Multi-Source Data Fusion for the Estimation of Mean Shipping Densities

Yvan Gauthier
Maritime Forces Pacific Operational Research Team

Peter Minter
Co-op student, University of Victoria

DRDC CORA TM 2005–18
August 2005

Defence R&D Canada
Centre for Operational Research and Analysis

MARPAC OR Team

National Defence Défense nationale

Canada

Multi-Source Data Fusion for the Estimation of Mean Shipping Densities

Yvan Gauthier
Maritime Forces Pacific Operational Research Team

Peter Minter
Co-op student, University of Victoria

DRDC – Centre for Operational Research and Analysis

Technical Memorandum

DRDC CORA TM 2005–18

August 2005

Author

Yvan Gauthier

Recommended by

P. Sutherland
MARPAC DCOS OR

Approved for release by

J. Evans
DOR(MLA)

The information contained herein has been derived and determined through best practice and adherence to the highest levels of ethical, scientific and engineering investigative principles. The reported results, their interpretation, and any opinions expressed therein, remain those of the authors and do not represent, or otherwise reflect, any official opinion or position of DND or the Government of Canada.

Abstract

Maintaining continuous tracks of all vessels crossing a large oceanic area is a very difficult task to perform. For this reason, the true mean shipping densities over a given period of time can rarely be directly calculated; they need to be estimated from a limited number of observations made at discrete points in time. The completeness and correctness of these observations may vary from one data source to another. This paper proposes a Bayesian fusion method for producing shipping density maps of high resolution based upon vessel position reports obtained from various sources of maritime traffic data. An example involving data from three sources with different characteristics is given. A computer program implementing the fusion algorithm is also presented.

Résumé

Suivre à la trace tous les navires traversant une large aire océanique est une tâche très difficile. Pour cette raison, la densité exacte de navires durant une certaine période de temps peut rarement être directement calculée et ne peut qu'être estimée que sur la base d'un nombre limité d'observations faites à divers points dans le temps. Ces observations ne sont pas nécessairement toujours complètes ou correctes d'une source à une autre. Ce rapport propose une méthode de fusion Bayesienne pour produire des cartes à haute résolution représentant la densité de navires en utilisant les rapports obtenus de diverses sources de données reliées au trafic maritime. Un exemple impliquant trois sources de données avec des caractéristiques différentes est présenté. Un programme implémentant l'algorithme de fusion est aussi décrit.

This page intentionally left blank.

Executive summary

Shipping density maps describe the mean geographic distribution of vessels within a given region. Applications of these maps are numerous: they can be used to analyze and simulate maritime traffic, perform risk analyses associated with environmental and marine safety issues, evaluate the completeness of the maritime picture compiled by traffic management authorities, help in defining surveillance requirements, model sources of ambient noise, and so on.

For small regions, accurate measurements of shipping densities can often be achieved through a single sensor or source of traffic data. For large oceanic areas however, this task can be a lot more challenging because there is generally no convenient way to maintain continuous tracks of all vessels, and shipping densities need to be estimated from a certain number of observations made at discrete points in time. Various methods have been proposed in the past to perform this kind of estimate, but the accuracy and/or resolution of the resulting maps have been shown to be relatively limited.

This paper proposes a Bayesian fusion method for producing shipping density maps based upon vessel position reports obtained from various sources of maritime traffic data. In essence, the algorithm uses data coming from "sea-truthing" sensors such as maritime patrol aircraft to calibrate report density maps obtained from other sources, and then merges data from all sources into a single map representing the most-likely levels of shipping activity for the period of interest. The procedure is flexible enough so that position reports from almost any type of sensor or traffic data source can be utilized. Sources known to miss vessels, produce duplicate reports of a same vessel, or include false alarms can still be used to a certain extent. Through the Bayesian approach, initial estimates of densities can be added to the mix in order to take advantage of historical data and increase the fidelity of the fused map. Another benefit of this approach is that probable ranges for the shipping densities are straightforward to compute and offer clear measures of confidence in the estimated densities.

The utility of the fusion method is demonstrated in an example involving data from three sources with different characteristics. Fictitious density measurements off the West Coast are presented and then, using two different metrics, it is shown that the algorithm can derive density maps that are significantly closer to the true map than any of the individual maps used as input. A Java® tool implementing the fusion algorithm is described in Annex A.

The method described herein was developed for a study of the maritime traffic in the North-East Pacific conducted by Maritime Forces Pacific Operational Research Team in support of Operation SEA LION. This report is mainly intended for analysts who want to know the detail of the fusion method, or apply it to their own studies.

Yvan Gauthier, Peter Minter; 2005; Multi-Source Data Fusion for the Estimation of Mean Shipping Densities; DRDC CORA TM 2005–18; DRDC – Centre for Operational Research and Analysis.

Sommaire

Les cartes de densité de navires décrivent la distribution géographique moyenne des navires dans une région donnée. Les applications de ces cartes sont nombreuses : elles peuvent être utilisées pour analyser et simuler le trafic maritime, pour procéder à des analyses de risque associées à l'environnement et à la sécurité maritime, pour évaluer la couverture et les besoins d'un centre de surveillance, modéliser les sources acoustiques, etc.

Dans de petites régions bien circonscrites, des mesures précises de la densité de navires peuvent souvent être obtenues par l'intermédiaire d'une seule source de données. Toutefois, pour de larges aires océaniques, cette tâche s'avère beaucoup plus complexe car il n'y a généralement pas de moyen facile pour suivre à la trace tous les navires et la densité de navires doit être calculée sur la base d'un certain nombre d'observations. Diverses méthodes ont déjà été proposées pour faire ce genre d'estimation, mais la précision et/ou résolution des cartes qui en résultent sont relativement limitées.

Ce rapport propose une méthode de fusion Bayesienne pour produire des cartes à haute résolution représentant la densité de navires en utilisant les rapports obtenus de diverses sources de données reliées au trafic maritime. Essentiellement, l'algorithme utilise les données provenant de sources très fiables pour calibrer la cartes produites à partir d'autres sources, avant de fusionner toutes les mesures en une seule carte représentant les densités de navires les plus probables pour la période d'investigation. La procédure est assez flexible pour que des rapports d'a peu près n'importe quel types de senseur ou source de données soient utilisés. Les sources étant reconnues pour ne pas rapporter tous les navires, ou celles produisant de faux rapports peuvent être utilisées dans une certaine mesure. Grâce à l'approche Bayesienne, des estimations initiales de la densité de navires peuvent être fusionnées de manière à prendre avantage des données historiques et augmenter la fidélité des résultats. Un autre bénéfice de cette approche est que des intervalles de probabilités sont facilement calculables et offrent une mesure de confiance dans les densités estimées qui est claire à comprendre.

L'utilité d'une telle méthode est démontrée par un exemple impliquant trois sources de données avec des caractéristiques différentes. Des mesures de densités obtenues au large de la Côte Ouest sont d'abord simulées, puis par l'utilisation de deux métriques différentes, il est démontré que l'algorithme permet d'obtenir des mesures de densité significativement plus proches de la réalité que n'importe quelle mesure initialement disponible. Un outil en Java® implémentant l'algorithme de fusion est présenté en annexe.

La méthode présentée ici a été développée dans le cadre d'une étude du trafic maritime dans le Nord-Est du Pacifique, laquelle est conduite par l'équipe de recherche opérationnelle des Forces Maritime du Pacifique en support à l'opération SEA LION. Ce rapport est principalement rédigé à l'intention des analystes voulant en savoir plus sur la méthode employée, ou l'appliquer à leurs propres travaux de recherche.

Yvan Gauthier, Peter Minter; 2005; Multi-Source Data Fusion for the Estimation of Mean Shipping Densities; DRDC CORA TM 2005–18; RDDC – Centre d’analyse et de recherche opérationnelle.

Table of contents

Abstract i

Résumé i

Executive summary iii

Sommaire v

Table of contents vii

Tables ix

Figures ix

Acknowledgements x

1 Background 1

1.1 Challenges of shipping density measurements 1

1.2 Sources of shipping data 2

1.3 Aim 3

1.4 Outline 3

2 Deriving and fusing density maps 4

2.1 Report density maps from individual data sources 4

2.2 Bayesian fusion of report density maps 5

2.3 Estimating the parameters 6

2.3.1 Estimating α_i and β_i 6

2.3.2 Estimating σ_i 8

2.4 Confidence measures 9

2.5 Summarizing the algorithm 10

3 An example 11

3.1 Fictitious density measurements off the West Coast 11

3.2 Results of fusion 14

3.2.1 Picture quality metrics 14
3.2.2 Contribution of prior knowledge 15
3.2.3 Contribution of sea-truthing source 15
3.2.4 Contribution of additional samples for source calibration . . 16
4 Concluding remarks 17
4.1 Pros and cons 17
4.2 Future look 18
References 19
Annex 21
A SEASCAPE 21
A.1 Getting started 21
A.2 How to use SEASCAPE 21
A.2.1 The main menu 21
A.2.2 Create a setup file 21
A.2.3 Fusing Maps 23
A.2.4 Visualization 24
Appendix 1: SEASCAPE setup file format 28
Appendix 2: Grid file format 30
List of symbols/abbreviations/acronyms 31

Tables

1 Comparing density maps to the true mean shipping densities. 15

2 Benefit of data fusion as the number of observations by Source A increases. . 16

3 Benefit of using additional samples for calibrating sources 16

Figures

1 Using a local analysis window of $L \times L$ cells (left) for estimating α_i and β_i in the center cell (right). 8

2 "True" mean shipping densities used to produced fictitious reports. 12

3 Report density maps a) from Source A (MPA), b) from Source B (satellite), c) from Source C (self-reporting vessels). Figure in d) represents shipping densities measured a year earlier. 13

4 Fused density map representing the most likely mean shipping densities for the period of interest. 14

A.1 Setup file editor screen. 22

A.2 Choosing a setup file. 24

A.3 Choosing output file. 25

A.4 Map fusion in progress. 25

A.5 Visualization options. 26

A.6 Examples of visualization screens. 27

Acknowledgements

The authors would like to thank Mr. Ed Emond for his insightful comments and suggestions regarding Bayesian methods, as well as Mr. Sean Bourdon for his observations on the picture quality metrics used in this work. Many thanks also to Mr. Paul Sutherland for his reviews of draft versions of the report.

1 Background

1.1 Challenges of shipping density measurements

Shipping density maps describe the mean geographic distribution of vessels within a given region. These maps constitute essential references for analyzing a large range of economic, environmental, and security issues. In the recent past, such maps have been instrumental to many CORA studies of maritime surveillance systems [1, 2, 3, 4], which all required detailed descriptions of the seaborne activity along Canadian coasts in order to produce credible results.

For small regions, the measurement of shipping densities can often be achieved through a single source of data. For instance, in the immediate vicinity of a port where all vessel movements are monitored in near-real time (e.g., via a radar station, an automated identification system, or other means), the data gathered by traffic management authorities may be sufficient to produce accurate density maps. It is simply a matter of averaging the number of vessels in various locations, over a certain period of time.

However, producing shipping density maps for large oceanic areas is a lot more challenging. In Canada, the Maritime Forces have multiple responsibilities with regard to the defence of North America, the maintenance of the country's sovereignty and interests, as well as the provision of search and rescue services along Canada's coasts. The execution of these responsibilities requires a comprehensive situational awareness of the seaborne activity occurring in areas totalling more than 10 millions square kilometers over three oceans. Although various surveillance assets (e.g., ships, aircraft, ground-based radars, etc.) are available to the navy for monitoring vessel movements, measuring the average shipping activity everywhere inside such a large area remains difficult.

A method has been proposed by Dickinson and Boivin [5] for estimating traffic levels using vessel reports received from various sources at a maritime surveillance centre. The method involves the analysis of report continuity during three-day time periods (the approximative duration of transit for vessels in the area). It assumes that over such a period, the number of vessels leaving the surveillance area will roughly balance out the number entering the area. After counting which ships have been reported and which have not on day 1, day 2, and day 3, it is possible to approximate how many ships are not reported on a typical day, and by adding to those reported, produce a crude estimate of what the true average population of vessels is in the area. This is unfortunately not sufficient to determine precisely how many vessels are not tracked, nor where vessel detection fails. It also requires reports to include the identity of the vessels, a piece of information that is not always available.

Approximative density maps can also be produced by estimating route envelopes from vessels' departure and destination points, using the shipping statistics released by major ports. This approach is used by the U.S. Naval Oceanographic Office to generate worldwide maps of the expected shipping densities at different times of the year, and for various types of vessels [6]. It has the disadvantage of focusing exclusively on large

commercial vessels bound for a limited number of ports, and because it relies on predicted routes instead of actual routes, the precision of the derived maps is relatively poor. Moreover, the method cannot be applied to smaller vessels, such as fishing boats or pleasure craft, since their trajectories cannot be predicted on the sole basis of their origins and destinations.

One way to produce more meaningful representations of the shipping activity is to perform aerial surveys of the region of interest. Maritime patrol aircraft (MPAs), when using the right sensors and following appropriate flight patterns, can report all vessels sailing inside a given area with a high degree of confidence in the ship count and positions. Yao and Barnes [7] describe statistical techniques that can be applied to derive densities from multiple flights with even or uneven coverage of the area. The problem is that even long-endurance aircraft can only cover tiny portions of an ocean during a single patrol, and the number of flights required to obtain many statistical samples over large regions can easily become prohibitive, limiting the achievable resolution of the derived maps. For this reason, taking advantage of other sources of data is desirable to produce accurate, high-resolution shipping density maps.

1.2 Sources of shipping data

In the context of this work, sources of ship position reports are divided into two broad categories: “sea-truthing” sources and the others. A source is said to describe sea truth when it correctly reports[1] all vessels within a given area at a certain time, with no spurious reports or false alarms. As previously mentioned, MPAs have the ability to produce very accurate surveys and fall under this category, but the number of patrols realistically obtainable in a given period and their spatial extent are both limited. Similarly, seaborne platforms such as Navy or Coast Guard vessels can accurately monitor shipping activities, but because of their slow speeds and short horizons, they can only describe sea truth in their immediate neighbourhood.

On the other hand, wide-area sources of traffic data exist, but the completeness of the information provided is not guaranteed. For instance, most vessels are reporting their own positions to authorities on a regular basis, but they are not always required to do so, especially in international waters. Surveillance satellites represent another source of reports, and they have the ability to cover very broad regions several times per day, depending on their orbits, payloads and swath widths. The drawback of these systems is that they are far from the surface of the Earth, and depending on the detection mode, they can miss vessels or produce false alarms biasing the number of counted ships. Ground-based sensors installed along the coast, such as radar stations, are also abundant sources of traffic data. The coverage offered by these sensors is limited to areas adjacent to the land. The completeness of the data obtained is also uneven, being typically better at short range than at long range.

[1] For simple density measurements, the reports only need to include the positions of the vessels. For other types of studies, sea truth could involve the correct reporting of vessels’ speeds, courses, identities, or other parameters.

1.3 Aim

It is obvious that each type of shipping data source has its own advantages and limitations. The aim of this paper is to present a Bayesian algorithm that can be used to merge data coming from various sources (sea-truthing and other), and produce density maps of higher resolution and higher precision than what individual sources can produce. This kind of technique is increasingly employed in fields where images from multiple sensors need to be combined, such as medical diagnosis [8, 9], navigation guidance [10], and image reconstruction [11], just to cite a few. The complexity of the fusion model can vary quite a lot, and depends on the sensors' characteristics as well as the type of phenomenon analyzed.

The method described herein was developed for a study of the maritime traffic in the North-East Pacific conducted by Maritime Forces Pacific Operational Research Team (MARPAC ORT) in support of Operation SEA LION [12]. This report is mainly intended for analysts who want to know the detail of the fusion method, or apply it to their own studies. Results of the MARPAC ORT traffic analysis will be published in a subsequent report.

1.4 Outline

This paper is divided into two parts. First, the mathematics of the method and an algorithm implementing the fusion technique are presented. In the second part, simulated measures of the North-East Pacific vessel activity are used to demonstrate the potential of the method. The example uses data from fictitious surveillance activities to keep the paper at an unclassified level. Two metrics are employed to evaluate the quality of the fused density map by comparing it to the modelled sea truth. Strengths and weaknesses of the method are then discussed in the conclusion. A user's guide to a Java® application implementing the fusion algorithm is included in Annex A.

2 Deriving and fusing density maps

2.1 Report density maps from individual data sources

Consider an oceanic area of interest divided into small square cells (e.g., $1^\circ \times 1^\circ$). The number of ships $s(x, y, t)$ sailing inside a cell centered at longitude x and latitude y is a stochastic function of time t. The goal here is to determine what the average value of $s(x, y, t)$ is over a pre-determined period of time T, say one month. This average is called the *mean shipping density* $\rho(x, y)$ and is mathematically defined as:

$$\rho(x, y) \equiv \frac{1}{T} \int_{t_0}^{t_0+T} s(x, y, t)dt \tag{1}$$

In practice, an accurate and continuous monitoring of $s(x, y, t)$ inside all cells is generally not possible over long periods, and ρ must be estimated on the basis of a certain number of observations made at discrete points in time $\{t_j, j = 1, \ldots, N\}$ during the period of interest $[t_0, t_0 + T]$. Moreover, the number of vessels reported by the ith data source, $n_i(x, y, t_j)$, may be different from the actual number of ships $s(x, y, t_j)$ depending on the quality of the data provided by the source. Assuming $N_i(x, y)$ counts of the number of vessels are performed for a given cell, the *report density map* derived from the ith data source can be expressed as:

$$d_i(x, y) = \sum_{j=1}^{N_i} n_i(x, y, t_j)/N_i(x, y) \tag{2}$$

It is assumed that the time interval between each observation is long enough for the population of ships inside the cell to change, so that all samples obtained from a given source can be considered independent of each other.

The fusion model proposed in section 2.2 is based on the assumption that to the first degree of approximation, the relationship between $d_i(x, y)$ and $\rho(x, y)$ can be expressed as:

$$d_i(x, y) = \alpha_i(x, y)\rho(x, y) + \beta_i(x, y) + \epsilon_i(x, y) \qquad (\alpha_i > 0, \beta_i \geq 0) \tag{3}$$

The parameter α_i represents a scale factor describing by how much a certain data source is underestimating or overestimating the real shipping activity at a given location. A source can underestimate the number of vessels for many reasons: inefficient sensors, unfavourable weather conditions, reporting failures, etc. On the other hand, shipping density can be also be overestimated, for example in the case where multiple vessels are reported by a given source, but are actually representing a single ship. Another possibility is that a source will produce false alarms and report non-existent vessels. This phenomenon is represented by the positive bias β_i. The parameter ϵ_i is an error term representing random variations in the quality of the data provided by each data source. This term is assumed to be normally distributed with zero mean and variance $\sigma^2_{\epsilon_i}$.

Because each report density map is the result of an averaging process, the central limit theorem[2] shows that the probability distribution function $\mathcal{P}$ for $d_i(x, y)$ tends to a normal distribution as the number of observations increases. Dropping the reference to spatial location, this function is written as:

$$\mathcal{P}(d_i|\rho) = \frac{1}{\sigma_{\epsilon_i}\sqrt{2\pi}} e^{-\frac{(d_i-\alpha_i\rho-\beta_i)^2}{2\sigma_{\epsilon_i}^2}} = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{(\rho_i-\rho)^2}{2\sigma_i^2}} \tag{4}$$

where $\rho_i \equiv \frac{d_i-\beta_i}{\alpha_i}$ is an estimate of ρ with variance σ_i^2. The derivation of α_i, β_i, and σ_i will be discussed in section 2.3.

2.2 Bayesian fusion of report density maps

Here is essentially the fusion task that needs to be performed: given a set $\mathbf{d}$ containing M report density maps $\{d_i, i = 1, \ldots, M\}$ produced from M different sources of data, we want to derive a single shipping density map $\hat{\rho}$ approaching as much as possible the true map ρ.

There are many ways to combine the density maps and to infer the true level of shipping activity. A Bayesian formulation is particularly convenient because it allows the incorporation of *a priori* knowledge of this activity. For instance, it is known that ships are not uniformly distributed over the ocean surface. High vessels densities are expected to be found in well-defined areas, such as fishing zones or shipping lanes, whereas low densities are expected in "exclusion zones" set up by governmental authorities. Taking advantage of such information is useful, especially when the number of available data sources is limited.

Suppose $\mathcal{P}(\rho)$ represents the prior knowledge of what shipping densities are before data is actually received and analyzed. The *a priori* probability distribution function for ρ can also be expressed as a normal distribution:

$$\mathcal{P}(\rho) = \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{(\rho-\rho_0)^2}{2\sigma_0^2}} \tag{5}$$

For instance, ρ_0 could represent the shipping densities measured in a previous study, or the densities predicted from port shipping statistics. The parameter σ_0 determines the error associated with this initial information.

One method to estimate true densities is to maximize the probability of a certain shipping activity level given the set of report density maps $\mathbf{d}$ on hand. Using Bayes' rule, the *a posteriori* probability distribution function after deriving density maps from various sources is

$$\mathcal{P}(\rho|\mathbf{d}) = \frac{\mathcal{P}(\rho)\mathcal{P}(\mathbf{d}|\rho)}{\mathcal{P}(\mathbf{d})} = \frac{\mathcal{P}(\rho)\prod_i \mathcal{P}(d_i|\rho)}{\prod_i \mathcal{P}(d_i)} \tag{6}$$

[2]The central limit theorem states, in essence, that the distribution of an average tends to be normal, even when the distribution from which the average is computed is decidedly non-normal.

where the denominator is a normalizing factor. The equation is only valid when density maps are produced from independent sources.

Given equations 4 and 5, the log likelihood of Equation 6 can be written as:

$$\begin{aligned}\mathcal{L}(\rho) &= \log \mathcal{P}(\rho|\mathbf{d}) \\ &\propto \sum_i \left(\frac{d_i - \alpha_i\rho - \beta_i}{\sigma_{\epsilon_i}}\right)^2 + \left(\frac{\rho - \rho_0}{\sigma_0}\right)^2 \\ &\propto \sum_i \left(\frac{\rho - \rho_i}{\sigma_i}\right)^2 + \left(\frac{\rho - \rho_0}{\sigma_0}\right)^2 \end{aligned} \tag{7}$$

By differentiating $\mathcal{L}$ with respect to ρ and equating to zero, the density map $\hat{\rho}$ being the most likely to represent the true level of shipping activity is obtained:

$$\boxed{\hat{\rho} = \frac{\sum_i \frac{\rho_i}{\sigma_i^2} + \frac{\rho_0}{\sigma_0^2}}{\sum_i \frac{1}{\sigma_i^2} + \frac{1}{\sigma_0^2}}} \tag{8}$$

This estimator of ρ is a shrinkage estimator also known as Stein's estimator [13, 14]. It is intuitively appealing because it takes the form of a weighted mean where individual density estimates with low variances carry more weight than those with high variances.

2.3 Estimating the parameters

In order to use Equation 8, the unknown parameters α_i, β_i and σ_i first need to be estimated. The following paragraphs explain how this can be achieved.

2.3.1 Estimating α_i and β_i

As discussed before, the parameters α_i and β_i define the relationship that exists between the actual shipping densities and the reported shipping densities for the ith source . Depending on the data available, various approaches can be taken for estimating these two parameters. Two cases are covered here.

Case 1: Simultaneous ship counts available. Let $\{t_k, k = 1, \ldots, K\}$ be a random sample of K times at which the ship count n_i from a possibly inaccurate source and the ship count s from a sea-truthing source (e.g., MPA) are both available. Such simultaneous vessel counts are often performed during the technical evaluation phase of a maritime sensor. Assuming a linear model similar to the one of Equation 3,

$$n_i(t) = \alpha_i s(t) + \beta_i + \epsilon_i \qquad (\alpha_i > 0, \beta_i \geq 0) \tag{9}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma_{\epsilon_i}^2)$, we can estimate the values of α_i and β_i by performing a linear regression on the values of $n_i(t_k)$ and $s(t_k)$. Note that the sample of times at which ship counts are obtained does not have to be restricted to the interval $[t_0, t_0 + T]$. A wider time

frame might increase the value of K, and thus lead to more precise estimates of the parameters. However, if samples are obtained from a time frame different than $[t_0, t_0 + T]$, then there must be evidence that the estimated values of α_i and β_i are still representative of the source's characteristics for the period of interest.

The most common method for performing linear fits is to minimize the sum of squared residuals:

$$\sigma^2_{\epsilon_i} = \frac{1}{K} \sum_{k=1}^{K} \left[\alpha_i s(t_k) + \beta_i - n(t_k)\right]^2 . \tag{10}$$

Numerous algorithms exist to do such a linear fit numerically. The one used in the example of section 3 and implemented in the application of Annex A is the Nelder and Mead simplex regression method. The method is not the fastest, but it is robust and can be employed to perform non-linear regressions if more complicated models happen to be used. The fit parameters can also be constrained between pre-defined limits if necessary. The reader can refer to [15, 16] for additional details.

There may be instances where a source is known not to produce false reports, which simplifies the fitting procedure since β_i can be set to zero. In any case, it must be ensured that the results of the regression ($\alpha_i \pm \sigma_{\alpha_i}$ and $\beta_i \pm \sigma_{\beta_i}$) are credible. If $\alpha_i \leq 0$ or $\beta_i < 0$, then the data coming from the ith source should be excluded from the fusion model previously described.

Case 2: Simultaneous ship counts unavailable. In practice, the kind of calibration data described above may not be available. Simultaneous ship counts by a data source and a sea-truthing reference are often difficult to obtain, especially when analysts have no control on the origin of the reports (e.g., voluntary reports from ships, data obtained from an external agency). Moreover, as the chosen cell size gets smaller, the time for the vessel population to change within a given cell also diminishes, forcing ship counts to be well synchronized. For this reason, we must find an alternate, more convenient way to calibrate the sources by using only the data on hand.

Let's go back to the set of report density maps $\mathbf{d}$ produced by averaging ship counts obtained during the period of interest $[t_0, t_0 + T]$ (see Equation 2). If one of these maps d_i is the product of a sea-truthing data source, then we can use it as a reference map $\widetilde{\rho}$ against which other report density maps will be calibrated. If two or more of the available sources are assumed to report sea truth, then a linear combination of these report density maps can be used to define $\widetilde{\rho}$.

Now consider a local analysis window composed of $L \times L$ cells centered on the cell for which α_i and β_i need to be estimated, as shown in Figure 1. We can again estimate these two parameters by performing a linear fit, this time between the $L \times L$ values of d_i and $\widetilde{\rho}$. The larger the local analysis window, the larger the number of points for the regression will be, potentially increasing the precision of the parameters. However, the size of the window must be kept reasonable, depending on how the quality of data provided by the ith source varies over space. If the source demonstrates the same level of data completeness

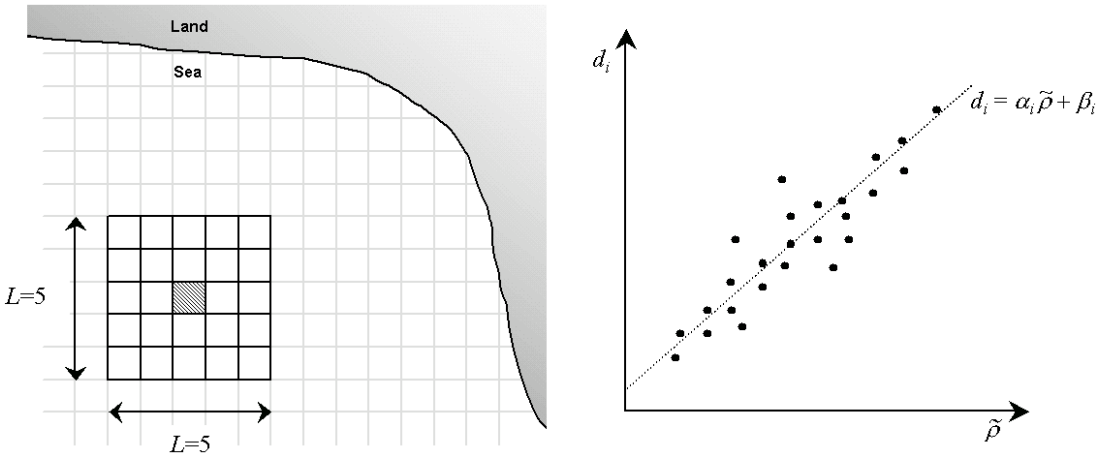

Figure 1: *Using a local analysis window of $L \times L$ cells (left) for estimating α_i and β_i in the center cell (right).*

and correctness everywhere, then a large value can be used for L. On the other hand, if the quality of the data markedly changes from one location to another (e.g., the data from a ground-based radar), then the value of L must be carefully chosen to ensure that α_i and β_i remain roughly the same all over the local analysis window. Note that when cells of the window completely overlap with a land area, their values must be excluded from the regression.

Note also that as for Case 1, the report density maps used in the estimation of α_i and β_i dot not necessarily need to be produced from observations made during the the period of interest only. Better estimates can generally be obtained if a larger time frame involving more observations is used, but it must be ensured that the values obtained are still representative of the source's characteristics for the interval $[t_0, t_0 + T]$.

2.3.2 Estimating σ_i

The parameter σ_i is essentially the standard error associated with the estimate of shipping density, ρ_i, produced from the ith data source. Alternatively, one can interpret $1/\sigma_i^2$ as the level of confidence, or weight, attached to each value of ρ_i in Equation 8.

Remember that $\rho_i \equiv \frac{d_i - \beta_i}{\alpha_i}$. By applying basic rules of error propagation [17], σ_i can be approximated as follows:

$$\begin{aligned} \left(\frac{\sigma_i}{\rho_i}\right)^2 &\approx \left(\frac{\sigma_{d_i-\beta_i}}{d_i - \beta_i}\right)^2 + \left(\frac{\sigma_{\alpha_i}}{\alpha_i}\right)^2 \\ &\approx \frac{\sigma_{d_i}^2 + \sigma_{\beta_i}^2}{(d_i - \beta_i)^2} + \frac{\sigma_{\alpha_i}^2}{\alpha_i^2} \end{aligned} \tag{11}$$

so that

$$\sigma_i^2 \approx \frac{\sigma_{d_i}^2 + \sigma_{\beta_i}^2}{\alpha_i^2} + \frac{\sigma_{\alpha_i}^2 (d_i - \beta_i)^2}{\alpha_i^4} \tag{12}$$

This equation neglects any correlation between the parameters. This is of course appropriate when the values of d_i, α_i, and β_i are produced independently of each other, from different data sets. Even when the parameters are obtained by means of regression through a local analysis window (Case 2), this remains a reasonable first-order approximation because the local value of d_i for the central cell has little influence on the values of α_i and β_i derived from all window cells.

All parameters of Equation 12 have already been determined so far, except $\sigma_{d_i}^2$. Going back to the initial averaging of ship counts (see Equation 2), the central limit theorem allows us to express $\sigma_{d_i}^2$ as

$$\sigma_{d_i}^2 \approx \frac{\text{Var}(n_i)}{N_i} = \frac{1}{N_i(N_i - 1)} \sum_{j=1}^{N_i} (n_i - d_i)^2 \tag{13}$$

when the value of N_i is large. If this is not the case, then the equation can still be used by considering a time frame wider than $[t_0, t_0 + T]$ in the evaluation of Var(n_i).

2.4 Confidence measures

The Bayesian framework of the present estimation method offers, through the posterior distribution of Equation 6, a direct access to the state of knowledge about ρ. As discussed in [14], this is very convenient because we can directly compute the probability for ρ to lie within a specific range $[\rho_{\text{low}}, \rho_{\text{high}}]$, without relying on the classical concepts of confidence intervals. This probability can be expressed as:

$$\mathcal{P}(\rho_{\text{low}} \leq \rho \leq \rho_{\text{high}} | \mathbf{d}) = \int_{\rho_{\text{low}}}^{\rho_{\text{high}}} \mathcal{P}(\rho | \mathbf{d}) d\rho, \tag{14}$$

or more explicitly as:

$$\mathcal{P}(\rho_{\text{low}} \leq \rho \leq \rho_{\text{high}} | \mathbf{d}) = \frac{\int_{\rho_{\text{low}}}^{\rho_{\text{high}}} \exp\left[-\frac{(\rho - \rho_0)^2}{2\sigma_0^2} - \sum_i \frac{(\rho - \rho_i)^2}{2\sigma_i^2} \right] d\rho}{\int_0^{\infty} \exp\left[-\frac{(\rho - \rho_0)^2}{2\sigma_0^2} - \sum_i \frac{(\rho - \rho_i)^2}{2\sigma_i^2} \right] d\rho}. \tag{15}$$

This equation is very easy to compute numerically and the Bayesian confidence intervals derived from it represent valuable information that can be presented alongside a shipping density map to illustrate the level of uncertainty associated with the estimates.

2.5 Summarizing the algorithm

In a nutshell, the following steps need to be followed in order to apply the fusion method previously described:

1. Select two or more independent sources of vessel position reports;
2. Divide the area of interest into cells of a certain size (map resolution);
3. For each data source i:
   (a) Produce a report density map $d_i(x, y)$ (Eqn. 2) and compute the standard deviations, $\sigma_{d_i}(x, y)$ (Eqn. 13);
   (b) If the values of $\alpha_i(x, y)$ and $\beta_i(x, y)$ cannot be obtained from a proper technical evaluation of each source (as in Case 1 of section 2.3.1), then:
      i. Select one of the report density maps produced from a sea-truthing source as the reference map, $\widetilde{\rho}(x, y)$, or combine maps from sea-truthing sources into a single reference map.
      ii. Define the size of the local analysis window (Fig. 1);
      iii. Estimate $\alpha_i(x, y) \pm \sigma_{\alpha_i}(x, y)$ and $\beta_i(x, y) \pm \sigma_{\beta_i}(x, y)$ by performing a linear regression of $d_i(x, y)$ values over $\widetilde{\rho}(x, y)$ for all cells within the window that are not completely overlapping with land areas. For sea-truthing sources, assume $\alpha_i = 1$ and $\beta_i = 0$.
   (c) Compute the estimated shipping densities, $\rho_i(x, y) = [d_i(x, y) - \beta_i(x, y)]/\alpha_i(x, y)$;
   (d) Compute the error on estimated shipping densities, $\sigma_i(x, y)$ (Eqn. 12);
4. If an *a priori* estimate of the shipping density is available, it can be incorporated during the fusion process:
   (a) Produce a shipping density map, $\rho_0(x, y)$, reflecting prior information.
   (b) Determine $\sigma_0(x, y)$. This parameter can either be defined as the estimated error on $\rho_0(x, y)$, or it can be arbitrarily set by the analyst to control the amount of prior information entering the fusion process, in light of the $\sigma_i(x, y)$ values for the other data sources.
5. Compute the density map that is the most likely to represent the true shipping activity, $\hat{\rho}(x, y)$ (Eqn. 8);
6. If desired, compute a Bayesian confidence interval for $\rho(x, y)$ using Eqn. 15.

This algorithm was implemented in Java® by the authors. A description of the *Statistical Evaluator of Average Shipping and Coastal Activity Picture* (SEASCAPE) is included in Annex A.

3 An example

3.1 Fictitious density measurements off the West Coast

The following paragraphs demonstrate how the algorithm summarized in section 2.5 can be applied to measure the average marine activity levels within a very large region. In order to keep this example at the unclassified level, **no real surveillance data is used**. Data sets are derived from simple probabilistic models of fictitious sources. The performance levels and concepts of operations of the sensors described herein are different from those of real surveillance assets. They have been deliberately defined to exaggerate the strengths and weaknesses of each type of data source. Nevertheless, they suffice to illustrate the validity and utility of the fusion technique.

Consider the North-East Pacific region shown in Figure 2, which encompasses the area of interest to Commander MARPAC. Suppose one wants to determine the mean number of ships everywhere in this region at a certain time of the year, say the month of June. We will assume here that the shipping densities of Figure 2 represent the true shipping densities in the region during this period (although the map is actually just an approximation obtained from [6]). These shipping densities represent what one would obtain if it were possible to track all vessels in a continuous fashion for the full month, and compute the average number of vessels in each cell over that period.

Now consider three hypothetical and independent sources of data available to achieve this task:

- Source A consists of the data obtained from an MPA squadron. We will suppose that a series of patrols were performed such that the entire region was surveyed $N_A = 3$ times during the period of interest, each time providing an exact count and correct positions of vessels, and that Figure 3a is the resulting density map from these patrols.
- Source B is a surveillance satellite capable of covering the entire area once a day, such that $N_B = 30$ samples of the vessel population are produced during the month. However, we will suppose that each satellite pass only detects a certain fraction of the vessels, and is prone to reporting false contacts. The report density map obtained from Source B is shown in Figure 3b.
- Finally, Source C represents vessels that are periodically reporting their own positions to authorities (for legal reasons, weather-reporting purposes, or other). For the sake of the example, we will assume that the fraction of vessels that are self-reporting monotonically decreases with the distance from the coast, but that all vessels are giving their exact location (no false alarms). In a situation where the most current positions are recorded every eight hours, a total of $N_C = 90$ "snapshots" of the maritime picture can be made during the month. The report density map obtained from this data is shown in Figure 3c.

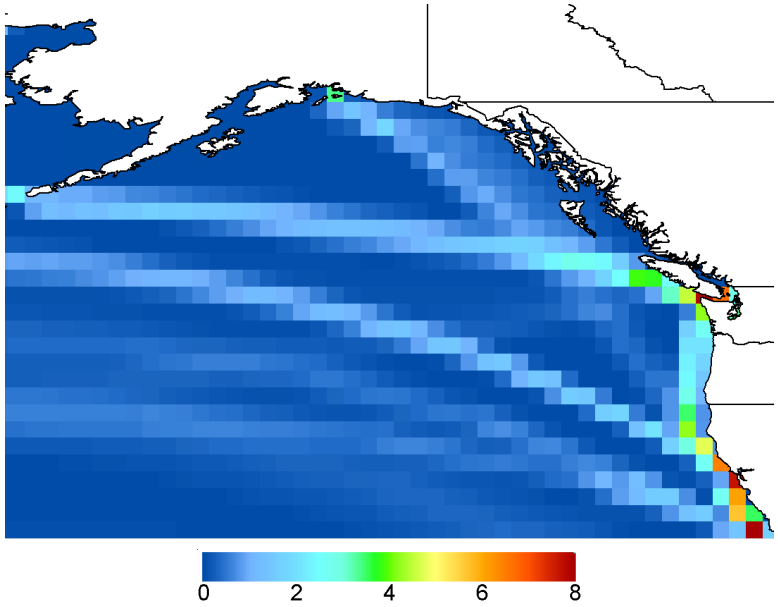

Figure 2: *"True" mean shipping densities used to produced fictitious reports.*

For simplicity, each data source offers here the same number of traffic observations for all cells. In real life, the value of N would likely vary over space, and the fusion method can accommodate situations where $N = N(x, y)$.

A quick look at Figure 3 shows that the report density map produced from Source A is the noisiest one. This was expected, given the low number of samples obtained by MPAs. One way to reduce the noise in this map would be to increase the surface of grid cells, and thus lose on resolution. In the case of Source B, the definition of shipping lanes is much clearer, but after comparing to the first map, it is evident that densities are underestimated by a large margin, whereas the impact of false alarms is difficult to determine. For the report density map produced from source C, the underestimation of densities is obvious, especially in deep ocean.

In addition to the maps produced from sources A, B, and C, suppose that shipping densities in this area have been somehow measured the year before, at the same period, with the result shown Figure 3d. This prior knowledge (ρ_0) of the seaborne activity is usable data. In the present case, we will assume that similar densities are expected for the current year, within a margin of error of $\pm 50\%$ (such that $\sigma_0 = 0.5\rho_0$).

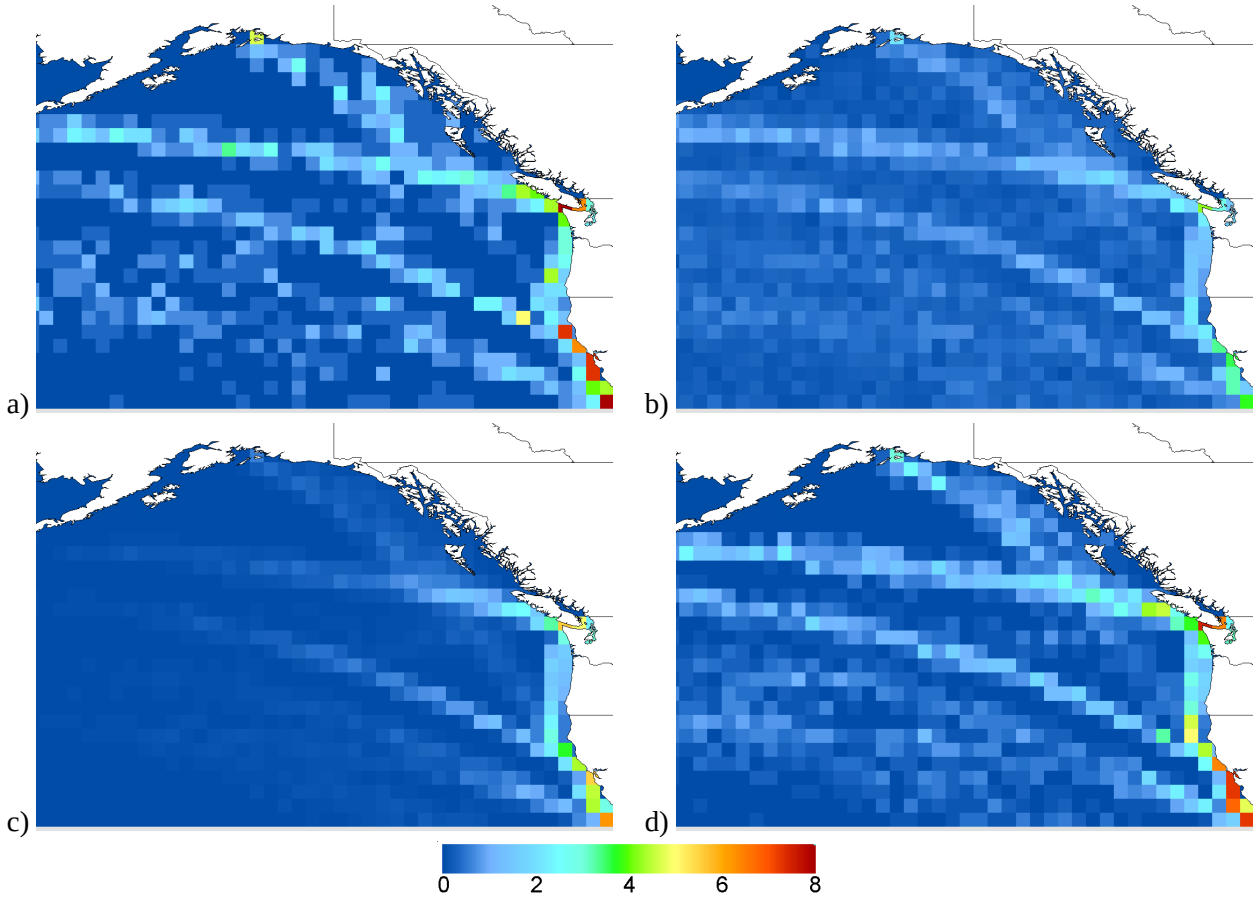

Figure 3: *Report density maps a) from Source A (MPA), b) from Source B (satellite), c) from Source C (self-reporting vessels). Figure in d) represents shipping densities measured a year earlier.*

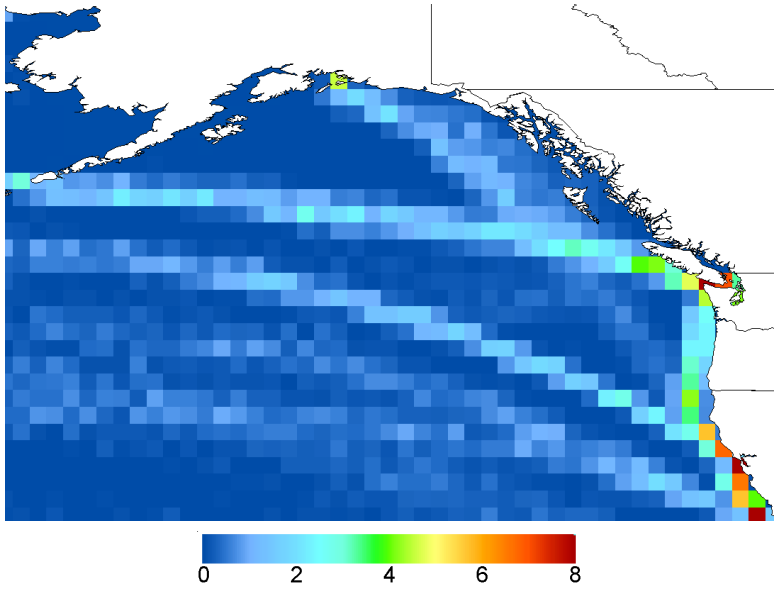

Figure 4: *Fused density map representing the most likely mean shipping densities for the period of interest.*

SEASCAPE was utilized to fuse the density maps. The report density map produced from Source A was used as the reference $\widetilde{\rho}$ to determine α and β values of sources B and C. A local analysis windows with $L = 25$ was used in performing the regression for Source B. Since the two parameters are not expected to vary much over space, other values could actually have been chosen for L, but a value of 25 was arbitrarily selected. In the case of Source C however, a smaller local window with $L = 5$ had to be utilized because Figure 3c demonstrates that for larger values of L, the quality of the data provided by the source would significantly change within the window.

3.2 Results of fusion

3.2.1 Picture quality metrics

Figure 4 shows the shipping densities obtained from the fusion of the maps of Figure 3. In order to quantify the improvement offered by the fusion technique over individual report density maps, the fused map was compared to the true mean shipping densities (Figure 2) using two metrics. First, the mean absolute difference per cell,

$$\Delta = \frac{\sum_{x,y} c(x,y)|d(x,y) - \rho(x,y)|}{\sum_{x,y} c(x,y)} \tag{16}$$

where $c(x,y)$ equals zero if a cell centered at (x,y) completely overlaps with land areas, and equals one otherwise. This metric is essentially the absolute difference between the true density and the calculated most-likely density, averaged over all relevant cells.

The second metric is the linear correlation coefficient between the two maps, also called Pearson's r, which can here be expressed as:

$$r = \frac{\sum_{x,y} c(x,y)[d(x,y) - \overline{d}][\rho(x,y) - \overline{\rho}]}{\sqrt{\sum_{x,y} c(x,y)[d(x,y) - \overline{d}]^2} \sqrt{\sum_{x,y} c(x,y)[\rho(x,y) - \overline{\rho}]^2}} \tag{17}$$

This coefficient provides an indication of how similar the shipping patterns of both maps are. It focuses on the relative distribution of vessels over space instead of the absolute number of ships.

Table 1 presents the values of the metrics for all input maps, as well as for the fused map. The fused map is the closest from the true densities for both metrics. The fusion process reduces the value of Δ by more than a half when the resulting map is compared to the one produced from the sea-truthing sensor (Source A) only. The correlation coefficient also increases significantly, from 89.7% to 98.2%. It is thus beneficial to supplement the data from Source A with data from other sources, even if these sources do not provide perfectly complete or accurate information when taken individually .

Table 1: *Comparing density maps to the true mean shipping densities.*

	Δ	r
Source A map (d_A)	0.238	89.7%
Source B map (d_B)	0.226	95.2%
Source C map (d_C)	0.281	91.7%
Prior map (ρ_0)	0.148	95.6%
Fused map ($\hat{\rho}$)	**0.106**	**98.2%**

3.2.2 Contribution of prior knowledge

How accurate would be the fused map in the absence of a prior map? The algorithm was re-run to answer this question. The value of Δ increased to 0.127 and the value of r decreased to 97.4%. Despite these changes, the level of marine activity remains more accurately described by the fused map than any of the input maps. This shows that the incorporation of historical data or initial density estimates can significantly improve the resulting map, even when the margin of error on these estimates ($\pm 50\%$ in the example) is relatively large.

3.2.3 Contribution of sea-truthing source

In the present example, each cell of the grid was surveyed only 3 times by MPAs, the only sea-truthing source available. The following question naturally follows: if more observations could have been done, would it still be advantageous to fuse the report density map from Source A with those produced from non-truthing sources?

Computations were repeated for two other scenarios where the number of MPA observations is increased to $N_A^{\dagger} = 10$ or $N_A^{\ddagger} = 50$. No changes were made to the data sets from other sources or to the fusion algorithm. Table 2 shows how an improved sea-truthing influences the utility of the fusion method.

Table 2: *Benefit of data fusion as the number of observations by Source A increases.*

	Δ	r
Source A map ($N_A = 3$)	0.238	89.7%
Fused map	0.106	98.2%
Source A map ($N_A^{\dagger} = 10$)	0.125	97.1%
Fused map†	0.078	98.9%
Source A map ($N_A^{\ddagger} = 50$)	0.057	99.3%
Fused map‡	0.046	99.6%

When the number of observations is increased from 3 to 10, the fusion algorithm still reduces Δ by almost 40%, and its impact on the correlation coefficient is also positive. If the number of MPA observations is 50, then trying to incorporate data from non-truthing sources does not improve the picture as much, since the density map produced from MPA reports is already very accurate. However, one can question the feasibility of producing so many samples from MPA patrols in a single month. Moreover, the *raison d'être* of the fusion method is to compensate for the scarcity of sea truth samples, so it is acceptable for the method not to be as beneficial in this kind of situation.

3.2.4 Contribution of additional samples for source calibration

For all of the previous results, the values of α and β for sources B and C were derived as explained in Case 2 of section 2.3.1, using only the data obtained during the month of interest. The fusion method was re-run to verify what happens if the calibration of sources is done by using data covering a wider period of a few months instead. Assuming that the performance characteristics of the sources do not change during these months and that a total of $N_A^* = 15$, $N_B^* = 150$, and $N_C^* = 500$ observations are available for calibration purposes, then the quality of the fused map previously obtained can be improved, as shown in Table 3.

Table 3: *Benefit of using additional samples for calibrating sources*

	Δ	r
Fused map (no additional samples)	0.106	98.2%
Fused map (additional samples)	**0.082**	**98.7%**

4 Concluding remarks

4.1 Pros and cons

Since the continuous tracking of all vessels within large regions is generally difficult to perform over long periods of time, true mean shipping densities can rarely be directly derived. They most often need to be estimated from a limited number of observations made at discrete points in time. Various statistical techniques can be used for this estimation process, but the one proposed in this work has many advantages:

- **Multi-source capability**. The algorithm is flexible enough so that position reports from virtually any type of sensor or shipping data source can be utilized. Sources known to miss vessels, produce duplicate reports of a same vessel, or include false alarms can still be used to a certain extent. Moreover, prior estimates of densities can be added to the mix in order to take advantage of historical data and increase the fidelity of the fused map.

- **Reduced requirement for sea truth samples**. Despite the fact that reports coming from sea-truthing sensor(s) are required by the algorithm (when α and β values of sources are *a priori* unknown), its ability to tap into a number of sources increases the achievable precision and resolution of the resulting density maps, while reducing the amount of sea truth data normally needed by conventional sampling techniques to achieve similar results.

- **No need for track reconstruction**. The presented method requires only position reports as an input. Ship trajectories do not need to be rebuilt from the sensor data. This avoids any incorrect extrapolations of vessel movements. It is particularly beneficial when investigating small vessel activity (e.g., fishing vessels, pleasure craft), since the paths followed by these ships are in general very difficult to predict, especially along coastlines.

- **Flexible map size and resolution**. The example presented in the previous section describes the computation of a shipping density map covering a million square nautical miles with a resolution of $1^\circ \times 1^\circ$, for a specific one-month period. The same method could be applied to produce maps of smaller regions, with a much finer cell size (e.g., $1' \times 1'$). It could also be used to precisely describe the level of seaborne activity during shorter periods matching the time frame of scientific experiments or military exercises.

- **Straightforward computation of confidence measures**. The Bayesian approach allows analysts to determine very easily the probability that true shipping densities fall within pre-defined intervals. In the context of maritime surveillance, the limits of such intervals could be used as boundaries outside which observed shipping densities would be considered unusual or suspicious.

In the example of section 3, it was demonstrated by two different metrics that the algorithm can produce density maps that are significantly closer to sea truth than any of

the individual maps used as input. There are thus benefits to fuse data from sea-thruting sources with data obtained from other, potentially inaccurate sources of information.

However, care must be taken in applying the fusion method. If the ship counts are somehow biased towards a certain portion of the period of interest (e.g., if they are all made during the first few days of a one-month time frame), or if they tend to favour certain categories of ships over others, then the fused map will likely exhibit a similar bias. A minimal understanding of the data sources is also needed to determine the size of the local analysis windows, when required.

4.2 Future look

The performance of the fusion algorithm could potentially be improved by making it better adapted to specific sensors and data sources. For instance, the linear assumptions of the fusion model could be revised to take into account particular characteristics of sensors. Another option could be to filter the density maps using the continuity equation. Vessel speeds and courses would then need to be added to the positional information when producing the fused maps.

In the near-term, the present version of the algorithm will be used to analyze shipping flows, densities, and patterns along the West Coast, using real surveillance data collected from various assets. Possible applications of the resulting maps within the maritime forces and DRDC are numerous. High-fidelity maps could be used to evaluate the completeness of the recognized maritime picture compiled by surveillance operation centres. They could also be employed for modelling seaborne activity along the Canadian coasts, and help in defining surveillance requirements. Finally, because maritime traffic is a predominant source of low-frequency ambient noise, density maps could be used to refine assessments of the impact of shipping on sonar system performance.

References

1. Fong, V. and Doré, S. (2003). High Frequency Surface Wave Radar Site Optimization Study. ORD Project Report PR 2003/20, Department of National Defence, Canada.

2. Gauthier, Y. and Bourdon, S. (2004). Performance, Benefits, and Costs of Long Endurance UAVs for Domestic Maritime Roles. ORD Project Report PR 2004/14, Department of National Defence, Canada.

3. Fong, V. (2005). A Simulation Study of Multi-Sensor Maritime Surveillance using SIMLAB (draft). DRDC-CORA Technical Report, Centre for Operational Research and Analysis, Ottawa, Canada.

4. Hill, A. (2003). Integration of RADARSAT and Vessel Monitoring System for Maritime Surveillance. Unpublished DOR(Joint) study, Department of National Defence, Canada.

5. Dickinson, R. G. and Boivin, S. E. (1996). CANLANT Surface Surveillance : Analysis of Reporting. Research Note RN-0396, Maritime Forces Atlantic Headquarters, Halifax, Canada.

6. U.S. Naval Oceanographic Office (2000). Database description for the Historical Temporal Shipping Traffic Database - Variable Resolution (HITS-V 1.1). Technical Report. Mississippi 39522-5001.

7. Yao, J. Z. and Barnes, A. E. (1975). On the Estimation of Shipping Densities from Observed Data. Planning Systems Incorporated Report PSI-TR-004018.

8. Lange, K., Bahn, M., and Little, R. (1987). Theoretical Study of Some Maximum Likelyhood Algorithms for Emission and Transmission Tomography. *IEEE Trans. on Medical Imaging*, **6**(2).

9. Mohammad-Djafari, A. (2003). Bayesian Approach for Data and Image Fusion. AIP Proc. of the 22nd International Workshop on Bayesian Inference 659(1) pp.386-408.

10. Sharma, R. K. (1999). Probabilistic Model-Based Multi-Sensor Image Fusion. PhD dissertation. Oregon Graduate Institute of Science and Technology. Department of Electrical and Computer Engineering.

11. Yu, P. L. H. (1994). A Simple Statistical Project: Image Reconstruction. *The American Statistician*, **48**(1).

12. MARPAC OR Team Program of Work (2005). Support to Operation SEA LION. Maritime Forces Pacific Headquarters, Victoria, Canada.

13. James, W. and Stein, C. (1961). Estimation with Quadratic Loss. In *Proceedings of the Forth Berkeley Symposium on Mathematical Statistics and Probability*, Number 1, pp. 311–379.

14. Emond, E. J. (1988). An Alternate Approach to the Problem of Estimation, Ranking, and Multiple Comparison of Mean Values. ORAE Staff Note 2/88, Department of National Defence, Canada.

15. Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1990). Numerical Recipes in C - The Art of Scientific Computing, Cambridge University Press.

16. Flanagan, M. T. (2004). Java Library Regression Class. http://www.ee.ucl.ac.uk/∼mflanaga/java/Regression.html.

17. Taylor, J. R. (1997). An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements, Sausalito, CA: University Science Books.

Annex A
SEASCAPE

A.1 Getting started

The *Statistical Evaluator of Average Shipping and Coastal Activity Picture* (SEASCAPE) is a tool implementing the algorithm of section 2 and allowing its user to fuse report density maps obtained from multiple sources into a single picture - a picture that represents the most-likely shipping density for the time period of the original maps. A copy of the tool can be obtained by contacting MARPAC ORT.

The program needs no installation, but it requires that you have the Java® Runtime Environment (JRE) installed on your computer in order to run (preferably the latest version, at least version 1.4.2). It also requires Java® Advanced Imaging (JAI) libraries to be installed on your machine. Both JRE and JAI can be downloaded for free from http://java.sun.com.

Other than these requirements, SEASCAPE does not need any software to be installed. Simply place the seascape.jar file in the directory from which you plan to run the program and double-click on it to start.

A.2 How to use SEASCAPE

A.2.1 The main menu

The main menu is the control centre for SEASCAPE. Data fusion operations are launched from this menu. SEASCAPE requires that you have a setup file to describe the density maps that will be merged to derive the most likely density map. Once the fusion is complete, the data can be ported to ESRI® ArcToolBox for producing high quality maps; however, SEASCAPE can also give a quick overview of the report density map used in input, as well as those produced in output.

A.2.2 Create a setup file

In order to create a setup file you can either go to the *Setup File Editor* screen or create one from scratch using a text editor (see Appendix 1 for a description of the file format). To access the editor, choose *Create/Modify Setup Files* under the *File* menu. Figure A.1 shows a snapshot of the setup file editor.

You can both edit and create setup files using the editor. To create a new setup file, press the *New Setup File* button; to load an existing setup file press the *Load* button. To save the current setup file, press the *Save* button. If you wish to save changes made to a setup file in a new file, you can also use the *Save As…* button.

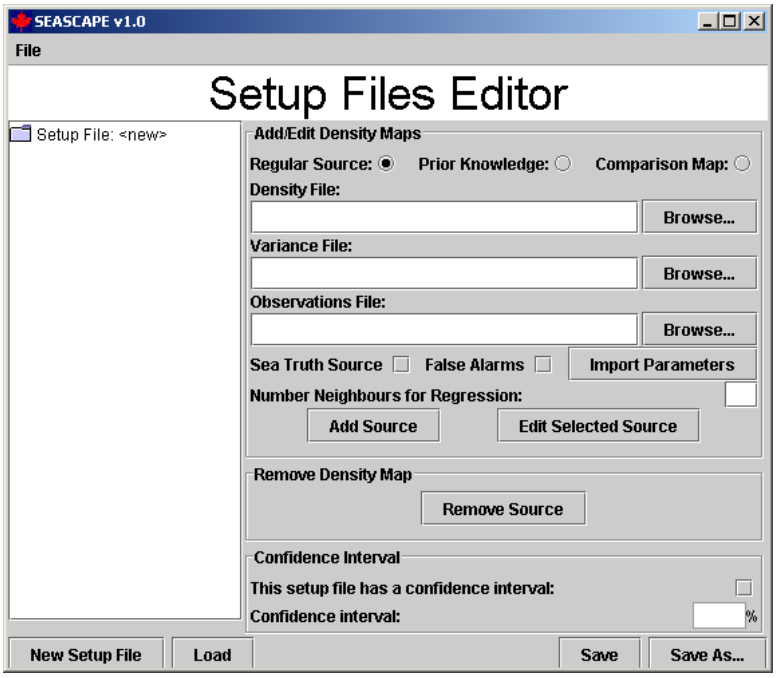

Figure A.1: *Setup file editor screen.*

Three files per data source are minimally required to run the program:

1. A **report density file**, which for the ith data source provides the mean number of ships counted $d_i = \sum_{j=1}^{N_i} n_i(t_j)/N_i$;

2. A **variance file**, which for the ith data source provides the variance on the number of ships counted $\sigma^2_{n_i} = \frac{1}{(N_i-1)} \sum_{j=1}^{N_i} [n_i(t_j) - d_i]^2$. Note that this is not the variance $\sigma^2_{d_i}$ presented in Equation 13;

3. An **observation file**, which for the ith data source provides the number of counts N_i.

These files describe the values of d_i, $\sigma^2_{n_i}$, and N_i over space, respectively. They are all required to use the same grid file format presented in Appendix 2.

The *Sea Truth Source* checkbox allows you to mark any source assumed to describe sea truth in terms of the ship counts provided. If one source is selected as such, then the reference map used to calculate α and β for the non-calibrated, non-sea truthing sources is simply the density map from this source. Otherwise, a weighted mean from all sources identified as sea-truthing is computed to define the reference map, with the values of $N_i/\sigma^2_{n_i}$ as weights. The values for α and β for the sources identified as sea-truthing are respectively set to 1 and 0 by default.

The *Potential for False Alarms* checkbox allows you to mark a source possibly counting ships that do not really exist. If this box is not checked, then β_i is assumed to be zero.

In cases where the values of $\alpha_i \pm \sigma_{\alpha_i}$ and $\beta_i \pm \sigma_{\beta_i}$ have already been determined in the past, the values can be imported by pressing the *Import* button. You will then be asked to provide the location of the grid files containing the parameters. If the parameters are not readily available, then they must be estimated through local analysis (Case 2 of section 2.3). The *Number Neighbours for Regression* is the number of neighbours on each side of a given cell that will be used as the local analysis window during the regression. For example, if you enter '2' as a value, the local analysis window will be five cells by five cells i.e. two above, two below, two to the right, and two to the left of the cell that is having the regression conducted on it. The entered value must be an integer. The box cannot be left empty, so dummy values should be entered even if the source is sea-truthing or if α_i and β_i values are imported.

A prior knowledge map can also be entered using the set-up screen. A prior knowledge map is a map displaying an initial estimate of the densities for the period of interest. For prior knowledge, the variance file must include the squared errors on the initial estimates of densities. SEASCAPE requires both a density file and a variance file, but not an observations file. Remember to click the *Prior Knowledge* radio button before adding the source.

A comparison map can be provided here as well. If provided, then the two metrics of section 3.2.1 will be computed to determine how the fused map measures up against the comparison map. Since the metrics involve only densities, SEASCAPE requires that you provide only the density file.

Once you have entered the data for a source, you can add it with the *Add Source* button. If you had already selected a source before you entered the data, you can choose instead to edit that source with the *Edit Selected Source* button. You may also remove a source simply by selecting it and pressing the *Remove Source* button. Note that the *Edit Selected Source* button must be pressed in order to apply changes to a source before saving.

The bottom portion of the set-up screen determines if Bayesian confidence intervals for shipping densities should be computed in addition to most-likely densities. If so, the box must be checked and the level of confidence entered, in percentage. The confidence intervals must be interpreted here as 'probable ranges'. For instance, a 90% interval will be limited by the values of ρ_{low} and ρ_{high} such that $\mathcal{P}(\rho \leq \rho_{\text{low}}|\mathbf{d}) = 5\%$ and $\mathcal{P}(\rho \geq \rho_{\text{high}}|\mathbf{d}) = 5\%$.

A.2.3 Fusing Maps

Once you have a setup file created and saved, you can use it to create a most-likely density map. Map fusion is launched from the main screen. To fuse maps you must enter the setup filename (see Figure A.2) that describes the input maps you want to fuse. This can be done either by typing the absolute path and filename in the text box or by browsing using the *Browse...* button.

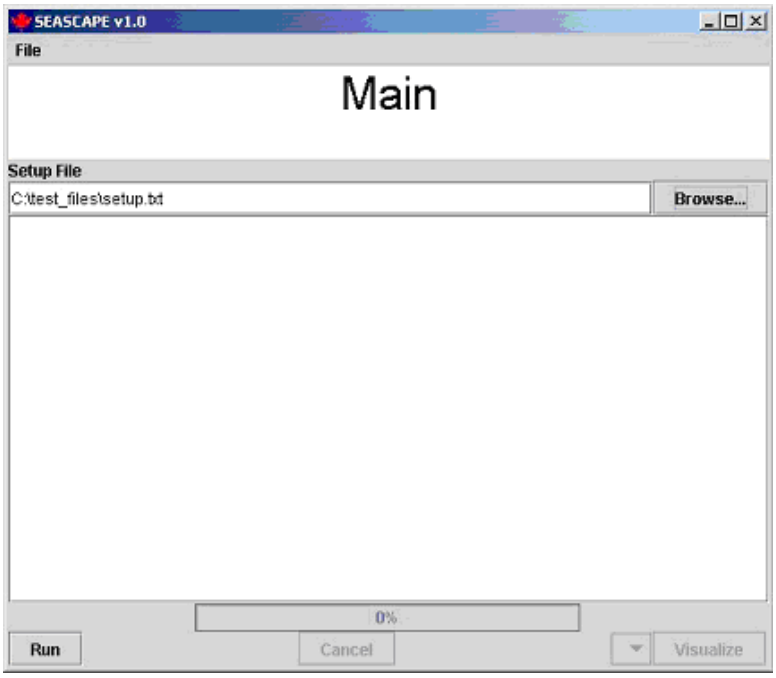

Figure A.2: *Choosing a setup file.*

Once a setup file has been chosen, you must press the *Run* button to begin fusing the maps. When you press the *Run* button, a dialog is presented for you to choose the file that will store the fused density map, as shown in Figure A.3. The directory of the selected file will also be used to save other grid files, including the estimated values of α_i, σ_{α_i}, β_i, and σ_{β_i}, as well as the boundaries of the confidence intervals.

After the output file name and destination have been chosen, the system will begin fusing the maps (Figure A.4). If an error should occur, the reason will appear on the display. Progress is also recorded on the display. The fusion can be cancelled at anytime and the output file will not be created (or if it already exists, it will not be overwritten). If you provided a comparison map, SEASCAPE will compare it to the generated map as well as the input maps when the fusion is complete. The metrics values of section 3.2.1 will be displayed. Confidence intervals will also be computed if requested.

A.2.4 Visualization

After map fusion is complete, you can visualize the results. To view the fused map, choose *Most Likely Density* from the dropdown menu beside the *Visualize* button, as shown in Figure A.5. If you wish to compare any (or all) of the source maps or the comparison map, you can visualize many of them simultaneously by choosing the appropriate maps from the dropdown menu and pressing the *Visualize* button again. Note that maps representing calibration parameters or confidence interval limits can be displayed as well.

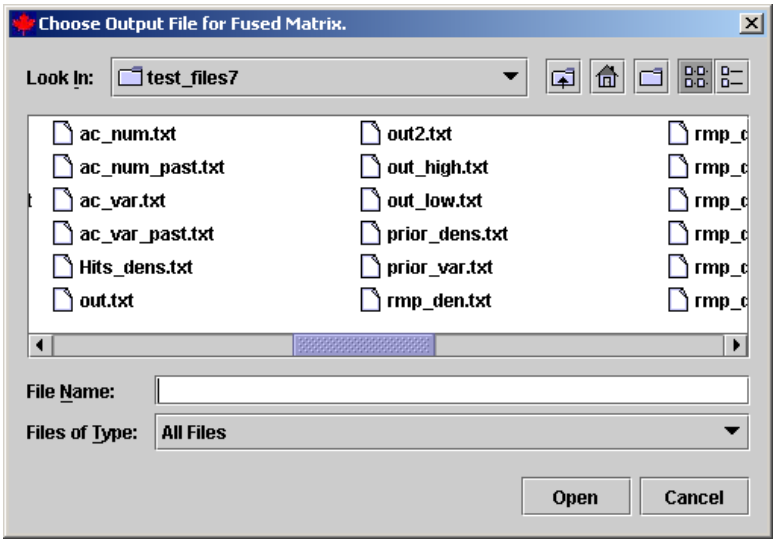

Figure A.3: *Choosing output file.*

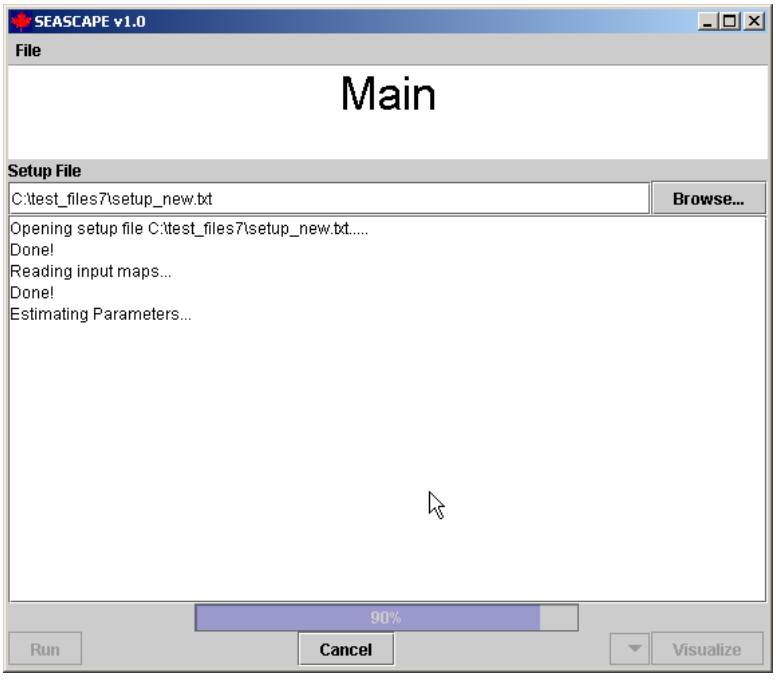

Figure A.4: *Map fusion in progress.*

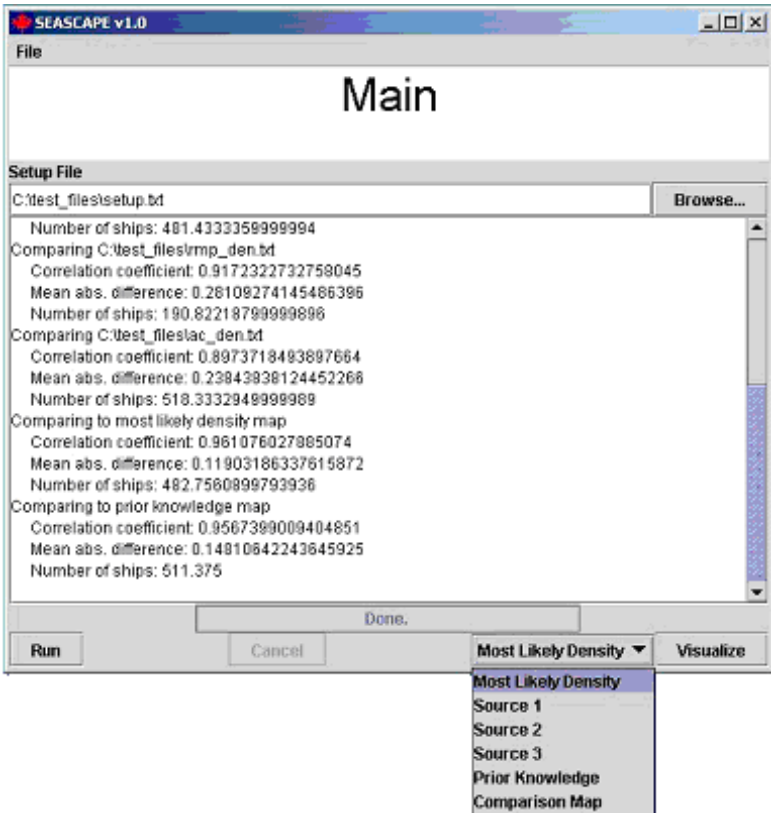

Figure A.5: *Visualization options.*

Examples of visualization screens are shown in Figure A.6. The density in a specific cell can be obtained just by placing the mouse cursor over the desired location. The latitude, longitude, and density (between brackets) appear on the lower-right corner of the screen. The number of observations, $N_i(x, y)$, used to produce the map can be displayed instead of the densities. By right-clicking on the screen, it is also possible to magnify, rotate, move or zoom on the displayed maps.

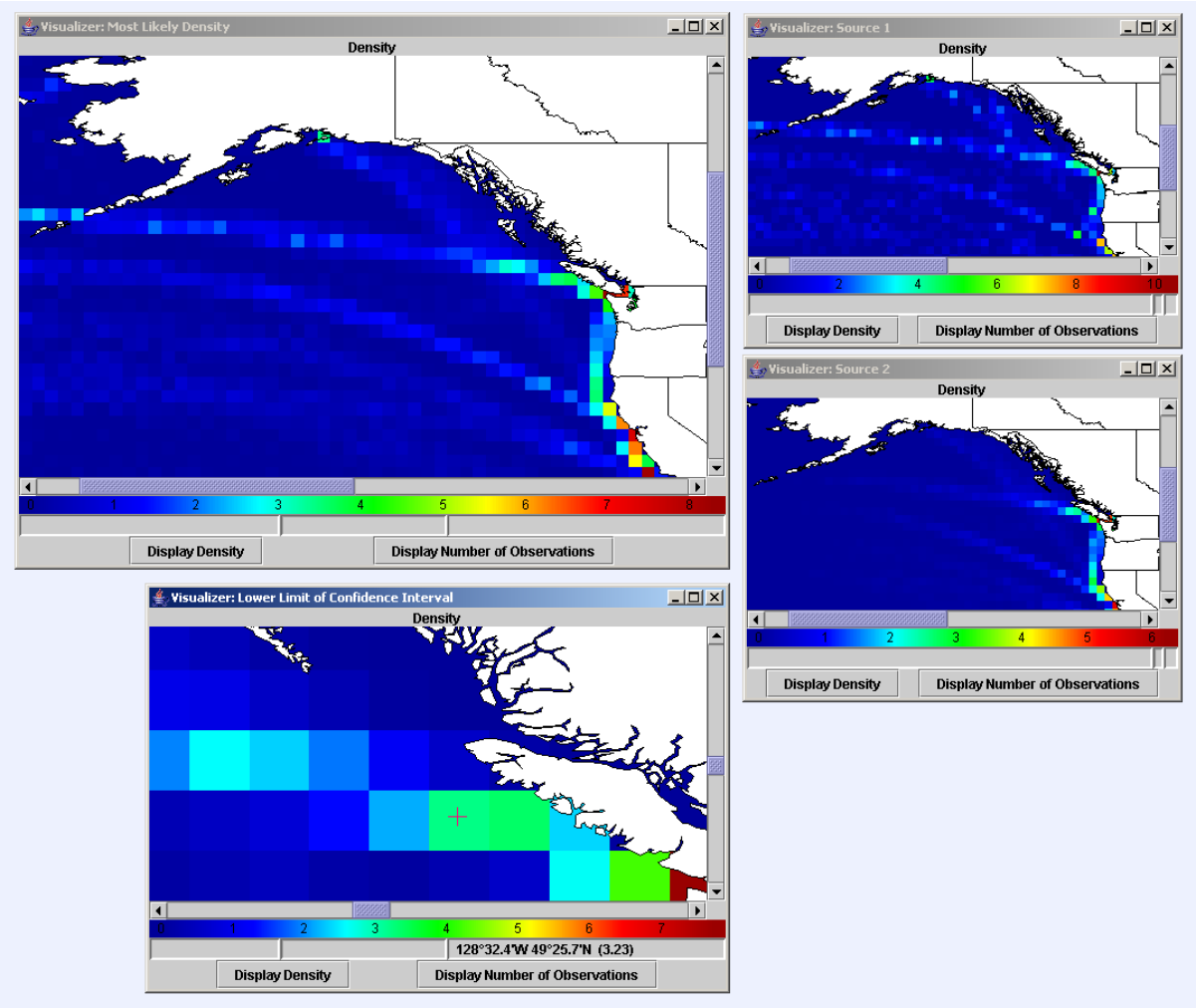

Figure A.6: *Examples of visualization screens.*

Appendix 1: SEASCAPE setup file format

SEASCAPE setup files are split into four main sections. The first section describes the input maps to be fused. The second section describes the prior information (prior map). The third section describes the availability of a comparison map. Finally, the last section describes the size of the confidence intervals to be produced. As a user, you have the choice to build the setup files manually using a text editor like Notepad, or use the built-in setup file editor of SEASCAPE. User-created files and editor-created files will both work the same in SEASCAPE and its setup file editor.

The opening section must tell the system how many sources (or input density maps) will be fused. This is accomplished with the line 'N_SOURCES <*number of sources to be fused*>'. Then, for each source, the setup file must describe some important information. First, the three files that describe the density map; this is accomplished with the three following lines: 'DEN_FILE <*absolute file path*>', 'OBS_FILE <*absolute file path*>, and 'VAR_FILE <*absolute file path*>'. Second, whether or not the source describes sea truth; this is accomplished with the line 'TRUTH <*YES/NO*>'. Third, whether or not the source could produce false alarms; this is accomplished with the line 'FALSEALARMS <*YES/NO*>'. Fourth, the radius of surrounding cells that will be used for the local analysis window when the system does regression calculations; this is accomplished with the line 'N_NEIGHBOURS_FILTERING <*number of cells for local analysis window*>'. If the parameters $\alpha_i \pm \sigma_{\alpha_i}$ and $\beta_i \pm \sigma_{\beta_i}$ are already available, the user can import them directly to avoid estimating them. This is done with the line 'IMPORT <*NOTHING/ALPHA/ALPHA_BETA*>' establishing which parameters are available. The lines that follow provide the locations and filenames of the associated parameters.

The second section must establish if there is prior knowledge to be considered; this is done with the line 'PRIOR_KNOWLEDGE <*YES/NO*>'. If there is a prior map to consider, then there must be two lines that give the paths to the two files that respectively describe the density and variance matrices.

The third section describes on one line whether or not there is a file for comparison and what the file is. The syntax for that line is as follows: 'COMPARISON_MAP <*YES/NO*> <*absolute file path*>'.

Finally, the fourth section also describes on a single line whether or not confidence intervals should be computed in addition to the fused map. The syntax for that line is 'CONFIDENCE_INTERVAL <*YES/NO*> <*level of confidence*>', where the level of confidence should be greater than 50%, and is typically 90.0% or 95.0%.

An example of a setup file follows:

```
N_SOURCES 3

DEN_FILE C:\density_map_folder\sat_den.txt
NUM_FILE C:\density_map_folder\sat_num.txt
VAR_FILE C:\density_map_folder\sat_var.txt
TRUTH NO
FALSEALARMS YES
N_NEIGHBOURS_FILTERING 12
IMPORT NOTHING

DEN_FILE C:\density_map_folder\rmp_den.txt
NUM_FILE C:\density_map_folder\rmp_num.txt
VAR_FILE C:\density_map_folder\rmp_var.txt
TRUTH NO
FALSEALARMS NO
N_NEIGHBOURS_FILTERING 2
IMPORT NOTHING

DEN_FILE C:\density_map_folder\ac_den.txt
NUM_FILE C:\density_map_folder\ac_num.txt
VAR_FILE C:\density_map_folder\ac_var.txt
TRUTH YES
FALSEALARMS NO
N_NEIGHBOURS_FILTERING 5
IMPORT ALPHA_BETA
ALPHA C:\density_map_folder\sat_den_alpha.txt
SIGMAALPHA C:\density_map_folder\sat_den_sigmaalpha.txt
BETA C:\density_map_folder\sat_den_beta.txt
SIGMABETA C:\density_map_folder\sat_den_sigmabeta.txt

PRIOR_KNOWLEDGE YES
DEN_FILE C:\test_files\prior_dens.txt
VAR_FILE C:\test_files\prior_var.txt

COMPARISON_MAP YES C:\test_files\true_dens.txt

CONFIDENCE_INTERVAL YES 90.0
```

Appendix 2: Grid file format

The grid files used by SEASCAPE are plain text (ASCII) files. The formatting of the files is compatible with the conversion utilities of ESRI® ArcToolBox. The file template is the following:

NCOLS <*value*>
NROWS <*value*>
XLLCENTER <*value*>
YLLCENTER <*value*>
CELLSIZE <*value*>
NODATA_VALUE <*value*>
<*row 1*>
<*row 2*>
.
.
.
<*row n*>

where NCOLS is the number of columns in the ASCII file, NROWS is the number of rows, XLLCENTER is the x coordinate (longitude) for the center of the lower left most cell in the grid, YLLCENTER is the y coordinate (latitude) for the center of the lower left most cell in the grid, CELLSIZE is the length of a cell's edge, NODATA_VALUE is the value in the ASCII file representing unknown values, and <*value*> are numbers. Note that the cell values are space-delimited.

Here is an example of a file suitable for SEASCAPE and conversion with ArcToolBox:

NCOLS 53
NROWS 33
XLLCENTER -169
YLLCENTER 34.5
CELLSIZE 1
NODATA_VALUE -9999
0.233333 0.200000 0.066667 -9999.000000 -9999.000000 -9999.000000 -9999.000000 . . .
0.133333 0.133333 0.200000 0.200000 -9999.000000 -9999.000000 -9999.000000 . . .
0.400000 0.366667 0.966667 0.633333 0.166667 -9999.000000 -9999.000000 . . .
.
.
.

List of symbols/abbreviations/acronyms

CORA	Centre for Operational Research and Analysis
CF	Canadian Forces
DND	Department of National Defence
DRDC	Defence Research and Development Canada
JAI	Java Advanced Imaging
JRE	Java Runtime Environment
MARPAC	Maritime Forces Pacific
MPA	Maritime Patrol Aircraft
OR	Operational Research
ORT	Operational Research Team
SEASCAPE	Statistical Evaluator of Average Shipping and Coastal Activity Picture

This page intentionally left blank.

DOCUMENT CONTROL DATA

(Security classification of title, body of abstract and indexing annotation must be entered when document is classified)

1. ORIGINATOR (the name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g. Centre sponsoring a contractor's report, or tasking agency, are entered in section 8.)

DRDC – Centre for Operational Research and Analysis
MARPAC HQ, PO Box 17000 Stn Forces, Victoria BC V9A 7N2

2. SECURITY CLASSIFICATION (overall security classification of the document including special warning terms if applicable).

UNCLASSIFIED

3. TITLE (the complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S,C,R or U) in parentheses after the title).

Multi-Source Data Fusion for the Estimation of Mean Shipping Densities

4. AUTHORS
(Last name, first name, middle initial. If military, show rank, e.g. Doe, Maj. John E.)

Gauthier, Yvan ; Minter, Peter

5. DATE OF PUBLICATION (month and year of publication of document)	6a. NO. OF PAGES (total containing information. Include Annexes, Appendices, etc).	6b. NO. OF REFS (total cited in document)
August 2005	44	17

7. DESCRIPTIVE NOTES (the category of the document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of report, e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered).

Technical Memorandum

8. SPONSORING ACTIVITY (the name of the department project office or laboratory sponsoring the research and development. Include address).

DRDC – Centre for Operational Research and Analysis
MARPAC HQ, PO Box 17000 Stn Forces, Victoria BC V9A 7N2

9a. PROJECT OR GRANT NO. (if appropriate, the applicable research and development project or grant number under which the document was written. Specify whether project or grant).	9b. CONTRACT NO. (if appropriate, the applicable number under which the document was written).
N/A	

10a. ORIGINATOR'S DOCUMENT NUMBER (the official document number by which the document is identified by the originating activity. This number must be unique.)	10b. OTHER DOCUMENT NOs. (Any other numbers which may be assigned this document either by the originator or by the sponsor.)
DRDC CORA TM 2005–18	

11. DOCUMENT AVAILABILITY (any limitations on further dissemination of the document, other than those imposed by security classification)

(X) Unlimited distribution
( ) Defence departments and defence contractors; further distribution only as approved
( ) Defence departments and Canadian defence contractors; further distribution only as approved
( ) Government departments and agencies; further distribution only as approved
( ) Defence departments; further distribution only as approved
( ) Other (please specify):

12. DOCUMENT ANNOUNCEMENT (any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, where further distribution beyond the audience specified in (11) is possible, a wider announcement audience may be selected).

13. ABSTRACT (a brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), (R), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual).

Maintaining continuous tracks of all vessels crossing a large oceanic area is a very difficult task to perform. For this reason, the true mean shipping densities over a given period of time can rarely be directly calculated; they need to be estimated from a limited number of observations made at discrete points in time. The completeness and correctness of these observations may vary from one data source to another. This paper proposes a Bayesian fusion method for producing shipping density maps of high resolution based upon vessel position reports obtained from various sources of maritime traffic data. An example involving data from three sources with different characteristics is given. A computer program implementing the fusion algorithm is also presented.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus. e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus-identified. If it not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title).

Bayesian statistics
Data fusion
Maritime traffic
Operational research
Shipping densities

DRDC CORA

www.drdc-rddc.gc.ca