Notes on R scripts used in the analysis of the Elephant Trade Information System data to report to the 18th meeting of the CITES Conference of the Parties (CoP18)

GitHub repository ETIS_CITESReporting_RCode v.CoP18

Fiona M. Underwood

2019

Fiona M. Underwood
www.fmunderwood.com
fiona@fmunderwood.com

This document marks the first time that the analytical code for the ETIS analysis to a CITES Conference of the Parties has been placed in the public domain. It should be cited as: Underwood FM (2019). Notes on R scripts used in the analysis of the Elephant Trade Information System data to report to the 18th meeting of the CITES Conference of the Parties (CoP18): GitHub repository ETIS_CITESReporting_RCode v.CoP18. Zenodo https://doi.org/10.5281/zenodo.3334771

Thanks to Tom Milliken, Thomasina Oldfield, Steven Broad and Antony Bagott of TRAFFIC International, Bob Burn and the MIKE-ETIS TAG for reviewing and commenting on this document to help improve its clarity.

Contents

This document is a set of notes to accompany the R scripts in the GitHub repository https://github.com/fmunderwood/ETIS_CITESReporting_RCode version v.CoP18. This version contains the R scripts used to produce the analysis of the Elephant Trade Information System (ETIS) data for a report (Milliken *et al.*, 2018) to the 18th meeting of the CITES Conference of the Parties (CoP18).

Background

ETIS, managed by TRAFFIC International, was mandated by CITES to track levels and trends in the illegal trade in elephant ivory and other elephant products under CITES Resolution Conf. 10.10 to inform decision making on elephants. ETIS comprises reported records of seizures and subsidiary data on covariates related to the illegal ivory trade. ETIS is required to produce reports to CITES Standing Committees and CoPs.

Methodological Notes

The purpose of this document is to describe the process by which the R scripts are used to produce outputs for reporting to CITES. It is not to present the methods behind the ETIS analysis in any kind of detailed manner. Familiarity with the literature describing the ETIS methods is assumed.

The underlying methodology used for analysis is described in the Technical Report Burn & Underwood (2013). The modelling framework and detailed results are also described in the peer-reviewed paper Underwood *et al.* (2013). Results from using this methodology were first presented in a CITES report at CoP16 in Milliken *et al.* (2013).

Some revisions to the methodology have occurred since 2013. These have generally been documented in the relevant reports to CITES Standing Committees and CoPs and in CITES Information Documents in response to interventions by CITES Parties. The key methodological changes are:

- The model to estimate seizure weights from number of pieces no longer includes a time trend (change first included in analysis for SC66)
- The model to estimate the weight per seizure (for raw or worked ivory) no longer includes a time trend (CoP17)
- Revised criterion to select the countries to include in the analysis (CoP17)
- Derivation of a new variable – the Trade Chain Index – for consideration as a proxy variable for the seizure rate (CoP18)
- Five rather than six ivory classes are used to model the Transaction Index (CoP17)
- Accounting for multiple countries of origin (CoP17)
- A revised set of variables for use in the cluster analysis (CoP18)
- Development of a methodology to explore the robustness of the cluster analysis (CoP18)
- Bayesian modelling is carried out using JAGS rather than WinBUGS (CoP17)

A note on the R scripts

The R scripts for the ETIS analysis were originally developed under Darwin Initiative project 17-020 over the period 2009 – 2013. Those R scripts were written by RW Burn and FM Underwood. Since then, the ETIS analysis has been revised, as described above, and the R code improved and updated.

Because of the nature of ETIS funding, all improvements to date have been made within an analysis cycle. Thus the R code has been adapted iteratively over time. All attempts are made to update scripts in a consistent manner but some inconsistencies between scripts remain – for example in some scripts weights are identified as `wts` and other times `wgts`.

A full review and revision of the ETIS code would include streamlining some processes and scripts. For example: although much of the data management now uses functions from the R package `tidyverse`, this is not uniformly the case because of the time required to revise and thoroughly check all such revisions; some scripts are already functions and others could be combined (because they do similar tasks) and revised into functions to form an R package; RMarkdown could be used to create living documents that link existing scripts together.

The set of R scripts provided here show, to all intents and purposes, the working R code that was used to transform the ETIS data into the outputs for the CoP18 report. There has been some minimal editing of the R scripts. Specifically: more detailed descriptions of each script have been provided at the top of each R script; file names have been edited - the iterative process of development led to many versions of the file and long file names to track these changes; exploratory work to develop, for example, new variables (Trade Chain Index) or new methods (sensitivity analysis) have been removed because they do not contribute to the scripts needed to reproduce the analysis.

Data and Definitions

The ETIS data are held in a PostgreSQL relational database.

Basic data

For every seizure in the ETIS database the minimal information for its inclusion in the analysis is that it describes:

- The year in which the seizure was made
- The country that made the seizure – the country of discovery (which can be any country in the trade chain or, sometimes, the only country known in a trade chain)
- Quantity – either weight (kgs) or number of pieces – of raw and/or worked ivory
- Is status 3 or above[1] – indicating that the record has been validated by ETIS

For most analytical purposes seizures that contain both raw and worked ivory are counted as two separate seizures.

[1] There are eight categories (-2, -1, 0, 1, …, 5) that denote the status of ETIS records; each record is assigned a status and only status 3, 4 or 5 records, i.e. those which have been validated, are eligible for analysis.

Countries on the trade chain

Other information about the seizure that is used in the analysis concerns other countries that are listed on the trade chain. For any seizure the following information may also be provided:

- Countries of origin – where the ivory originated from and the proportion of ivory in the seizure that comes from this country
- Countries of export – where the shipment was, or was to be, exported (or re-exported) from
- Countries of transit – where the shipment passed, or was passing through, between export and destination. In each case it is noted whether or not the country falls within the chain of custody; in other words, the consignment was in the legal jurisdiction of the country for a period of time and could have been seized if detected[2]. If it was, it is marked as having a seizure opportunity.
- Country of destination – the final destination of the shipment

This information is not always available. Sometimes the same country can have multiple roles in the trade chain (for example, it could be the country of origin, export and discovery, or the country of discovery and destination).

Country Codes

All countries have a unique identifier code and are also described using ISO 3166-1 alpha-2 country codes. In the database they are recorded in capitals – e.g. US, KE. Country codes are converted to lower case within R. This makes it easier to track Namibia (ISO code NA – which means Not Available in R).

In some cases a pseudo country is listed – for example a country of origin marked as XF means that it came from a country in Africa but the particular country is not known. These pseudo-countries are excluded when analysis is carried out at country level.

Seizures in and out

For analysis, a seizure that is made in a particular country is listed as a ***seizure in*** for that country (the country of discovery). A seizure is listed as a ***seizure out*** for countries on the trade chain that are not the country of discovery. For example, consider the following hypothetical shipment:

> The shipment of one tonne of raw ivory was seized in SG. The ivory was from UG (20%) and KE (80%). It was exported from KE and then passed through AE where it was part of the chain of custody and MY where it was not part of the chain of custody before being seized in SG. The destination of the shipment was CN.

There are two ways of counting *seizures out* depending on what is being measured.

[2] For example, a plane carrying illegal ivory as air freight may refuel en route in country X. Because the cargo is not offloaded there is no seizure opportunity and thus country X is not considered to be part of the chain of custody.

- **To estimate a measure of law enforcement and effectiveness** the seizure is counted as a *seizure out* only for countries that had an opportunity to make a seizure. Currently this is all countries listed on the trade chain prior to the country of destination[3]. To do this, in several R scripts the flag `ctry_dest` is set to `FALSE`. In the hypothetical example for the countries UG, KE and AE this seizure would count as a *seizure out*. It does not count as a *seizure out* for CN because there was no chance for law enforcement activity in CN because the consignment never reached CN.

- **To estimate trade flows** the seizure is counted as a *seizure out* for all countries on the trade chain (that are not the country of discovery) because the intent is to capture the intended known pathway of each transaction. The flag `ctry_dest` would be set to `TRUE`. In the hypothetical example it would count as a *seizure out* for UG, KE, AE **and** CN.

In neither case would this seizure count as a *seizure out* for MY because it was never part of the chain of custody and so the consignment was never officially within the jurisdiction of MY. In general, transit countries with no seizure opportunity are not included in any analysis unless specifically noted.

Weights in and out

In the cluster analysis ***weights in*** and ***weights out*** are also calculated for each country. A seizure's weight counts towards the *weight in* total for the country of discovery. So in the hypothetical example the shipment contributes a *weight in* of one tonne to SG.

A seizure's weight contributes to the *weights out* total for the countries in which it counts as a *seizure out*[4]. However, for countries which are only listed as a country of origin, only the portion of the ivory that comes from that country is included. In the hypothetical example above, a *weight out* of 200kg (rather than one tonne) is used for UG. For all other countries – KE, AE and CN – a weight of one tonne is used. Although KE was the country of origin for 800kg of the ivory, the whole shipment including the 200kg from UG was exported from KE. So for this shipment the *weight out* for KE is one tonne.

Analysis

A local copy of the database is provided to the analyst prior to the analysis starting. This copy is used for the analysis so any new seizure records that come into the database after this time are not included. The copy includes the seizure records, related tables and the subsidiary data.

[3] Generally speaking, the records in ETIS have, at most, listed a country of destination after the country of discovery. However, in recent years, trade chains have sometimes become longer following investigations and have occasionally listed both transit and destination countries after the country of discovery. A future planned update to the ETIS database structure will make it possible to account for this in the analysis.

[4] In the R code the *weights out* are calculated using both the Law Enforcement and Trade Flow definition of *seizures out* but as described in Stage 4 – Cluster Analysis the *weights out* that is used is based on the Trade Flow definition.

Updates to seizure records are only carried out in extreme cases when a correction to a seizure is provided between the date of the local copy being provided and the data is signed off after further data checks. The stages at where these changes can be made are documented and all changes are recorded in relevant R scripts.

Access to the local copy of the database is managed using the R package `RPostgreSQL` and the file **`PG settings.R`** contains relevant information.

All of the ETIS analysis is carried out using R. R scripts are stored in the directory `C:/ETIS/analysis/R code` and outputs are stored in a number of different directories (partly to do with backing up processes). The main directory is `C:/ETIS/analysis/Processed Data`

There are four principle stages comprising an analysis of the ETIS data for a CoP report. Those stages are:

Stage 1: Preparing the data for analysis;

Stage 2: Deriving the Transaction Index (TI) to describe smoothed relative trends in illegal ivory trade transactions;

Stage 3: Deriving the Weight Index (WI) to describe smoothed relative trends in the quantity of illegal ivory in trade;

Stage 4: Undertaking a cluster analysis to comparatively assess the role of countries in the illegal ivory trade for a specific time period.

The work in each of these stages consists of a series of steps that can be generically described as **data preparation**, **statistical modelling** and the production of **model outputs**. Figure 1 shows the main steps in each of the first three stages and how they link together.

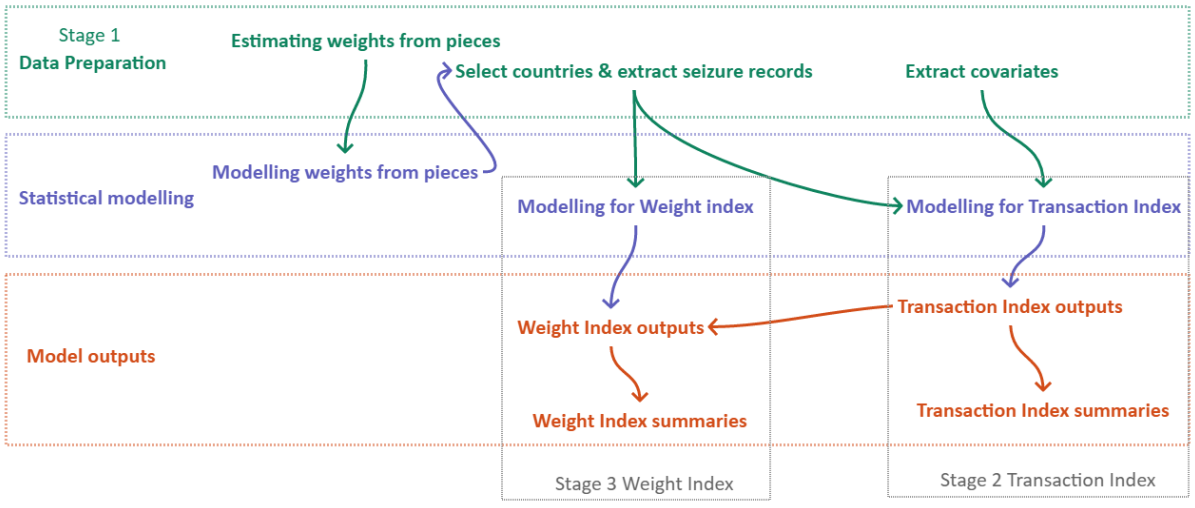

Figure 1: Main stages in converting seizures data into Transaction and Weight Indices

More details of each stage are described below. For each stage a figure describes how the different R scripts link together. In these diagrams the following codes are used:

- `clusternm` – indicates the specific set of variables used for the cluster analysis
- `ddmmyyyy` – day, month, year
- `dest_-` which definition of *seizures out* is used
- `modnm` – which TI model is used
- `pms` – model parameters, rather than listing them all separately
- `size_` – used when only seizures with a weight of at least the limit defined are included
- `y1_y2` – indicates year range
- `*` - denotes that there are several options for this part of the file name. This means that the R script that produced this output must be run more than once with different starting inputs.

Analysis of the illegal ivory trade using ETIS data is an ongoing iterative process. Each analysis builds on the previous analyses and uses the existing R scripts as much as possible. In many cases the start of each R script includes variables whose values can be specified; for example, the years to include in the analysis. These values can then be updated in each successive analysis.

Stage 1: Data Preparation

The main data used for this analysis are ivory seizure records and subsidiary data corresponding to seizures made in the period 2008 – 2017 that were reported to ETIS before 6 June 2018. Additional data from other years were used at various times as detailed in the methodological notes below.

Step 1.1 Estimating weights from pieces.

For the TI analysis, the data needed are the number of ivory seizure records for each country and year in each of six ivory classes - raw and worked ivory for each of three weight categories – small (less than 10kg), medium (10kg to less than 100kg), large (at least 100kg). Not all seizures report the seizure weight (they may only report the number of pieces) and so in these cases the weights need to be estimated.

Missing weights also need to be estimated so that: (a) total *weight in* and *weight out* can be calculated in the cluster analysis; and (b) the ETIS team can produce weight summaries where needed for reporting purposes outside of the CoP cycle.

Missing weights are estimated by building a model using data from seizure records where both the number of pieces and seizure weight are recorded. The model is revised and updated when preparing an analysis for a CoP report. The current procedure is that in between CoPs, for example for analyses for Standing Committee reports or when the ETIS team need to make other summary reports, the same model is used for estimation. Then, with each successive analysis for a CoP report, a new model is fitted using existing data plus the new data that has arrived since the last CoP.

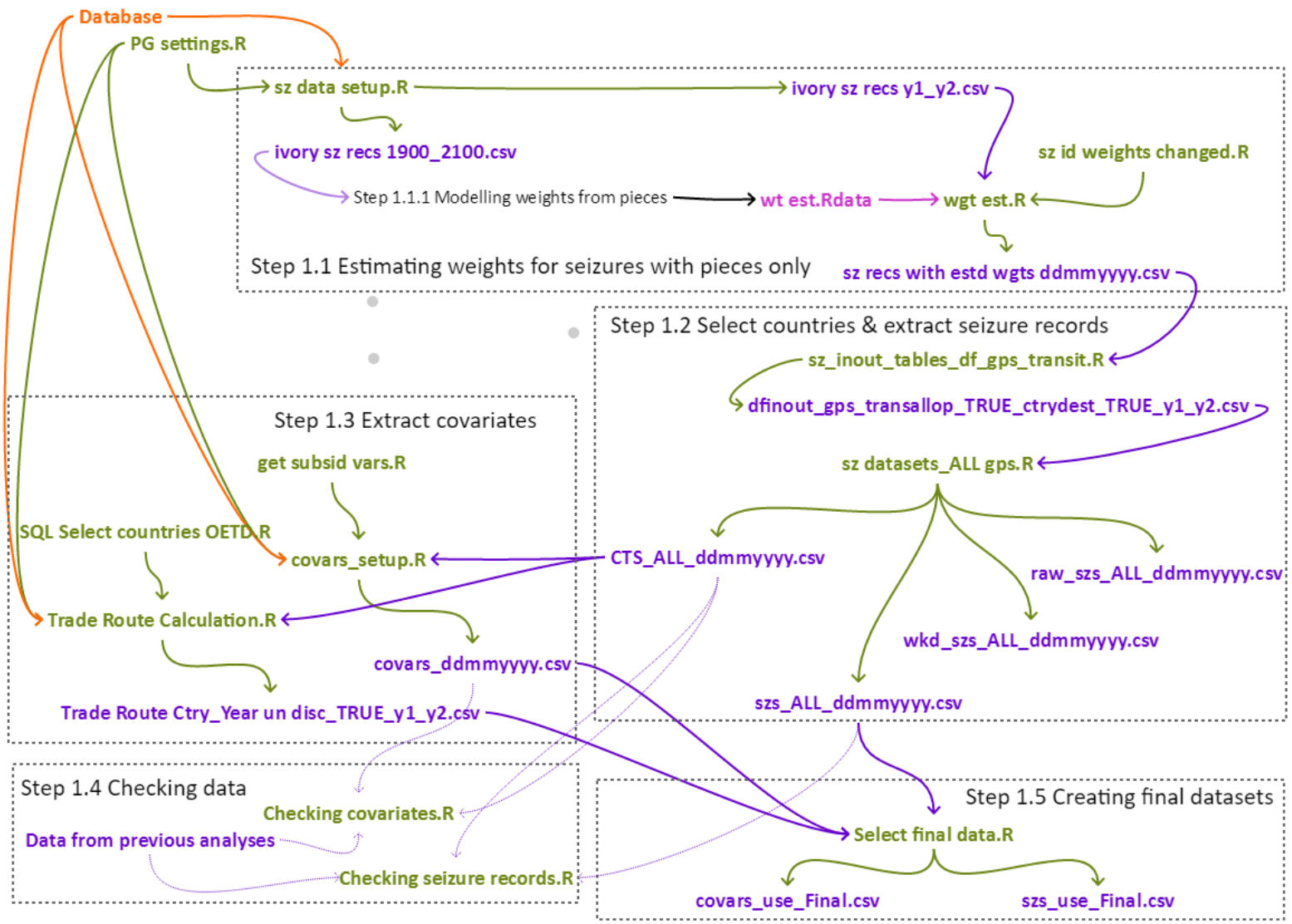

Figure 2: Stage 1 – preparing data for analysis.

All estimated weights in the seizures database are therefore revised each time the model is updated. The data used to build the model to estimate weight from pieces therefore comes from a larger set of years than those used for the CoP analysis. The file `sz data setup.R` is used to extract all data in the database. Relevant years are selected in the modelling process and the modelling process is described in Step 1.1.1 below and illustrated by Figure 3. The final output is the file `wt est.Rdata` that contains the details of the weight estimation model.

The file `wgt est.R` uses the weight estimation model (`wt est.Rdata`) to estimate the seizure weight for seizures where only the number of pieces are recorded. The file also converts all (known and estimated) weights to Raw Ivory Equivalent (RIE) weights[5].

The seizures to which the weight estimation model are applied are seizures made in the years to be considered for the analysis. The relevant seizures are extracted from the database using the file `sz data setup.R`.

The file `sz id weights changed.R` contains the list of seizures where the weight was provided but they were subsequently amended based on revised information. These adjustments are also made in `wgt est.R`. The weight of seizure record 109137 was changed from 512kg (the weight provided prior to the cut-off date) to 1286 kg (the revised weight which was provided shortly after the cut-off date). Given the size of this shipment and its significance for some countries, this correction was considered reasonable. The weights of a few other small shipments were also re-estimated[6].

Step 1.1.1: Modelling weights from pieces

The model follows that described in Burn and Underwood (2013) except that the year term is no longer included[7]. The model uses seizures data where both the weight and number of pieces are recorded. After initial model fitting some odd values are identified, examined and removed (`sz id remove weight pieces model.R`). Further fitting identified outliers and influential points some of which are removed. The final recommendations for modelling are then saved in `Raw wt from pieces.Rdata` and `Worked wt from pieces.Rdata`.

Separate models are fitted for raw and worked weights (`wgt est models.R`) and the model form and parameters are stored in the file `wt est.Rdata`.

[5] This means that worked ivory weights are divided by 0.7 (Milliken et al., 2002).

[6] Seizures 101114 and 107677 were also adjusted. For seizure 101114 the total weight of the seizure was given, the worked weight was estimated from the number of items seized and the raw weight calculated as the difference between this worked weight and the total seizure weight that was provided. For seizure 107677 the weight was given originally as 0.405kg for 71 pieces but the note for this seizure was that there was one piece of 0.405kg and another 70 pieces. These 70 were estimated using the model to give a total weight of 5.95kg. In these two cases the revised weights kept the seizures in the same ivory class as previously.

[7] One reason for this is because the current model will be used to estimate weights of seizures made in 2018 – 2021. If the model included a year term these estimated weights would be based on extrapolation.

Step 1.1.1 Modellng weights from pieces

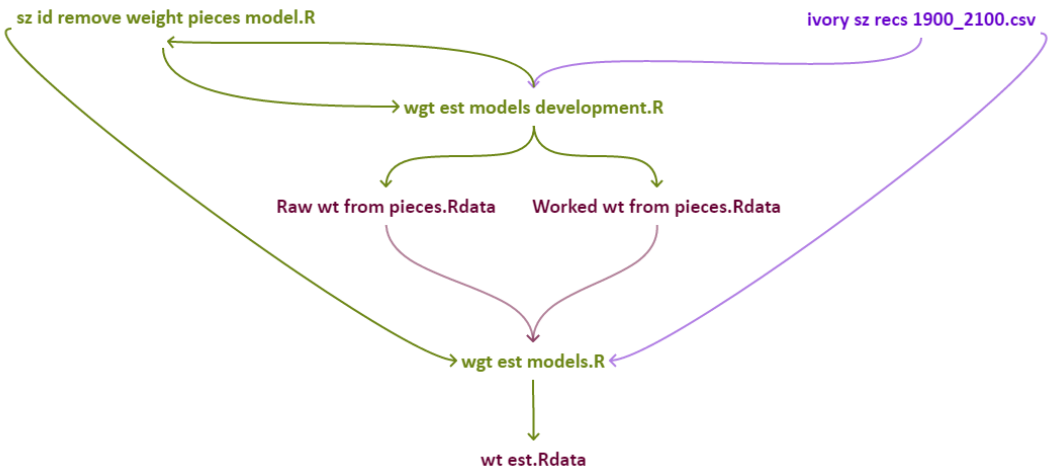

Figure 3: Step 1.1.1 Modelling seizure weights from number of pieces

Step 1.2 Select countries and extract seizure records

Not all CITES Parties are included in the analysis[8]. The strategy to select countries uses the same approach as in the CoP17 report. Countries were included if, over the period 2008 – 2017, they scored at least 100 using the following formula[9]:

1 * Number of small seizures (less than 10kg)

+ 10 * Number of medium seizures (between 10kg & less than 100kg)

+ 100 * Number of large seizures (at least 100kg)

For the purposes of the selection criterion only, all countries listed in the trade chain, whether or not there was a seizure opportunity, are implicated in a seizure. The file `sz_inout_tables_df_gps_transit.R` extracts the relevant summaries.

The file sz datasets_ALL gps.R calculates the score for each country and the score is based on seizures made in the last ten (complete) years; in this case from 2008 - 2017. Once the countries have been identified the number of seizures in each ivory class made by each selected country in each year is calculated.

[8] Modelling is challenging when countries are included that show very little evidence of any ivory trade activity – either because they don't make, or are not implicated in, seizures of any size. More stable modelling results are achieved without their inclusion.

[9] For example, if a destination country is implicated in a single seizure of over 100kg then it would score 100 and be included. If a country was implicated in nine seizures of 20 kg each and also made 12 seizures of 5kg each, this country would score 102 and be included in the analysis.

Step 1.3 Extract covariates

These are the covariates to be considered as candidate variables for modelling the seizure and reporting rates. Data for most of the covariates are extracted directly from the ETIS subsidiary database. Some of the data extracted from the subsidiary database requires further processing – specifically the CITES reporting score and the Data Collection Score. For some variables there was no recent information and so they were not considered in the analysis. Data extraction, processing and exclusion are carried out in `covars_setup.R`

For CoP18 a new variable, the Trade Chain Index (TCI), was derived to account for the fact that countries that mainly act as destination countries will have a higher LE ratio than countries that are not mainly destination countries. The TCI is calculated separately in `Trade Route Calculation.R`.

The TCI for a country in a particular year is calculated as the log of the ratio of two scores:

> Destination Score: The number of times that the country is listed as a country of destination divided by the number of seizures in which a country of destination is given. This is calculated separately for seizures of raw and worked ivory and the average of these two values is taken. There are over three times as many worked ivory seizures as raw ivory seizures records which report more than just the country of discovery. Without calculating the score separately for raw and worked ivory the index would be biased towards the role that a country plays in the trade chain for worked ivory.
>
> Non-destination Score: The average of the equivalent score for origin, export and transit countries. For example, the Export Score is the number of times that the country is listed as a country of export divided by the number of seizures in which a country of export is provided. Again, this is calculated separately for raw and worked ivory and the average of the raw and worked scores is taken.

Step 1.4 Data checking

Using the file `Checking seizure records.R`, graphical and numerical summaries of the seizures data (in the file `szs_ALL_ddmmyyyy.csv`) are compared to similar summaries of the data used in the previous analysis – in this case for Standing Committee 69.

Using the file `Checking covariates.R` a similar process is carried out to compare some of the derived covariates (in the file `covars_ddmmyyyy.csv`) with those used in the previous analysis.

These comparisons are returned to the ETIS team for checking to ensure that any differences in records from past years can be explained. If there are any problems, these are resolved.

Once the ETIS administrators are happy with the data, it is signed off for analysis.

Step 1.5 Creating final datasets

Once signed off, a final set of edits to tidy up the data are carried out to produce summaries of seizure records (`szs_use_Final.csv` used for the TI – step 2 - and WI – step 3) and

covariates (`covars_use_Final` used for the TI – step 2 - and cluster analysis – step 4). In this analysis, South Sudan meets the selection criteria, but because the country has only been recognised since 2011 onwards, no data (not even zeros or missing values) for the years 2008 to 2010 are included in the analysis. Portugal is excluded from the analysis because the covariates are currently under review; it was not implicated in any large-scale seizures over this time period.

Stage 2: Transaction Index

This uses the seizures data `szs_use_Final.csv` and covariates `covars_use_Final` created in Step 1.5.

Step 2.1 TI Model fitting

The statistical modelling is carried out in the file `sz_JAGS_model_Final.R`. The heart of the model is this equation describing the expected value of y_{ikt}:

$$E[y_{ikt}] = \lambda_{ikt}\phi_{it}\theta_{it}$$

where:

y_{ikt} is the number of seizures of ivory class k made by country i in year t and has a negative binomial distribution

λ_{ikt} is the expected number of unobserved class k ivory transactions in country i and year t – this is a function of time

ϕ_{it} is the seizure rate for country i in year t –a function of covariates

θ_{it} the reporting rate for country i in year t – a function of covariates

The final model for CoP18 had four polynomial terms for the temporal smoothing and the following covariates:

- seizure rate: the one-year lagged Law Enforcement ratio[10](LE1) and the Trade Chain Index
- reporting rate: Data Collection Score and the CITES Reporting Score

[10] The Law Enforcement Ratio is $seizures\ in/(seizures\ in + seizures\ out)$

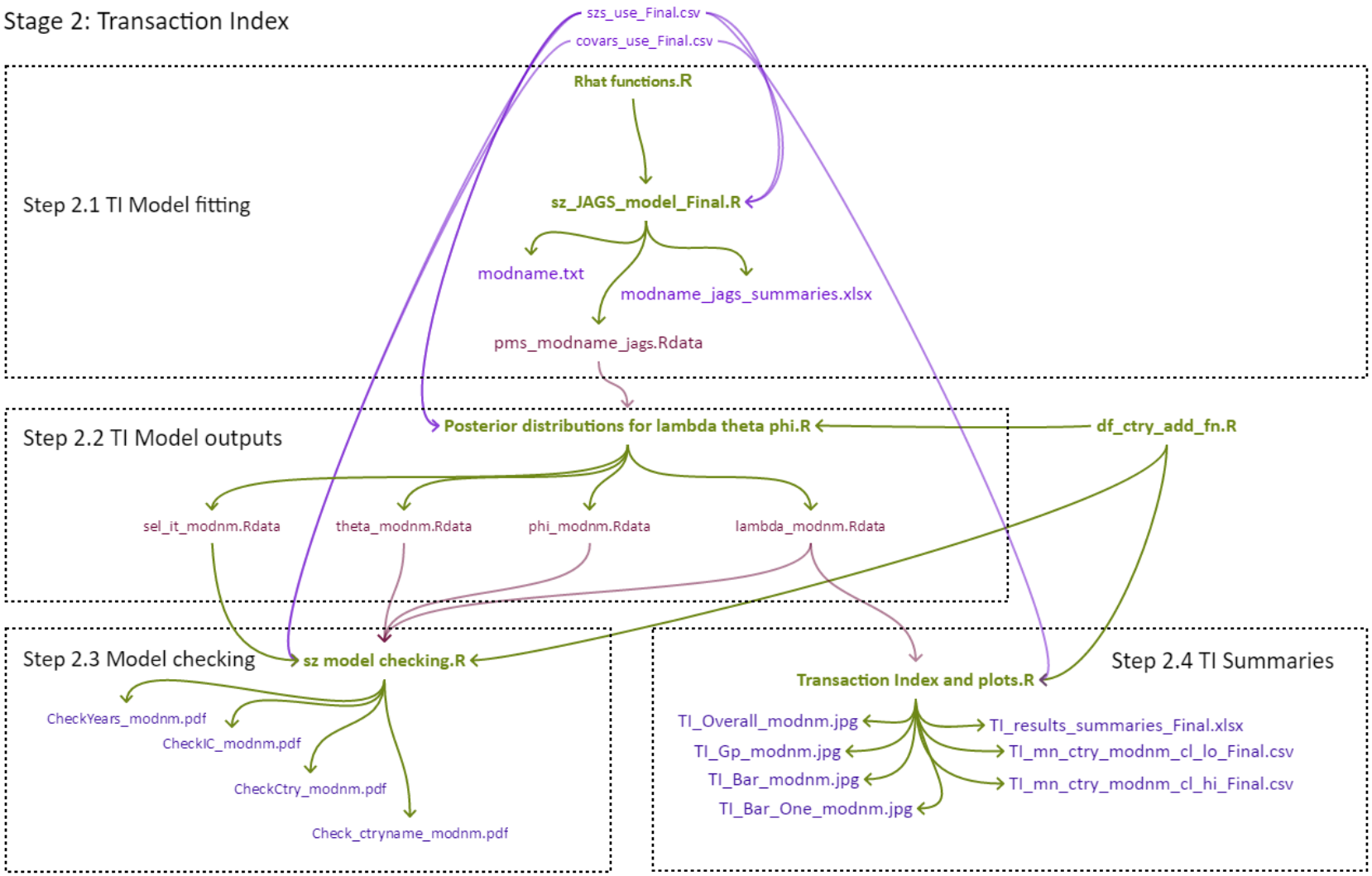

Figure 4: Stage 2 – calculating the Transaction Index

The script also includes the code used to check for convergence and inspect parameters values. This R script can be easily adapted to allow other covariates and polynomial terms to be included in, or excluded from, the model. Model comparisons were based mainly on credible intervals and model checking. Model checking is carried out in Step 2.3 and requires draws from the posterior predictive distribution of the y's. These are calculated using the TI model outputs (created in Step 2.2)

Step 2.2 TI Model outputs

Posterior distributions of the seizure rate, ϕ_{it}, reporting rate,θ_{it}, and ivory transactions λ_{ikt}s are needed to calculate the posterior predictive distributions of the y's for model checking (Step 2.3) and for TI summaries (Step 2.4). Posterior predictive distributions of these parameters are calculated using 10,000 draws from the posterior distributions of model parameters.

Step 2.3 Model checking

Model checking (`sz model checking.R`) calculates draws from the posterior predictive distribution of the y's and summarises them by, for example, country, year and ivory class. The mean, median and credible intervals are then compared to the equivalent data summaries. As well as 90% intervals, 50% intervals are also used, to clearly explore model behaviour.

Step 2.4 TI Summaries

The Transaction Index outputs used for reporting purposes are based on summaries of the posterior distributions of the λ_{ikt}s (Step 2.2). The λs are summarised by country, ivory class and year.

Stage 3: Weight Index

This uses the seizures records with estimated weights `sz recs with estd wgts yr1_yr2.csv` created in Step 1.1 (although only seizures with known weights are used in this part of the process) and the λs created in Step 2.2.

Step 3.1 WI model fitting

Separate models are fitted to estimate the distribution of weights of individual seizures for raw (`raw wt_dist_jags.R`) and worked (`wkd wt_dist_jags.R`) ivory separately. Only seizures with a known rather than estimated weight are used. Weights are truncated so that they can be no more than ten tonnes – currently no shipments of more than ten tonnes have been seized. The distribution of raw or worked weights is a T-distribution with priors on the mean, variance and degrees of freedom.

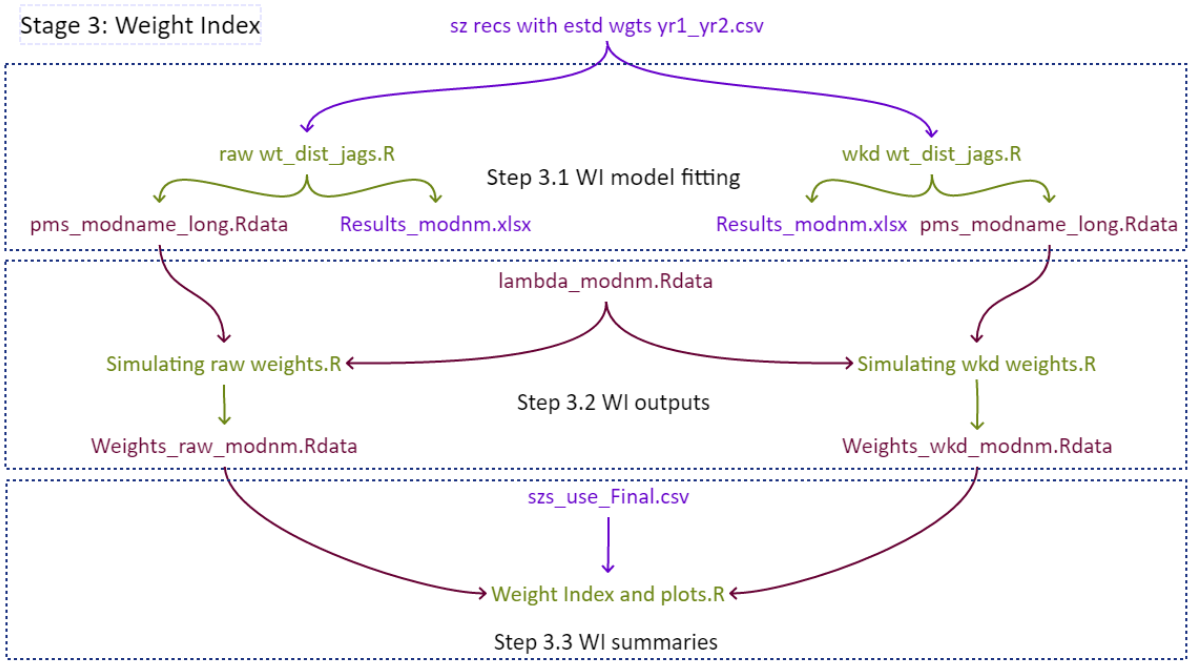

Figure 5: Stage 3 – calculating the Weight Index

Step 3.2 WI outputs

For each of the 10,000 draws from the posterior distributions of the parameters created in Step 3.1, and each of the 10,000 draws from the posterior distributions of the λ's (from Step 2.2), a weight is simulated for every transaction of raw (`Simulating raw weights.R`) and worked (`Simulating wkd weights.R`) ivory.

Step 3.3 WI summaries

The WI summaries used for reporting are based on summaries (mean and quantiles) of the 10,000 draws from the posterior distribution of the total weight of ivory by ivory class and year that were calculated and saved in Step 3.2.

Stage 4 Cluster analysis

The underlying methodology for the cluster analysis is the same as in previous analyses. The sensitivity analysis is new. The sensitivity analysis uses the posterior distributions of the bias-adjusted variables to explore the robustness of the cluster analysis.

The variables used in the most recent cluster analysis are described in the CoP report. In brief, they are the following summary variables for the period 2015-2017 for all countries selected in Step 1.2:

- Transaction Index for each ivory class
- Seizures out (less than 500kg and at least 500kg) – using LE definition of *seizures out*
- Weights in (less than 500kg and at least 500kg)
- Weights out (less than 500kg and at least 500kg) – using Trade Flow definition of *seizures out*

All variables are bias-adjusted as described below.

Note that in previous cluster analyses (for CoP16 and CoP17) the weight limit was one tonne rather than 500kg. For CoP18 there were not that many seizures of at least one tonne and so it was decided to use a cut-off of 500kg instead. Although many file and variable names retain the 1T suffix the calculations were based on the 500kg value.

Bias adjustment

The Transaction Index (TI) variables – the λ_{ikt}s – are already bias adjusted values. Other bias-adjusted variables are created by summarising bias-adjusted values of individual seizures. The bias-adjusted value of a seizure is:

- the seizure weight multiplied by the bias-adjustment factor - for calculating the *weights in* and *out*
- (one multiplied by) the bias-adjustment factor - for calculating the *seizures out* variable

The bias-adjustment factor is the reciprocal of the product of the country of discovery's seizure (ϕ) and reporting (θ) rate for that year. That is for seizures discovered in country i in year t the bias adjustment factor is:

$$1/\phi_{it}\theta_{it}$$

Step 4.0 Data Preparation for bias adjustment

Some of the variables used in the cluster analysis are summaries of *seizures out* or *weights out*. Some seizures may count as a *seizure out* for a country that is to be included in the cluster analysis but the country of discovery (which is used to calculate the bias-adjustment factor for that seizure) may be one of the countries excluded from the modelling in Step 1.2. Because the country of discovery was excluded from the modelling, the posterior distributions of the seizure and reporting rates and therefore also the bias adjustment factor have not yet been calculated

To obtain these posterior distributions the values of the covariates used to calculate the seizure and reporting rate are needed for all countries that made a seizure.

These covariates are extracted by carrying out similar steps to those in Step 1.3. Specifically the script `covar setup_ALL_cluster.R` extracts and processes covariates that are held in the subsidiary data (lagged LE ratio, data collection score and CITES reporting score). The script `Trade Route Calculation_All.R` calculates the TCI.

The script `Select covars cluster ALL.R` brings these covariates together ready for calculating the bias adjustment factor.

Figure 6: Stage 4 – Cluster analysis and sensitivity analysis

Step 4.1 Bias adjustment

Using the covariates (from Step 4.0) and the posterior distributions of the parameters (from Step 2.2) required to calculate the seizure and reporting rates, the posterior distribution of the bias adjustment factor is calculated for all countries that made seizures (`Sx x rep rate for all countries for cluster analysis Final.R`).

Given these values the posterior distributions of adjusted *seizures* and *weights in* and *out* for each country in each year are calculated (`sz_inout_tables_adj_sims.R`). Some of the variables used in the cluster analysis (*seizures out*) use the law enforcement definition of *seizures out*. Others (*weights out*) use the Trade Flow definition of *seizures out*. Furthermore, the variables are calculated separately for shipments of at least 500kg and shipments of less than 500kg. Hence this R script is run four times: country of destination set to TRUE or FALSE and in each case for seizures of all sizes and seizures of at least 500kg to produce four sets of outputs.

For one seizure – 109462 – country 188 was a transit country but the shipment was seized prior to it reaching that country. It was marked as being in the chain of custody but it needed to be treated like a country of destination when calculating seizures out. Note that this will lead to a future modification in the database to account for this more directly in the future.

Step 4.2 Accounting for multiple countries of origin

In addition to the bias adjustment it is also necessary to account for multiple countries of origin when calculating *weights* and *seizures out*. As described in the Data and Definitions Section of this document, when a country is listed as one of several countries of origin for a shipment (and nowhere else on the trade chain) then the *weight out* is only the portion of the shipment that comes from that country.

The database is queried to extract information about the trade route for all seizures with more than one country of origin (`mult_ctries.R`). The file `Multiple mentions all wtlimit sims.R` uses this information to correct the bias-adjusted *weights out* and *seizures out* for countries which are only one of several countries of origin. The *seizures out* correction is required because in some cases a seizure may be over 500kg but an individual country may contribute less than 500kg. This script is run four times to use the four outputs from Step 4.1.

Step 4.3 Cluster analysis and sensitivity analysis

The file `Creating variables for cluster analysis sims Final.R` creates posterior distributions of all variables to be used in the cluster analysis using the outputs from Step 4.2 and Step 2.2.

The cluster analysis and sensitivity analysis are carried out in the same file `cluster analysis sims Final 500.R` The cluster analysis is carried out using the posterior mean values of each variable. A PCA and biplot are also provided to help the ETIS team with understanding the data. The number of groups to be used for descriptive purposes is decided and summary statistics of the variables for the different groups calculated.

The sensitivity analysis repeats the cluster analysis using every draw from the posterior distribution of the variables. Using the same number of groups as specified above, the countries that share group membership are saved for each draw. The proportion of times that each country occurs with every other country is calculated. These are provided in matrix format (graphically and numerically) to help the ETIS team interpret the cluster analysis.

License

Authorship of R scripts

The R scripts were written by Robert W Burn and Fiona M Underwood and are maintained by Fiona M Underwood (fiona@fmunderwood.com).

Citation

The GitHub repository ETIS_CITESReporting_RCode and all releases are available here https://github.com/fmunderwood/ETIS_CITESReporting_RCode. The repository has been given a DOI (Digital Object Identifier) via the data archiver Zenodo. The GitHub repository should therefore be cited as:

> Underwood FM, Burn RW (2019) fmunderwood/ETIS_CITESReporting_RCode, Zenodo. https://doi.org/10.5281/zenodo.3334771

This DOI represents all versions of the repository and will always resolve to the latest one. The specific DOI for any particular version – for example for v.CoP18 – is also listed on this page (https://doi.org/10.5281/zenodo.3334771) so that specific files from these versions, or the version as a whole can be cited.

For example, this document can be cited using the general DOI for this repository – as given on page 2, or using the specific DOI for the CoP18 release.

Funding

The original development of the R code was funded by the UK Government's Darwin Initiative project 17-020 Enhancing the Elephant Trade Information System to Guide CITES policy whilst the authors were working at the University of Reading. Development of the code for future

analyses has been funded by the European Union and the United States Fish and Wildlife Service's African Elephant Conservation Fund. The governments of Belgium, the Netherlands and the European Union funded the work to make the code available on-line.

References

Burn RW, Underwood FM (2013). A new statistical modelling framework to interpret ivory seizures data: A Technical Report describing the new modelling framework for analysing seizures data from the Elephant Trade Information System. Mathematics Report series (1/2013) Department of Mathematics and Statistics, University of Reading, UK

Milliken T, Underwood FM, Burn RW, Sangalakula L (2018). The Elephant Trade Information System (ETIS) and the Illicit Trade in Ivory: A report to the 18th meeting of the Conference of the Parties to CITES. CITES Secretariat, Geneva, Switzerland. CoP18 Doc. 69.3 Annex 1

Milliken T, Burn RW, Underwood FM, Sangalakula L (2013). Monitoring the illegal trade in ivory and other elephant specimens. CITES Secretariat, Geneva, Switzerland. CoP16 Doc. 53.2.2 (Rev.1)

Milliken, T., Burn, R.W. and Sangalakula, L. (2002). A report on the status of the Elephant Trade Information System (ETIS) to the 12th meeting of the Conference of the Parties. CoP12 Doc. 34.1 Annexe 1. CITES Secretariat, Geneva, Switzerland

Underwood FM, Burn RW, Milliken T (2013). Dissecting the illegal ivory trade: an analysis of ivory seizures data. PLoS One 8(10): e76539. doi:10.1371/journal.pone.0076539