Analysis Plan: Vessel Operations and Compliance Statistical Analysis

Christopher J. Fonnesbeck

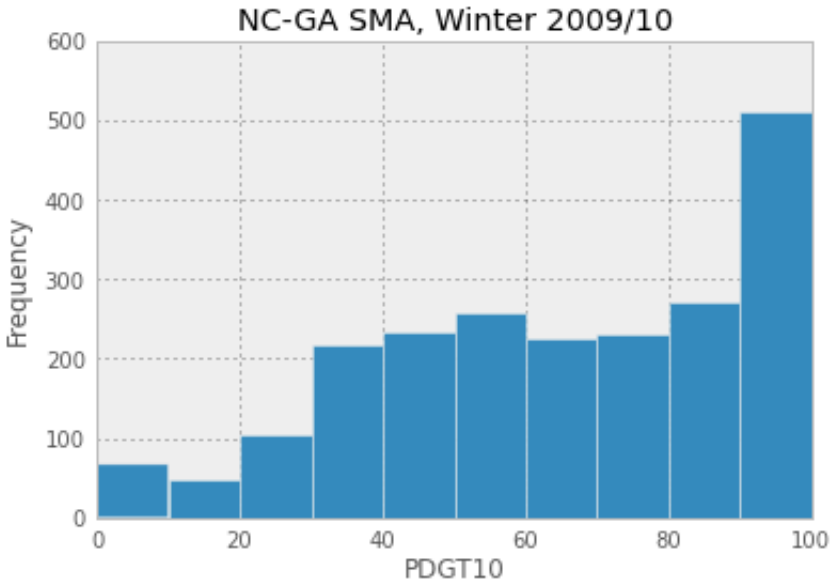
19 April, 2013

This document outlines an analysis plan for Automatic Identification System (AIS) data as it relates to the adherence of commercial vessels to speed restrictions in seasonal management areas (SMA) across the eastern seaboard. Specifically, we seek to estimate the relationship between outreach and notification programs and compliance with vessel speed rules.

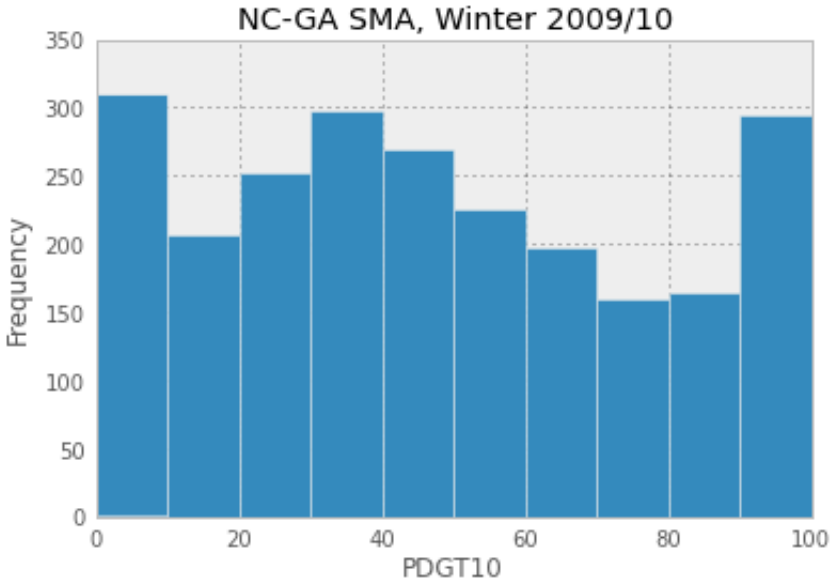
Estimation of Compliance Differences Between SMA Active Periods

A simple approach to summarizing the change in vessel behavior over time is to compare summary statistics of compliance between SMA seasons. I recommend estimating the difference in compliance (using metric of choice) between adjacent SMA seasons, along with an associated measure of uncertainty, such as a confidence interval or a Bayesian credible interval. An estimation approach is generally preferred to the use of statistical tests because, given the large dataset, many of the tests are going to be significant even if there are very small differences between seasons. Thus, an estimate of effect size is preferable because it is more informative; we know that any two seasons will not be exactly equivalent, so the question is “how different are they?”

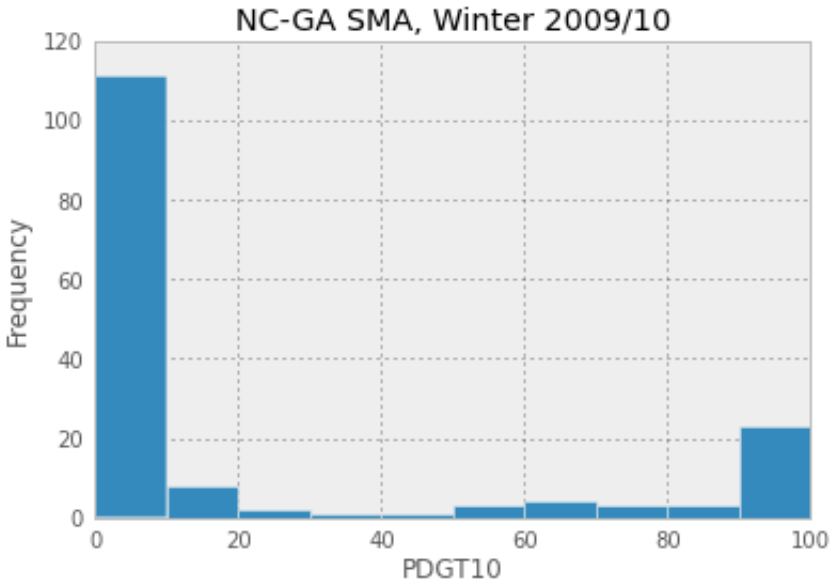
Based on a preliminary examination of the data, compliance appears to vary strongly enough that we should consider calculating mean compliance and differences within each SMA, and possibly for each ship type within SMA. For example, the distribution of PDGT10 for cargo vessels during the winter 2009/10 SMA was as follows:



Whereas, the distribution of the same ship type during the following year's SMA was:



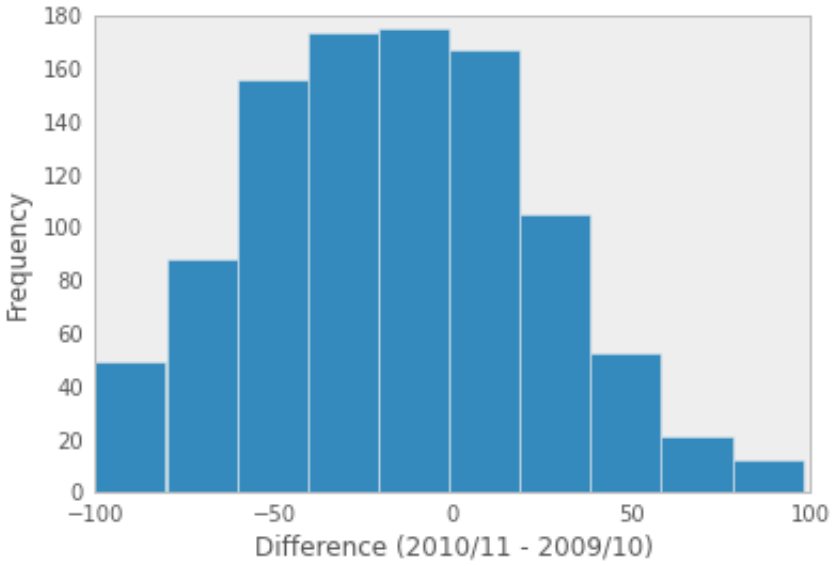
There is an even stronger contrast between ship types; the NC-GA 2009/2010 SMA PGDT10 distribution for all towing vessels was:



As these figures illustrate, it may not be very informative to simply calculate a mean difference between adjacent SMAs, even within ship type. Instead (or additionally), we need to be able to characterize the difference in *distributions* of compliance measures.

There are a variety of non-parametric tests that compare the two distributions, such as the Cramer-von Mises test and the Kolmogorov-Smirnov test. These compare two cumulative empirical distributions via some measure of similarity. However, these tests are non-directional; they simply look to see if two distributions are different (in a goodness-of-fit sense), and not whether one distribution is “larger” than another. For example, distributions may differ in shape, rather than location.

An alternate approach is to resample the empirical distributions of compliance for two SMAs, and for each pair of samples, calculate the difference. This results in a distribution of resampled differences from which a mean and standard deviation could be calculated. For example, the distribution of resampled differences between the two cargo ship SMA distributions above is shown in the figure below:



The mean of this distribution is -20.3, but the standard deviation is 40.8.

Change-point Analysis

If a particular enforcement program has been effective in increasing compliance with vessel speed rules, we might expect to be able to detect a relatively discrete change in behavior that corresponds to the implementation of a particular program. Consider a single effective outreach program: A change-point model considers a time series of observed vessel speeds or compliance rates being subdivided into two periods, one before the program implementation and another after, with the mean compliance lower in the first period relative to the second, as a result of the program's effectiveness. We may not be able to anticipate exactly when during the time series that the population of vessels changed behavior, or perhaps even which program was effective (if any). As a result, the change-point model treats the location of the switch as unknown, and thus a variable in the model to be estimated. The estimation procedure uses the information contained in the observed vessel speeds or compliance rates to yield the most likely location of the change-point. If the estimated location is within some reasonable period following the implementation of a particular enforcement or notification program, it may be considered evidence in favor of that program's effectiveness.

Specifically, we will model two means μ_{pre}, μ_{post} (though this can be expanded to three or more, if we believe there are more changepoints) representing the mean of the metric for compliance before and after the changepoint. For now, we will assume that PDGT10 is the metric of choice. Since PDGT10 is a proportion, it is modeled on the $[0, 1]$ interval, but we can logit- or probit-transform this value so that the mean can be modeled on the real line:

$$E(x_i) = \mu_j,\ j = pre, post$$

$$\text{logit}(\mu_j) = \theta_j$$

The simplest model for the mean θ is that it is a single, unknown value throughout the interval, in which case we assign some diffuse prior distribution to model our uncertainty regarding its true value:

$$\theta_j \sim N(0, 1e4)$$

More realistically, we may opt to model the expected value as varying randomly and/or as a function of one or more covariates:

$$\text{logit}(\mu_j) = \theta_j + X\beta + \epsilon_j$$

$$\epsilon_j \sim N(0, \tau)$$

Covariates that may lead to variation on compliance might include country of origin (flag) or ship type.

The change point can be modeled as the month, week or day that separates the first and second intervals. If there is no prior information regarding the location of the change point, this is most easily modeled as a discrete uniform random variable:

$$s \sim \text{DiscreteUniform}(t_0, t_n)$$

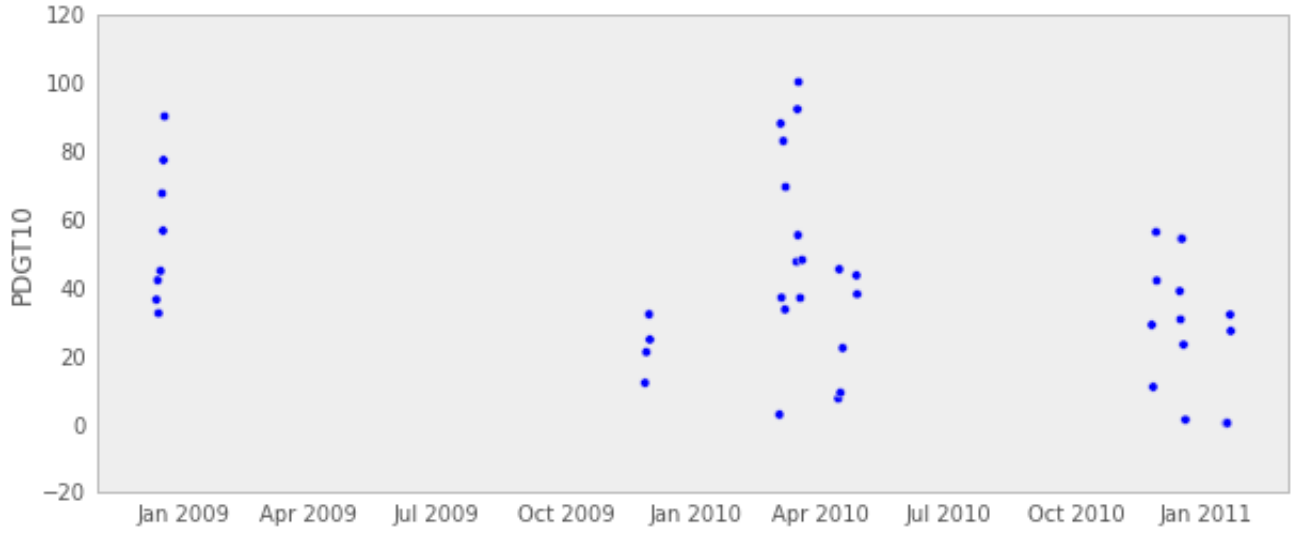
where t_0 and t_n are the first and last times (day, week, month) in the dataset. Hence the pre interval is $[t_0, s)$ and the $post$ interval is $[s, t_n]$.

A change-point model will be implemented across all SMA and ship types, in order to estimate the response by the entire regulated community. In addition, it may be worthwhile executing the model on prespecified subgroups of interest. These will likely include an analysis of cargo ships only, which make up the majority of the traffic in some SMAs and are among the fastest ships, as well as passengers vessels, which are also fast. These analyses can be replicated both across and within SMAs.

Vessel Speed Time Series Analysis

A third approach is to examine the AIS data longitudinally. That is, rather than treating each segment as independent, analyze the time series of segments across ships. Within the current dataset, there is a mean of 20 segments per mmsi record (median 6, max 1631), so most vessels have multiple segments and would hence be admissible for such an analysis.

For example, below is the time series of segment PDGT10 values for a randomly-chosen ship:



Thus, we have observations clustered within ships; ships, in turn, may appear in one or more SMAs. The simplest implementation of a time series model for such data posits a relationship between the outcome (a compliance metric) and a set of predictors that explicitly includes time. If we are using PDGT10, for example, we must also perform a logit transformation on the model.

$$\text{logit}[Pr(y_{it})] = \mu_i + \alpha_i t + X\beta_i + \epsilon_{it}$$

where μ_i is a mean compliance for mmsi record i, α_i is a parameter for the effect of time, X is a set of predictors (e.g. flag, ship type, etc.) and β a vector of corresponding parameters. ϵ_{it} is an observation random effect, which is typically distributed according to some parametric distribtuion, such as a Gausssian:

$$\epsilon_{it} \sim N(\mu_\epsilon, \sigma_\epsilon^2)$$

If we do not want to make parametric assumptions, we may instead implement some nonparametric distribution, such as a Dirichlet process or a Polya tree mixture. Finally, we may expect that changes in compliance may interact in some way with other factors. In this case, we can add additional interactive terms $X'(\beta' t)$, where X' is a subset of X that interact with time.

Of primary interest would be to check if α is negative, representing increased compliance over time. This parameter is indexed by mmsi so that it may be represented as a random effect, or as a function of ship-level covariates, since we do not expect all ships to respond in the same way to regulations and enforcement. The simple formulation above posits a linear relationship between time and compliance, but more complex relationships (e.g. quadratic, cubic, etc.) can be implemented easily.