

USING THE ARIMA MODELING CAPABILITIES OF SAS® TO FORECAST RATES OF FEBRILE RESPIRATORY INFECTION AT EIGHT U.S. MILITARY RECRUIT TRAINING CENTERS

Christian J. Hansen; Anthony W. Hawksworth;
Phillip I. Good PhD; Margaret AK Ryan, MD MPH

Department of Defense Center for Deployment Health Research
Naval Health Research Center

ABSTRACT

Auto-Regressive Integrated Moving Average (ARIMA) forecasting may be used as a method of detecting aberrations and quantifying excursions from normal behavior in time series data. Using SAS® system's PROC ARIMA, weekly counts of health care visits for febrile respiratory infection (FRI) were forecast using several years of FRI count data collected regularly from eight DoD recruit training centers. The macro processing capabilities of SAS® were used to increment successive forecasts, and to collect and evaluate the accuracy of multiple models. The capability of PROC ARIMA to include covariate data was used to expand the model to include total recruit population as well as the number of new recruits arriving weekly. Actual FRI counts were compared with ARIMA forecast counts, with differences measured in standard deviations from the forecast value. This paper describes some of the capabilities and techniques of ARIMA forecasting and macro processing through SAS®.

Introduction

Early detection of respiratory disease outbreaks as well as those that may be due to bioterrorism, is essential to the success of intervention methods. This project represents an attempt to improve current analysis of surveillance data to more quickly and accurately quantify deviations from typical behavior. Current methods utilize thresholds based on the number of febrile respiratory infections (FRI) per 100 healthy recruits. However, the use of fixed thresholds does not adjust for procedural differences between the different service branches. It may be possible to characterize the early behavior of an outbreak to provide more advanced indication of a significant event.

Data Sources

Data for this project were regularly collected as part of ongoing surveillance efforts at Naval Health Research Center. Data were collected from eight recruit training centers throughout the United States.

The outcome of interest and the target of the ARIMA forecast surveillance effort is weekly counts of febrile respiratory infections (FRI) at each recruit center. These counts represent recruits identified as having both a respiratory infection and an elevated oral temperature. The weekly totals are produced by medical service personnel or by research assistants at each sites medical facilities.

Estimates of weekly total base population were also available for each surveillance site. This allowed a rate of infection as a proportion of the total population to be estimated. Additional information provided by some sites included the weekly number of new recruits arriving at the recruit center.

Hurdles in Aberration Detection

A rate of 1.5 FRI cases per 100 trainees has been the conventional threshold used to define an outbreak event. However, each surveillance site has a unique distribution of weekly rates. At surveillance site 2, a rate above 1.5 may not occur for several years, while at site 4, the mean rate exceeds the 1.5 threshold. (Figure 1.)

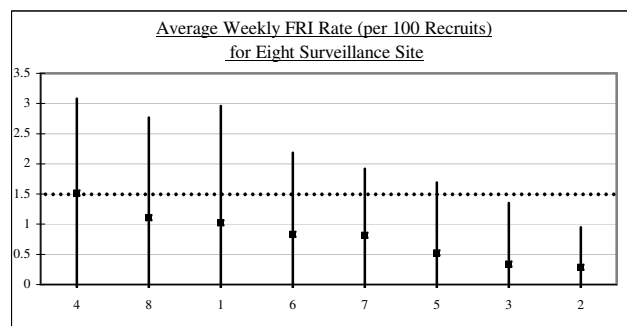


Figure 1. FRI rates at eight Department of Defense recruit center. The mean rate for recruit center 4 is above the conventional static threshold, while center 2 never achieved this threshold.

These differences may be the result of unequal surveillance effort between recruit centers, or procedural and cultural differences between military branches. Enhanced sensitivity in the identification of outbreak events requires a model inclusive of these differences. (Table 1.)

Recruit Center	1	2	3	4	5	6	7	8
Total Sick Counts	16,192	1,989	2,122	13,160	647	9,027	15,088	14,784
Average Weekly Population	9,003	4,020	4,501	4,846	737	6,138	9,906	7,284
Average FRI Rate	1.02	0.29	0.34	1.52	0.52	0.83	0.82	1.11

Table 1. Differences total FRI counts, population sizes, and average FRI rates at eight recruit centers.

Each of the eight recruit centers is unique. They are located in diverse geographic regions and maintain significantly different sized populations of recruits. However, most regularly experience a substantial increase in population during the summer months.

The duration of recruit training at each of the centers varies from eight to twelve weeks, depending on the branch of service. However, a percentage of recruits at each center do not complete training, while others require additional time and are present at each site longer than normal. (Table 2.)

Recruit Center	1	2	3	4	5	6	7	8
Person-years	31,336	13,993	11,946	16,773	2,566	21,363	34,480	25,355

Table 2. Total person time contributed by each surveillance site.

ARIMA Forecasting

Auto-Regressive Integrated Moving Average (ARIMA) forecasting is a statistical process developed by Stats Canada, and integrated into SAS Version 8. The process identifies and estimates correlations between points within time series data. Additionally, evaluation and inclusion of cross correlations with other predictive time series data are possible.

The relationships identified by ARIMA are used to produce a predictive model that generates forecast values beyond the end of the set. An error estimate is provided with each forecast value, enabling deviations from predicted values to be quantified.

In this model, weekly FRI counts were forecast using total base population as a covariate. When available, weekly estimates of new recruits arriving at each center were added as a covariate.

A limitation of ARIMA modeling is the requirement for a continuous set of time history data. Currently, all eight sites maintain continuous surveillance. However, as surveillance began on different dates, and each center maintains a different average population, total time contributed to each centers model differs.

ARIMA coding

In the following example, Proc ARIMA is used to evaluate two time history sets contained in the dataset "Fort" and produce forecast values.

```
Proc Arima data=Fort;
    identify var=Sick scan
alpha=0.1 crosscor=Pop;
*** Use an ARIMAX (3,0,0) model***;
    estimate p=3 q=0 input=Pop;
    forecast out=new;
```

The variable "Sick", a weekly record of FRI counts at recruit center 7, is the target variable to be forecast. The second variable, "Pop", is the weekly total base population, and is used as a predictive covariate. Additional predictive covariates may be easily added to the model and evaluated in this statement. Values for p and q in the estimate statement specify positions of auto-regressive factors and define the models moving average components.

Selection of the model parameters involved initial hand crafting to maximize forecast accuracy and reduce error

estimates. After a stable model had been chosen, the magnitude of the standard error estimates was evaluated to determine the offset thresholds used to prompt a change to yellow or red outbreak status.

Measures of deviation in error dimension

An advantage of ARIMA forecasting is the ability to quantify actual values with relation to the forecast values by using the error estimates of the forecast value. In this system, when other conditions are satisfied, instances when actual values exceeded the forecast value by 0.5 or 1 standard errors prompt the outbreak status to be elevated, often well below the traditional threshold (Figure 2).

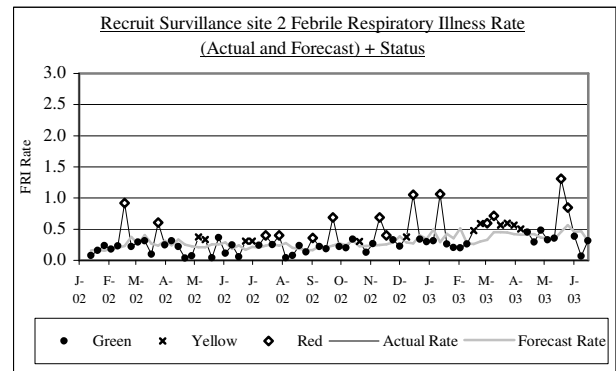


Figure 2. Time history of real FRI rates and week-ahead forecasts one Department of Defense recruit center. Note application of yellow and red status though at no point was the 1.5 threshold exceeded.

This technique allowed for a standard method quantifying the offset from different base populations. However, actual values exceed forecast values during periods when low counts are returning to more typical levels. During such circumstances, an increase in status may be indicated during a period when FRI rates are below average. This limitation was addressed by only identifying points when FRI rates were above a six month moving average (Figure 3).

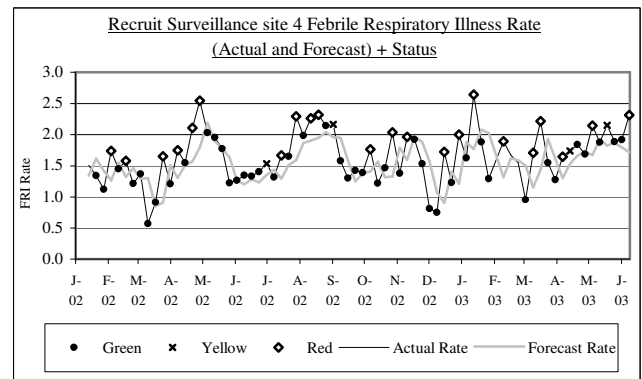


Figure 3. Note a substantial number of weeks lie above the traditional 1.5 threshold, but here are classified as green. Rapid increases returning rates towards the mean value were not allowed to elevate outbreak status.

Mixed and dynamic thresholds

The ARIMA model is specifically sensitive to early rate increases. It is important to remember that in this model, ARIMA prompts a change in outbreak status based on increasing FRI rates relative to forecasts based on recent behavior. A consistent, slow increase in FRI rates will not be identified as an event using ARIMA alone.

For all centers, it was determined that an absolute FRI rate should prompt an elevation of outbreak status if an ARIMA offset had not already done so. For each recruit center, a static threshold was set, representing the FRI rate exceeded less than approximately 5% of the duration of surveillance.

Macro to iterate and bundle output

The macro processing capability of SAS software was used to create and organize multiple successive predictions.

During the development of the model, it was necessary to evaluate the performance of different versions with archived data. Within the macro, a model is produced based on a truncated set of data, ending where forecasting is to begin. Each loop of the macro adds another value to the input set, and stores subsequent forecasts in an output dataset. The macro halts when values for the current week are added and a forecast for the future week is produced.

These simulations produce a history of forecasts that can be superimposed on real values. This technique helps to evaluate model accuracy using different collections of covariates, and during a variety of scenarios such as rapid changes in actual values, or when values are rapidly increasing or decreasing.

Predictably, when changes in FRI rates were unstable and rapid, the accuracy of forecasts was diminished. When rates rose and fell repeatedly, the models predictions often fell out of phase with actual counts. Under these conditions, the offset of predicted and actual counts would be accentuated, potentially yielding an incorrect classification of outbreak status.

Conclusion

There are many techniques for detecting aberrations in data. The flexibility of ARIMA forecasting in evaluating covariates is particularly useful when developing new models. ARIMA models are uniquely fitted to each situation and continuously adjust for changes in the population. In comparison with the use of fixed thresholds, ARIMA's moving average feature can more accurately quantify excursions from typical behavior by adjusting for trends.

The primary utility of this effort has been to create an aid to military health surveillance, and to evaluate conditions suspected of contributing to an outbreak event.

Acknowledgements

This project gratefully acknowledges the invaluable help of the following persons:

Naval Health Research Center

Jennifer Strickler
William K. Honner
LTC Joaquin Oronoz
Sandra Williams
MAJ Xiomara Brown
Johnnie Conolly
CPT Greg Martin
Robert Greenup
Susan Wolf
CAPT Aurelio Galati
CDR Sharon Ludwig
HS1 George McCall
CAPT Margan Zajdowicz
LT Justin Spackey
Ron Zupinski
LtCol James Neville
Maj John Lynch
LtCol Kevin Grayson
Dan Vestal
CDR Richard Williams
Robert Treston
CAPT Frank Chapman
LTJG Kurt Manley

Contact Information

Christian J. Hansen
Data Analyst
Department of Defense Center for Deployment Health Research
Naval Health Research Center
P.O Box 85122
San Diego, CA 92186-5512
(619) 553-7595
hansen@nhrc.navy.mil