than the Last special model. Displayed as a step symbol, indicating either a step up (_|‾) or a step down (‾|_).
- **Nonlinearity**. Generalises the linear Trend model by adding a quadratic term to provide a crude test for nonlinearity. The flag triggers if the quadratic term is statistically significant. The symbol displayed indicates whether the departure from linearity tends to be upwards or downwards.
- **Seasonality**. Models M1 to M4 are fitted with the addition of a simple annual cycle to investigate seasonality. This flag triggers if the seasonality is statistically significant, and the symbol ~~ is displayed to indicate the presence of an annual cycle.
- **Dispersion changed**. Overdispersion indicates extra variation from unexplained sources. Sometimes that variation can be explained by including other terms such as quadratic trend or seasonality, and the apparent overdispersion disappears in these models. If this occurs it is indicated by the symbol ~~AAA~~ and code to indicate which models reduced the dispersion.

Other Flags

There are two other flags which highlight other potential data issues. These are:

- **Excess zeros**. This flag triggers if the number of zeros in the data is unusually large for a Poisson distribution, and is displayed as 000. This could be evidence of under-reporting in some of the quarters. It is tested by fitting a zero-inflated Poisson version of model M1 and examining the statistical significance of the term for the zero inflation. Zero-inflation may be one explanation for overdispersion, but we cannot say if it is the sole cause.
- **Outliers**. If the data contains one or more gross outliers, this can affect other tests. For example a single extreme outlier can increase the variation so much that trends appear not significant and therefore would not be flagged. Therefore if this flag triggers, the reader should be aware that other characteristics of the data might be obscured and the graphical plot should be examined. The number displayed is the count of such outliers.

Module Six

Excel Template

Prepared for the

Civil Aviation Safety Authority

June 2014

Trend Analysis Output Generator Guide

Introduction

The Trend Analysis Output Generator (TAOG) is a Microsoft Excel Macro-Enabled Worksheet which facilitates a graphical/prettier display of the modelling output from the R Trend Analysis Tool. The TAOG is reasonably simple in its design and, consequently, can be modified without too much effort.

The TOAG has the following key features:

- Formulas;
- Conditional formatting;
- Named ranges;
- A macro; and,
- The use of the unusual format "Windings".

In this short guide we will look at how to operate the TOAG and provide a description of how it uses the above key features. With understanding of this and some basic excel background, one can troubleshoot and potentially modify the generator with little effort.

Basic Operation

When opening the TAOG, be sure to enable the macros. Once you have it open, you will see the following three worksheets:

- Formatted Display;
- Raw Data; and
- Lookups.

For basic operation, we are only concerned with the Formatted Display worksheet. From that worksheet you can select and view the formatted output csv file from the R Trend Analysis Tool. To do this follow these steps:

1. Click on the grey button labelled "Update Input Raw Data". This will open an Import Wizard.
2. In the import Wizard, navigate to the location of the csv file you wish to display and open it. You will notice that the wizard can only import csv files.

You should now be able to view the formatted display of the output csv file you loaded in the Formatted Display worksheet. To the right of the "Update Input Raw Data" button, you can also see the location and name of the output csv file which you loaded and are viewing. If you would like to change the csv file which is displayed, repeat the above two steps again.

In general, it is a good idea to keep a copy of the TAOG and the csv file which it's displaying in the same directory.

How It Works

Will now give brief overview of the internal workings of the TAOG to give you an understanding of what it does, how it works and how to troubleshoot and modify it.

The Import CSV Macro

Now that you know how to load a raw data csv file from the R Trend Analysis Tool using the "Update Input Raw Dataset" button, let's take a look at how it works. The button triggers a macro named "ImportCSV" which invokes an Import Wizard to let you navigate to, and select the csv file you wish to view. The selected file then gets copied into the "Raw Data" worksheet and the directory and name of the csv file are placed to the right of the button. You can select any csv file to load into the TAOG, but unless the csv file is output from the R Trend Analysis Tool, don't expect to see anything meaningful in the "Formatted Ouput" worksheet!

If for any reason the macro breaks, you can manually copy-paste the output csv file into the Raw Data worksheet.

Generating the Formatted Display

The Raw Data worksheet is merely a place to store the raw data - now we will look at how that it is displayed in the Formatted Display worksheet through the use of:

- Formulas;
- Conditional formatting;
- Named ranges; and,
- The use of the peculiar format "Windings".

The named ranges are located in the "Lookups" worksheet and dictate what character is displayed in the "Formatted Display" in place of the basic code used in the raw data. Not all the flags utilise a named range – generally, flags with more than two possible values use a named range. In the first instance, have a good look at the worksheet and examine the various codes and corresponding flags.

The most complicated of the flags is the "Last Special" flag – we will look at how that works, which should give you a feel for how the remaining flags are displayed.

The "Last Special" flag in the Formatted Display worksheet has the following features:

- A formula (essentially does a look up on the "last.special" named range);
- Conditional formatting (applied to colour the cells a particular colour);
- The font of that flag is "Windings" (displays the characters as arrows).

With some working knowledge of Excel, it should be straight forward to see that the symbol displayed in the Last Special flag can be altered by modifying the second column of the "last.special" named range; and the colouring of the flag can be changed by altering the conditional formatting rules. It would be unwise to change the font of the flag as that will no longer display the arrow character.

Column	Formula	Named ranges	Conditional formatting	Font
Occurrence Type	Yes		No	Standard
Average Count	Yes		No	Standard
Last Count	Yes		No	Standard
Model Fitted	Yes	model.fitted	No	Standard
Exposure	Yes		No	Standard
Trend	Yes	Trend	No	Standard
Last Special	Yes	last.special	Yes	Windings
Second Last Special	Yes	second.last.special	No	Windings
Last Quarter Changes Pattern	Yes	lqcp	No	Standard
Under-dispersion	Yes		No	Standard
Over-dispersion	Yes		No	Standard
Excess Zeros	Yes		No	Standard
Outliers	Yes		No	Standard
Change In Level	Yes	change.on.level	No	Standard
Nonlinearity	Yes	nonlinearity	No	Standard
Seasonality	Yes	seasonality	No	Standard
Over-dispersion removed	Yes	odr	No	Standard

Table 1.

Table 1 above summarises the excel features for each flag in the "Formatted Display" worksheet.