# Volume and Direction of the Atlantic Slave Trade, 1650-1870: Estimates by Markov Chain Monte Carlo Analysis

Patrick Manning, Yun Zhang, Bowen Yi

**Abstract**

This article presents methods and results in the application of the Markov Chain Monte Carlo analysis to a problem in missing data. The data used here are The Atlantic Slave Trade Database (tastd), 2010 version, available online. The article begins with background to the Bayesian statistical framework, Markov chains, and Monte Carlo methods, as compared with the frequentist statistical framework, still more widely used in economic (and demographic?) analyses.. It then describes the data, their analysis, the results, and a discussion of their strengths and weaknesses. The results provide a new estimate of the volume of African embarkations and American arrivals in the transatlantic slave trade for the period from 1650 to 1870, by decade, for eleven African regions of embarkation and seven American and European regions of arrival. These results are compared with earlier estimates of Atlantic slave trade volume by frequentist methods.

*Keywords:* Science, Publication, Complicated

The volume and direction of the Atlantic slave trade has long been a subject of importance and controversy. Detailed research has explored the numbers embarked on slave vessels at various African ports, the numbers who lost their lives in the course of the voyages of two months or more across the Atlantic, and the numbers who disembarked at the end of the voyage, mostly in the Americas. While a great deal of information has been gathered, especially within the past fifty years, the problem of missing data remains serious.

Markov Chain Monte Carlo analysis provides a systematic and comprehensive method for estimating missing data. We use it to present estimates

of the total number of captives embarked, by decade (from 1650 to 1870), for eleven regions of the African coast. In the same analysis, we estimate the number of captives who arrived in the same decades for six American regions and for Europe. In addition, the article discusses two other procedures for estimating the levels of African slave trade: one based on the combination of multiple methods carried out by our group; the other being the procedures used by David Eltis and his colleagues.

## 1. Studies of Atlantic Slave Trade: from Synthesis to Voyage-Based Data

Two major steps forward, and many smaller steps, have characterized the quantitative study of the Atlantic slave trade. First, Philip Curtins 1969 book, The Atlantic Slave Trade: A Census, presented comprehensive estimates of the volume of slave trade. He offered a total of 9.5 million persons–for the number of arrivals or disembarkations of captive Africans in the Americas from the fifteenth century into the nineteenth century. Curtins research was carried out through secondary works and with a wide range of methods. His estimated total, which was smaller than many expected, brought an outpouring of research into primary documents. This research brought distinction among such related variables as numbers of captures, embarkations, arrivals, and mortality, including distinction among migration in Africa, on the Atlantic, and in the Americas. Joseph Inikori sought to show that the estimate should be increased, and a debate ensued.(Inikori; Curtin) Increasingly, the analysis focused on documentation of Atlantic slaving voyages, though there remained numerous other methods for estimating aspects of the slave trade. Over twenty years of debate, the estimated volume of the slave trade had crept up by a million or so, and clearer distinctions were made on estimates of embarkations, arrivals, and mortality at various stages in the trade.(Lovejoy - numbers)

The second major step in documenting the Atlantic slave trade was a project to combine the many separate research projects on slaving voyages into a comprehensive dataset. David Eltis led this work, in association with several colleagues, and sustained it over more than twenty years. The initial report of this research was a 1999 CD-ROM with data on some 27,000 slave voyages, organized according to a systematic codebook.(Eltis et al 1999) This aggregative research clarified distinctions between (1) known voyages

and unknown voyages and (2) between known numbers of captive migrants–documented for known voyages–and missing values of captive migrants on known voyages. The succeeding versions of the dataset included large and expanding numbers of variables on the ships, ownership, crew, times and places of the voyages, and details on cargoes of captives. The dataset includes raw data, recorded directly from the sources, and imputed variables, calculated and estimated by the editors from the raw data. Most cells for the full dataset are empty for lack of specific data, yet the full collection of data is immensely rich and varied.

The 1999 CD-ROM was a great step forward and received wide attention.(Manning 1999) At first the editors emphasized the relative completeness of their data, which were indeed effective in identifying the great majority of voyages in the British, Dutch, and French slave trades. With time, however, and as submissions of new data continued, the editorial team recognized that there were many missing voyages–especially in Portuguese, Spanish, and Brazilian vessels–and energetically added the new information to their dataset. By 2008 the editors had published an expanded dataset in an online venue, "Slave Voyages." Since the online publication of the 2008 dataset, a subsequent and expaded 2010 dataset was added to the Slave Voyages website, and further significant modifications were published in 2015. By 2010, the number of voyages included had risen to nearly 35,000.

## 2. Estimates of Missing Migration Data: Alternative Methods

The effort to document the overall volume of the Atlantic slave trade–in time and space and with attention to gender, ethnicity, age, and mortality–has been the principal topic of interest in the overall analysis of the slave trade. The synthesis-based analyses of, most notably by Curtin and Lovejoy, assembled an eclectic range of data and analytical techniques, in which the distinctions among raw data, direct and indirect estimates were necessarily fuzzy. With time and expanded research, voyage-based data became increasingly important in slave-trade studies: the remarkably detailed archival studies of Jean Mettas on the French slaving voyages stood out in this regard.(Mettas)

Eltis, from his earliest work on British reports on slave trade in the 1840s, began and steadily developed a practice of estimating "imputed" numbers of captives for known voyages with missing data on numbers of captives.(Eltis in Gemery and Hogendorn; 1987 book)

The 1999 CD-ROM included raw data on captive migrants and also included "imputed" figures for missing data on captive migrants. Along with the dataset, presented in SPSS format, the authors made estimates of missing values and proposed overall totals of slave trade by region and by time period. Unfortunately, the methods for estimating totals in the slave trade left implicit the techniques for estimating missing values. In addition, the procedure did not include any estimates of error margins or tolerances in the estimates. This eclectic analysis based on projection of frequencies in known data. These methods, using various approximations for various situations, were developed in the course of work on the TransAtlantic Slave Database and the Slave Voyages project.: types of voyages, average number of captives per voyagea frequentist statistical approach. Some results are summarized in the Appendix, but the methods are not described in detail here. These estimates, doubtless drawn up with care, do not take advantage of valuable statistical techniques. [Recent results, based on 2010 data, are shown in the Appendix.]

The purpose of this analysis is to get a second and third opinion on missing data and totals in the volume of Atlantic slave trade, based on two sets of statistical methods. That is, working with a single set of data on known voyages and known captive migration (from the 2010 Slave Voyages dataset), this analysis will present three estimates of missing values and total volume of the transatlantic slave trade: the 2010 Eltis estimates, our estimates through multi-method frequentist analysis, and our estimates through Markov Chair Monte Carlo analysis.

The previous estimates may be labeled as using "frequentist" statistics to estimate missing values.

Of the analyses on which we report here–our "second opinion" and "third opinion"–one relies on a frequentist approach and the other relies on Bayesian statistics.

Bayesian statistics in general.

Distributions used in the analysis: multinomial, Dirichlet, Poisson, binomial, and Gamma. Give graphic representations and other discussion for each distribution

The Markov Chain in general.

Monte Carlo estimates in general. [what is a case? Distinguish MCMC-SI and MCMC-MI] Compare with start of strategy of multiple methods similar but distinct.

Some additional points must be specified regarding the regional character of the data and their analysis. The original data in Slave Voyages are coded geographically by port–the various ports of slave trade on the African coast and the ports of slave disembarkation or arrival in the Americas. For the estimation of missing values in numbers of captives, the ports are grouped into regions, both for the African and American coasts. Regions in the Eltis dataset were based on groupings of ports in the original data. We accepted the regions from Senegambia through the Bight of Benin, as defined in TASTD, but recoded the data for the region from the Bight of Biafra through Angola, reclassifying ports to yield four separate regions rather than the two regions identified by Eltis. This was not for statistical regions but to fit more tightly with the historical literature, which has distinguished the slave trade of the Congo basin from that of Angola more fully than is permitted by the very large West Central Africa region of the Slave Voyages website. We wanted to estimate slave exports for the regions we believe to be most commonly and usefully employed in analyzing export slave trade from the Bight of Biafra through Angola. (The totals should be unchanged by this reclassification of ports, but the distribution among regions is diferent.) Number of cells: 22 decades, within each decade there are 11 regions for embarkation and 7 regions for arrivals. Thus there are 242 cells in the output for embarkation and 154 cells in the output for arrivals. We used data from the following variables of the TASD dataset 2010 version, for the period from 1650 to 1870:

Most of the rest of this paper consists of two detailed descriptions of estimates of missing values: for the multiple-method analysis and for the MCMC analysis. Concluding sections and the Appendix present a comparison of the various estimates and suggestions for further investigation of this issue.

Figure 1: Eigenfunctions are on the left panel; Mean functions are on the right panel

## 3. New Method 1: multiple-method analysis of voyage-based data

In multiple-method analysis, we identify different types of missing data, then project missing data for each type according to an appropriate algorithm. Estimates for a given region within a certain decade

The primary goal is to impute the embarkation count for a given region within a certain decade. The voyages that may contribute to the total embarkation count within a given region and decade can be partitioned into sixteen categories based on the existence of records in the variables: embarkation port, decade, embarkation count and arrival count. The strategy is to compute the total embarkation coun t for each subset of data partitioned by missing type. Table 1 outlines the methods that will be employed in the treatment of each subset. Then the individual contributions are added to the current total and the estimated variance is updated by summing the variances. This implicitly assumes that these totals are independent, but also provides lower bounds of the estimate of variance.

*II. Data Preparation*

In the data preparation stage, we check for known inconsistencies in the dataset. One is the situation in which the arrival count is much greater than the embarkation count. We treat these cases by keeping arrival count but labelling embarkation count as missing. In fact, records with arrival count to embarkation count ratio between 0.9 and 1 are also suspicious. But we leave them unedited for now.

| | Port w/ Decade | Decade w/o Port | Port w/o Decade | Neither |
|---|---|---|---|---|
| Embark. | Direct Estimate | Multin. | Multin. | Multin. |
| Arriv. w/o Embark. | Ratio & SAE | Ratio & Multin. | Ratio & Multin. | Ratio & Multin. |
| Neither | Mean & SAE | Mean & Multin. | Mean & Multin. | Mean & Multin. |

Table 1: Types of estimates shown in the table: SAE - small area estimation, Ratio - ratio estimate, Multin. - propagation through the multinomial model, Mean - using the mean as an estimate

The figure below shows the type of estimates applied for voyages under three sets of embarkation-arrival conditions and four sets of combinations of port and decade. On the vertical axis, the table shows voyages with data for embarkations, voyages with data for arrivals without embarkations, and voyages with neither. On the horizontal axis, the table shows voyages documented for port and decade, for decate without port, for port without decade, and with neither.

*II. Model Specification* Note that all the models we introduce below are built for each decade separately except for the decade assignment model.

*(1) Ratio Estimate*

To illustrate the method, we perform a ratio estimate for the embarkation count($\{E_i\}$) against the arrival count($\{A_i\}$). We would like to estimate the ratio $\beta$ if the assumed model is

$$E_i = \beta A_i + \epsilon_i \tag{1}$$

where $\epsilon_i \sim n(0, \sigma^2)$. The estimate for $\beta$ with variance is given by,

$$\hat{\beta} = \frac{\sum_{i \in T} E_i}{\sum_{i \in T} A_i} \tag{2}$$

$$\widehat{\mathbf{V}(\hat{\beta})} = (1 - \frac{n}{N}) \frac{1}{n \bar{A}_N^2} \frac{\sum_{i \in T} (E_i - \hat{\beta} A_i)^2}{n - 1} \tag{3}$$

where $T$ is our training sample (those with both records for embarkation

and arrival count); $n$ and $N$ are the size of the training sample and all the samples with the arrival count respectively; $\bar{A}_N$ is the mean of the arrival counts amongst all those voyages with the count. The covariance between the individual port ratio estimate and the pooled ratio estimate will be needed in the shrinkage procedure, and is given by,

$$\widehat{Cov(\beta_p, \beta)} = \widehat{\mathbf{V}(\beta_i)} \frac{\sum_{i \in p} A_i}{\sum_{i \in \cup_j p_j} A_i} \tag{4}$$

Notice that the implicit assumption with this method is that the missingness of the embarkation count can be considered to be randomized, hence, making the training set a random sample of all the voyages with the arrival count.

In the cases with both the port and decade records small area estimation will be used, yielding a modified estimate of $\beta$. We are now able to estimate the net embarkation count($E_M$), for voyages containing the AC and missing the EC, with estimated variance by,

$$\hat{E_M} = \hat{\beta} \sum_{i \in M} A_i \tag{5}$$

$$\widehat{\mathbf{V}(E_M)} = (\sum_{i \in M} A_i^2) \widehat{\mathbf{V}(\hat{\beta})} \tag{6}$$

*(2) Mean Estimate*

The mean estimate is a very straightforward standard calculation. For a given embarkation region, $p$, we will denote the mean of the embarkation count in our training sample (those with records for embarkation) as $\bar{E}_p$ , and size of training sample as $|p|$

$$\widehat{\mathbf{V}(\bar{E}_p)} = \frac{\sum_{i \in p} (E_i - \bar{E}_p)^2}{|p| - 1} \frac{1}{|p|} \tag{7}$$

$$\widehat{Cov(\bar{E}_p, \bar{E})} = \widehat{\mathbf{V}(\bar{E}_p)} \frac{|p|}{\sum_j |p_j|} \tag{8}$$

This also undergoes the small area estimate modifications in the sparser cases, with both region and decade records.

*(3) Small Estimate*

We use small area estimation when performing either an estimation with a region mean, or via a ratio estimate. In both cases we use a technique, called composition or more generally shrinkage, by which the estimator is replaced by a linear combination($\hat{\theta_r^C}$) of the regional estimator and the estimator from the pooled regions.

$$\hat{\theta_r^C} = (1-b)\hat{\theta}_r + b\hat{\theta} \tag{9}$$

We will construct an optimal estimate for $b$, $b_r$, based on previously computed estimates for the region estimate variance, $\mathbf{V}(\hat{\theta}_r) = v_r$, the pooled regions estimate variance, $\mathbf{V}(\hat{\theta}) = v$, and the covariance between the region and the pooled region estimates, $cov(\hat{\theta}, \hat{\theta}_r) = c_r$.

$$b_r = \frac{v_r - c_r}{v_r + v - 2c_r + \sigma_B^2} \tag{10}$$

$$\hat{\sigma}_B^2 = \frac{1}{n}(S - \sum_{r=1}^{R} n_r v_r) \tag{11}$$

$$S = \sum_{r=1}^{R} n_r(\hat{\theta}_r - \hat{\theta})^2 \tag{12}$$

where $n_r$ and $n$ are the number of voyages in each region and the total number of voyages respectively. On the rare occasion that this estimator is negative then we conclude that the actual value is very small, in which case, $\hat{\theta}$ should be used to estimate $\theta_r$.

*(4) Region/Decade Assignment Model*

We will call the idea of assigning embarkation counts without region or decade or both values according to a multinomial distribution the "region/decade assignment model". Some mention should be made as to what actual imputations are being performed. The missing regions are not being imputed for given voyages, rather we would like an estimate of each region's contributions to the pool of voyages with missing region records. Consider that the region of departure may take one of several discrete values, which we index $1, 2...I$. Denote the embarkation counts from region i in decade j as $N_{ij}$, and the embarkation counts with missing region in decade j as $n_j$. Assuming

conditional on $n_j$, each region's contributions $n_{1j}, \cdots, n_{Ij}$ follow multinomial distribution with probabilities $q_1, q_2, ...q_I$, then the maximal likelihood estimates of $q_i$ and variances are given by

$$\begin{aligned}
\hat{q}_i &= \frac{N_{ij}}{\sum_{i=1}^{I} N_{ij}} && (13)\\
\hat{n}_{ij} &= n_j \hat{q}_i && (14)\\
\widehat{\mathbf{V}(\hat{q}_i)} &\approx \frac{\hat{q}_i(1-\hat{q}_i)}{\sum_{i=1}^{I} N_{ij}} && (15)\\
\widehat{\mathbf{V}(\hat{n_{ij}})} &\approx \widehat{\mathbf{V}(\hat{n}_j)}\widehat{\mathbf{V}(\hat{q}_i)} + \hat{n_j}^2\widehat{\mathbf{V}(\hat{q}_i)} + \hat{q_i}^2\widehat{\mathbf{V}(\hat{n}_j)} && (16)
\end{aligned}$$

One additional remark is that $N_{ij}$ and $n_j$ are computed using the most recent calculation(imputation) of the embarkation slave count (from those samples with no missing in region and decade, and samples with missing region respectively)

Identical procedures are performed to handle voyages with missing decades.

Consider decade of departure may take one of several discrete values, which we index $1, 2...J$. Denote the embarkation counts from region i in decade j as $N'_{ij}$, and the embarkation counts with missing decade for region i as $n_i$. Assuming conditional on $n_i$, each decade's contributions $n'_{i1}, \cdots, n'_{iJ}$ follow multinomial distribution with probabilities $p_1, q_2, ...p_J$, then the maximal likelihood estimates of $q_j$ and variances are given by

$$\begin{aligned}
\hat{q_j} &= \frac{N'_{ij}}{\sum_{j=1}^{J} N'_{ij}} && (17)\\
\hat{n'}_{ij} &= n_i \hat{q_j} && (18)\\
\widehat{\mathbf{V}(\hat{q_j})} &\approx \frac{\hat{q_j}(1-\hat{q_j})}{\sum_{j=1}^{J} N'_{ij}} && (19)\\
\widehat{\mathbf{V}(\hat{n'_{ij}})} &\approx \widehat{\mathbf{V}(\hat{n}_i)}\widehat{\mathbf{V}(\hat{q_j})} + \hat{n_i}^2\widehat{\mathbf{V}(\hat{q_j})} + \hat{q_j}^2\widehat{\mathbf{V}(\hat{n}_i)} && (20)
\end{aligned}$$

It should be noted that $N'_{ij}$ is not equal to $N_{ij}$ for the reason that $N'_{ij} = N_{ij} + \hat{n_{ij}}$. Equivalently, $N'_{ij}$ can be thought of the updated version of $N_{ij}$. Likewise, if both port and decade are missing then we estimate the percentage of these embarkation counts that contribute to each region and

decade in the analogous way using the most updated imputations.

This estimate of decennial slave embarkations by port applied a complex method with many independent parts. Many modeling assumptions were made, corresponding to each type of estimate. The most important assumption was that of random missingness. While this assumption is reasonable in most situations, the assumption that the missingness of a region is not confounded with the embarkation count of the region is somewhat questionable. The same considerations arise when information on the decade is missing. These are difficult assumptions to test, as the missing value may no be observed at the time they are missing. There are also instances where the procedure uses samples with a certain type of missingness repeatedly to train parameters that will allow the imputation of embarkation counts from samples with other types of missingness. For convenience, we ignore the variance brought by repeated measurements when we estimate the standard error of the imputations.

We apply the Decade assignment model pnly to decades with more than 200 voyages (which accounts for the vast majority of the voyages). This suggests underestimation of the embarkation count for decades with less than 200 voyages. More generally for the Region/Decade Assignment model, we take the missing region model as an example. The assumption that persons who embarked on voyages with missing region records fell independently into any regional category is invalid, because persons who embarked on the same voyage can fall in only one regional category. This can be remedied by using voyage as the unit instead of regionthat is, assuming that voyages with missing port records independently fall into any port category. All the calculations are similar. But to calculate contributions of embarkation count for each port, we need to assume further the number of persons embarked on each voyage are roughly the same, which cannot be validated. Two versions of the results are therefore presented, since either procedure has some drawbacks. Also note that the region/decade assignment model can be optimized further by taking advantage of our knowledge of the national flag of the voyage. The two methods, reported in the Appendix (technical or statistical?), show differences at the region level but are almost identical in their estimation of total embarkations.

## 4. New Method 2: Markov Chain Monte Carlo Analysis

Bayesian statistics in general.

Distributions used in the analysis: multinomial, Dirichlet, Poisson, binomial, and Gamma. Give graphic representations and other discussion for each distribution

The Markov Chain in general.

Monte Carlo estimates in general. [what is a case? Distinguish MCMC-SI and MCMC-MI] Compare with start of strategy of multiple methods similar but distinct.

*Imputing missing embarkation regions and arrival regions.*

The strategy of the analysis is to estimate the number of embarkations and the number of arrivals for each decade and each region, where the data are organized by voyage and where the length of the voyage has been previously estimated. That is, in the MCMC analysis, everything is known except the number of embarkations and arrivals for each decade and region these two variables are estimated in the analysis. In fact, there were missing data for region and voyage length as well as for embarkations and arrivals: these needed to be simulated in order to carry out the MCMC analysis for missing embarkations and arrival data; in contrast, there were no missing data for decade of departure from Africa.

Label data below and make them consistent for the two analyses.

*I. MCMC Imputation*

Our primary goal is to impute embarkation count by decade and embarkation region. Note that we have some missingness in almost every field, which includes embarkation count ($E$), arrival count ($A$), voyage length ($L$) and the ports of embarkation and arrival ($P_E, P_A$). Since imputed decade is used, there is no missingness in decade. All the models we introduce below are built for each decade separately.

*(1) Imputation for missing Embarkation Region and Arrival Region*

We first impute embarkation region and arrival region by MCMC method, particularly, the Gibbs-Sampler, assuming

$$\vec{\alpha} \sim Dirichlet(\vec{1}) \tag{21}$$

$$\vec{\beta} \sim Dirichlet(\vec{1}) \tag{22}$$

$$P_E \sim Multinomial(\vec{\beta}) \tag{23}$$

$$P_A \sim Multinomial(\vec{\alpha}) \tag{24}$$

This results in the conditional distribution:

$$\vec{\alpha}|P_A \sim Dir(\vec{1} + c_{P_A}) \tag{25}$$

$$\vec{\beta}|P_E \sim Dir(\vec{1} + c_{P_E}) \tag{26}$$

where $c_P$ is the count of voyage in region $P$. Given a value $\vec{\alpha}^{(t)}$ of *alpha* drawn at iteration $t$:
Imputation Step: Draw $P_{E,mis}^{(t+1)}$ with density $p(P_{E,mis}|P_{E,obs}, \vec{\alpha}^{(t)})$
Posterior Step: Draw $\vec{\alpha}^{(t)}$ with density $p(\vec{\alpha}^{(t)}|P_{E,obs}, P_{E,mis}^{(t+1)})$

Following a sufficient burn-in period, the iterative procedure can be shown eventually to yield a draw from the joint posterior distribution of PE,miss, a given PE, obs. Convergence diagnostics can also be conducted.

Missing values for regions. The first step of the analysis was imputing missing values for regions – embarkation regions and arrival regions – using the Gibbs Sampler. Assuming embarkation and arrival ports following multinomial distributions separately with parameters and , which have Dirichlet distributions, using Maximal likelihood estimates as the initial input parameters, we performed the Gibbs Sampler with 500 iterations and obtained our estimated value for and . And we imputed the missing values of embarkation regions and arrival regions based on multinomial distributions with parameter and respectively. Note that the estimated value is that of the final iteration, using rules for the Gibbs Sampler. Since there are no missing values with respect to decades, and among different decades, behaviors of embarkation and arrival can vary significantly, MCMC method is applied to each decade separately.

*Imputation for Missing Embarkartion and Arrival Counts, accounting for*

*Voyage Length.*

Once all the voyages have regions of embarkation and arrival, the main MCMC can take place (though it must also overcome any missing values of voyage length). The results of this work assign specific though randomly-generated regions to all of the missing embarkation regions and arrival regions. Calculations for assigning values to missing data are carried out separately for embarkation and arrival, but are carried out in the same program for convenience. How does this compare to multiple-method work?

Following a sufficient burn-in period, the iterative procedure can be shown eventually to yield a draw from the joint posterior distribution of $P_{E,obs}, \vec{\alpha}$ given $P_{E,obs}$. Convergence diagnostics can be then be conducted.

*(2) Imputation for missing Embarkation Count and Arrival Count*

The distributions that we assume for these variables are as follows:

$E|P_E \sim Possion(\lambda_{P_E})$

$L|P_E, P_A \sim Gamma(k, \delta_{R(P_E,P_A)})$

$A|E, L \sim Binomial(E, e^{-L\mu_{R(P_E,P_A)}})$

Exploratory analyses and graphics indicate that the assumptions of distributions are appropriate.

The fully conditional distributions are as follows:

$E|A, L, P_E, P_A, \mu, \lambda \sim A + Poisson(\lambda_{P_E}(1 - e^{-L\mu_{R(P_E,P_A)}}))$

$A|E, L, \mu, P_E, P_A, \sim Binomial(E, e^{-L\mu_{R(P_E,P_A)}})$

$f_L|A, E, P_A, P_E, \mu \propto L^{k-1}e^{-L(1/\delta_{R(P_E,P_A)}+\mu_{R(P_E,P_A)}A)}(1 - e^{-L\mu_{R(P_E,P_A)}})$

Parameters $\mu, \lambda, \delta$ are estimated from the voyages we have full records of. Imputations procedures proceed in much the same manner as imputation of region except that Multiple Imputation (MI) are implemented. We conduct posterior mean of the Monte Carlo samples imputation $m = 3$ times for each voyage, and then combine the results across the multiply imputed data.

Suppose that $\hat{Q}_j$ is an estimate obtained from data set $j(j = 1, ..., m)$ and $U_j$ is the standard error associated with $\hat{Q}_j$. The overall estimate is the

average of the individual estimates,

$$\bar{Q} = \frac{1}{m}\sum_{j=1}^{m}\hat{Q}_j \tag{27}$$

The between imputation variance is

$$B = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{Q}_j - \bar{Q})^2 \tag{28}$$

The total variance is

$$T = \frac{1}{m}\sum_{j=1}^{m}U_j + (1+\frac{1}{m})B \tag{29}$$

Missing data for voyage length must be estimated before embarkations and arrivals can be estimated. Metropolis-Hastings algorithm is used to draw $L_{mis}$ from the distribution where the normalizing constant is very difficult to compute. The term $(1-e^{-\mu_{R(P_A,P_E)L}})^{E-A}$ can be expanded into positive and negative terms and if we use the positive parts as the proposed distribution we obtain reasonable rejection thresholds. We are able to simulate from the distribution by noting that this proposal distribution may be thought of as a mixture. Namely,

$$f_{L|rest} \propto L^{k-1}e^{-L(1/\delta_R+\mu A)}(1-e^{-\mu L})^{E-A} \tag{30}$$

$$= L^{k-1}e^{-L(1/\delta_R+\mu A)}\Sigma_i^{[E-A]_0}(-1)^i e^{-i\mu L} \tag{31}$$

$$= L^{k-1}e^{-L(1/\delta_R+\mu A)}(f_0(L)-f_1(L)) \tag{32}$$

where $f_0(L) = \Sigma_{i,even}^{[E-A]_0}e^{-i\mu L}$ and $f_1(L) = \Sigma_{i,odd}^{[E-A]_0}e^{-i\mu L}$ and note that,

$$L^{k-1}e^{-L(1/\delta_R+\mu A)}f_0(L) \tag{33}$$

$$\propto \Sigma_{i,even}^{[E-A]_0}(1/\delta_R+\mu(A+i))^{-k}Gamma(L|k,1/\delta_R+\mu(A+i)) \tag{34}$$

This is clearly a mixture of Gamma distributions. The acceptance threshold is

$$\rho = min(1-\frac{f_1(V)}{f_0(V)})/(1-\frac{f_1(L_{t-1})}{f_0(L_{t-1})}),1 \tag{35}$$

where $V$ is a draw from the proposal distributions.

[Rewrite to fie] Secondly, voyage length (Gamma) is estimated with Gibbs Sampler and Metropolis-Hastings algorithm for unknown distributions. This requires making preliminary estimates of embarkations and arrivals in order to permit estimation of voyage length. The voyage length estimated in this process is then treated as known data in the next stage, the Monte Carlo estimation of embarkation and arrival totals. We performed imputation for missing values of embarkation numbers $E$, arrival numbers $A$ and voyage length $L$. With the assumption that embarkation number $E$ follows a Possion distribution with ${}_(P_E)$ related to the embarkation region, arrival number $A$ follows a binomial distribution with parameters associated to the number of embarkation population $E$, voyage length $L$, the port of embarkation $P_E$ and the port of arrival $P_A$, and voyage length $L$ follows a Gamma distribution with parameters related to embarkation region $P_E$ and arrival region $P_A$. Again we applied the Gibbs Sampler method to the data. Among those parameters, we paid special attention to the length of voyages. The reason is that compared with other unknown information, it is relatively complicated to derive the fully conditional distribution of voyage length. So we decided to use Metropolis-Hastings algorithm, which enables us to deal with unknown distributions. And we did 40 iterations for the voyage length within each single run of the Gibbs Sampler.

Standard errors are reported for embarkations and arrivals, but not for voyage length because its distribution is unknown and the algorithm for calculating variance and standard error is too complex. For each of 300 iterations in Gibbs Sampler, our procedure was to do 40 iterations for Voyage Length and take the final one as the voyage length.

*Estimation of embarkations and arrivals*
And overall, we performed 300 runs for Gibbs Sampler method, using the 50th to 99th, 150th to 199th and 250th to 299th to make imputation for the missing embarkation numbers E, arrival numbers A, voyage length L, simultaneously we computed standard errors for our estimates.

## 5. The Estimates: Comparison and Evaluation

Compare MCMC to Eltis 2013. That shows the difference  comment on thedifference. Also compare MCMC to Multiple Methods. Further, note Eltiss estimates of new voyages.

## 6. Conclusion

Other Africa and Other Destinations [review this with Bowen] Length of Voyage (compare early and late estimates; review the datacompare estimates to current data) More arrival regions  this can be done, though with less precision. But it wont be possible to use MCMC to estimate embarkation-arrival pairings.

## 7. Appendix

See appendix as noted on pg 7. The Appendix, an Excel sheet showing numbers of embarkations (and, in some cases, arrivals) by decade from as early as the 1650s to as late as the 1870s, presents results from the Multimethod analysis, the MCMC analysis, and two sets of estimates from the Slave Voyages dataset.

### Appendix 1. Embarkations

A.Comparison of MCMC estimates of embarkations with Slave Voyages estimates from 2015 and 2013. Length of Voyage figures do not seem dependable, though they are important in the overall estimation of embarkations.

B,Comparison of multi-method with MCMC, in numbers of captive migrants and in the standard error of the estimates. Discussion of variations in standard error.

C.Totals of each. Areas of relatively large difference. The question of Africa Other.

### Appendix 2. Arrivals

Summary of MCMC estimates for arrivals in Americas.