



THE PHILIPPINE --- **STATISTICIAN** ---

Volume 64, Number 2 (2015)

The Official Publication of the
Philippine Statistical Association, Inc.

Contents

Editorial	v
Acknowledgement	vi
Invited Paper in Computational Statistics	
Nonparametric Bootstrap Test in a Multivariate Spatial-Temporal Model: A Simulation Study	1
<i>Abubakar S. Asaad and Erniel B. Barrios</i>	
Application in Epidemiology	
Some Zero Inflated Poisson-Based Combined Exponentially Weighted Moving Average Control Charts for Disease Surveillance	17
<i>Robert Neil F. Leong, Frumencio F. Co, and Daniel Stanley Y. Tan</i>	
Application in Agricultural Monitoring	
Modelling Rice Yield in the Philippines using Dynamic Spatio-Temporal Models	43
<i>Stephen Jun V. Villejo</i>	
Biostatistics	
Statistical Evaluation of In Vivo Micronucleus Assays in Toxicology	59
<i>John Closter F. Olivo</i>	
Application in Sensory Evaluation	
Bootstrapping Penalty Analysis in Sensory Evaluation of Pizza Products	69
<i>Catherine Estiaga</i>	
Application in Social Statistics	
Incidence of Crimes and Effectiveness of Interventions in the National Capital Region: Evidence from Panel Data	95
<i>Michael Daniel C. Lucagbo, Lianne S. De la Cruz, Jecca V. Narvasa, Micah Jane A. Paglicawan, and Jerizza P. Pepito</i>	
Application in Test Development	
The Recursive Alpha (RALph) Coefficients: Quantifying Inter-item Cohesion under Indirect Range Restriction	115
<i>Michael Van B. Supranes, John Francis J. Guntan, Joy Pauline Adrienne C. Padua and Joseph Ryan G. Lansangan</i>	
Guidelines for Authors	130

Editorial

This issue marks a remarkable accomplishment for *The Philippine Statistician* (TPS) and statistics as a discipline in the country. After over 60 years of circulation, the journal now gets the approval of the Content Selection and Advisory Board (CSAB) of Scopus. To quote the board,

This journal brings out the fact that there is a very active and excellent community of statisticians in the Philippines. The contributions are of real interest and importance. The technical level of the articles is good. I think the work of this community should be better known to their peers. I recommend this journal for inclusion in Scopus.

Now, the papers published in TPS will appear into the feed of Scopus and will be seen by the International Community. Our contribution for the growth of statistical science will now be exposed to a bigger audience, our international peer.

In this issue, two papers are exploring characteristics of spatio-temporal models. There are also three papers presenting various modeling strategies and application involving count data. However, all these papers are valuable contributions in computational statistics.

I invited my PhD student (**A. Asaad**) to share the results of his work on a nonparametric test in a multivariate spatial-temporal model. **J. Olivo** shared his paper on assays in toxicology. Poisson regression is used to extract information from a micronucleus assay. A control chart based on zero-inflated poisson regression is used by **R. Leong, F. Co, and D. Tan** in surveillance of dengue. **S. Villejo** developed a methodology based on a dynamic spatio-temporal model in monitoring rice yield.

Sensory evaluation has become a valuable tool in food product development. **C. Estiaga** used bootstrap penalty analysis in the evaluation of pizza products. **M. Lucagbo, L. De La Cruz, J. Narvasa, M. Paglicawan and J. Pepito** used negative binomial regression in evaluating effectiveness of crime mitigation strategies given a panel data.

Finally, **M. Supranes, J. Guntan, J. Padua, and J. Lansangan** measured inter-item cohesion under indirect range restriction, illustrated how these methods can be used in psychometric test development.

I would like to extend my sincerest gratitude to the contributors, reviewers, editorial board, editors, and The Philippine Statistical Association, Inc., for the continuous support to the Journal.

Erniel B. Barrios
Editor

ACKNOWLEDGEMENT

The Philippine Statistical Association, Inc., and the editors of *The Philippine Statistician* would like to recognize the valuable contribution of the following who served as referees from 2010-2015. Some have served several times as referee.

Acusta, Andrei
Alfonte, Caryl
Asaad, Abubakar
Balinas, Vicente
Bersales, Lisa Grace
Billones, Lisa
Campano, Wendell
Cruz, Grace
De Leon, Alexander
Del Prado, Grace
Dumanjug, Charlotte
Esmenda, Mary Jane
Gonzaga, Alex
Guarte, Jacqueline
Isidro, Marissa
Jamon, Grace
Javellosa, Mark
Lansangan, Joseph Ryan
Lavado, Rouselle
Kabamalan, Maria Midea

Magadia, Joselito
Magallanes, Aries
Mancenido, Michelle
Mapa, Dennis
Maranon, Antonio
Martinez Jr., Arturo
Mina, Christian
Paltino, Maylin
Parcon, Jason
Redoblado, Jade Eric
Reyes, Celia
Roscom, Brigida
Santos, Kevin Carl
Sarol, Jesus
Sarte, Genelyn
Sotto, Critina
Tejada, Jeffry
Miralao, Viginia
Vargas, Gregorio

Thank you very much.

Nonparametric Bootstrap Test in a Multivariate Spatial-Temporal Model: A Simulation Study

Abubakar S. Asaad

*Department of Epidemiology and Biostatistics
University of the Philippines Manila*

Erniel B. Barrios

*School of Statistics
University of the Philippines Diliman*

The assumptions of constant characteristics across spatial locations and constant characteristics across time points facilitates estimation in a multivariate spatial-temporal model. A test based on the nonparametric bootstrap is proposed to verify these assumptions. The simulation studies confirm that the proposed test procedures are powerful and correctly sized.

Keywords: coverage probability, robustness, spatial-temporal model

1. Introduction

A phenomenon that varies over space and time is sometimes viewed only from the spatial viewpoint or over the temporal perspective. The essence of spatial analysis is that “space matters” and what happens in one region is related to events in neighboring regions. This has been properly stated in what Tobler (1979) refers to as the First Law of Geography that states: “Everything is related to everything else, but closer things more so.” In the statistical science, this is aptly captured in the concept of spatial autocorrelation.

Spatial dependence refers to the correlation of an attribute measured at two locations. In the absence of spatial dependence, the distance of the two locations does not influence the joint behavior of attributes observed at those two locations. When spatial dependence is present (for example, positive correlation), then nearby observations are more similar than those far apart. The analysis can be enhanced by drawing on temporal dependencies, in addition to spatial dependence. Failure to account for time dependencies means lost information. Temporal dependence

means that events at one time can be influenced by what has happened in the past alone, whereas, spatial dependence implies that events at any one point in time can be influenced by both the past and the future (Anselin and Bera, 1998).

Landagan and Barrios (2007) proposed a spatial-temporal model that accommodates irregularly shaped spatial units, with temporal observations made at equal intervals of time shown in Equation 1.

$$Y_{it} = X_{it}\beta + W_{it}\gamma + \varepsilon_{it}, \quad i = 1, \dots, n \quad t = 1, 2, \dots, T \quad (1)$$

where Y_{it} is the response variable from locations i at time t , X_{it} is the set of covariates from location i and at time t , w_{it} is the set of variables in the neighborhood system of location i at time t , and ε_{it} are the error components. The error component was examined for dependence structures, but only the autoregressive behavior was explored as $\varepsilon_{it} = \mu_i + v_{it}$, where $\mu_i \sim IID(0, \sigma_\mu^2)$ and the remainder disturbances v_{it} following stationary $AR(p)$. An estimation procedure that imbeds the Cochrane-Orcutt procedure into the backfitting algorithm of Hastie and Tibshirani (1990) for additive models was modified to simultaneously estimate group parameters at some point of the iterative process. Barrios and Landagan (2007) assumed: (i) constant covariate (β) effect across locations and time, (ii) constant temporal effect (ρ) across locations, and (iii) constant spatial effect (γ) across time.

Guarte and Barrios (2013) proposed a nonparametric bootstrap inference procedure based on the Landagan and Barrios (2007) spatial-temporal model to verify assumptions (ii) and (iii). On the other hand, Martines (2008) proposed a procedure to estimate parameters of multivariate spatial-temporal model by imbedding a multivariate regression and vector autoregressive (VAR) model in backfitting algorithm. The multivariate spatial model at a minimum, subsumed an autoregressive error process adopted from the work of Singh et. al. (2005) given by the model:

$$Y_i = X_i\beta + u_i, \quad u_i = W_i\rho + v_i \quad (2)$$

where Y_i is the $1 \times r$ vector of observation on the response variable from i^{th} location; X is the vector of covariates from i^{th} location; β is the vector of parameters; and u_i is the vector of error component; W_i is the neighborhood variable; ρ is the corresponding spatial effect on u_i ; and v_i is the remainder disturbance which is distributed with mean zero and constant variance. Martines (2008), in order to capture the temporal effect, examined the error disturbance for dependence structure, but he only explored the VAR behavior and postulated it as $v_t = \gamma v_{t-1} + \eta_t$. The multivariate spatial-temporal model of Martines (2008), assuming additivity is given in Equation 3.

$$Y_t = X_t\beta + W_t\rho + v_{t-1}\gamma + \eta_t \quad (3)$$

Following Landagan and Barrios (2007), Martines (2008) assumed constant temporal effect across locations, and constant spatial effect across time periods in model (3).

This study focuses on multivariate spatial-temporal model and proposes a nonparametric bootstrap test in verifying the assumptions of model (3) by extending the work of Guarte and Barrios (2013) to the case of multivariate spatial-temporal model. Recognizing the theoretical importance of determinants, in this paper, the determinant estimates of the bivariate VAR(1) coefficients matrix of either multivariate characteristics across spatial locations/time points were obtained for the purpose of representing and preserving the parameter estimates and in characterizing the behavior of a multivariate spatial-temporal model.

2. Methodology

Given the spatial-temporal VAR model:

$$\tilde{y}_i(t) = \phi_{1i}\tilde{y}_i(t-1) + \phi_{2i}\tilde{y}_i(t-2) + \dots + \phi_{pi}\tilde{y}_i(t-p) + \varepsilon_i(t), \quad (4)$$

in the same context as Martines (2008) assume (i) constant multivariate characteristics across spatial locations and (ii) constant multivariate characteristics across time points.

To verify these assumptions, the bivariate VAR(1) series characterized by the parameter $\phi_{li} = \begin{bmatrix} \phi_{11}^i & \phi_{12}^i \\ \phi_{21}^i & \phi_{22}^i \end{bmatrix}$ with innovations from normal distribution with mean 0 and variance $\sum_{\varepsilon_i(t)}$ were used. The bivariate VAR(1) series are characterized by,

$$\tilde{y}_i(t) = \phi_{li}\tilde{y}_i(t-1) + \varepsilon_i(t) \quad (5)$$

Considering the spatial units, the bivariate VAR(1) series are available for N locations each with T time points. The following hypotheses are tested:

H_0 : $\phi_{11} = \phi_{12} = \dots = \phi_{1N}$, i.e., all spatial locations have the same multivariate characteristics over time.

H_1 : $\phi_{li} \neq \phi_{lj}$ for at least one pair of $i \neq j$ i.e., at least one spatial location differs multivariate characteristics over time.

Algorithm 1:

Given these N time series each with 2 dimensions, the following procedures are used in testing the multivariate characteristics across spatial locations:

1. For each location i estimate VAR(1) process on a bivariate specification and conditioning on $\tilde{y}_i(t-1)$, the empirical distribution of the centered residuals is,

$$\varepsilon_i(t) \sim WN(0, MSE). \quad (6)$$

- Using the residuals (6), generate k bootstrap samples for each spatial location i of sample size $n(\varepsilon_i(1), \varepsilon_i(2), \dots, \varepsilon_i(n))$.
- For every bootstrap sample in Step 2, generate k time series for every i^{th} spatial location using the estimated model in Step 1.
- Estimate the bivariate VAR(1) model for every simulated time series in Step 3, then take the determinant of these bivariate VAR(1) coefficients matrix. Thus, there will be k determinants for k bootstrap samples.
- Compute the standard error of the determinant of the coefficients $\left| \widehat{\phi}_{li} \right|$ using the corresponding k bootstrap determinants $\left| \widehat{\phi}_{lij}^* \right|$ of the multivariate characteristics parameter estimates $\widehat{\phi}_{lij}^*$ generated in Step 4. Then we have,

$$\hat{\sigma}_{\left| \widehat{\phi}_{li} \right|} = \left[\frac{1}{k-1} \sum_{j=1}^k \left(\left| \widehat{\phi}_{lij}^* \right| - \left| \widehat{\phi}_{lij}^* \right| \right)^2 \right] \quad \left| \widehat{\phi}_{lij}^* \right| = \frac{1}{k} \sum_{j=1}^k \left| \widehat{\phi}_{lij}^* \right|, \quad (7)$$

where $j = 1, 2, \dots, k$ (bootstrap samples) and $*$ = bootstrap estimates,

$\left| \widehat{\phi}_{lij}^* \right|$ - determinant of the j bootstrapped estimates, in this case the estimated 2×2 square matrix of the multivariate characteristics across spatial locations, for $j = 1, 2, \dots, k$ bootstraps.

$\left| \widehat{\phi}_{lij}^* \right|$ - mean of the determinant of the j bootstrapped estimates, in this case the estimated 2×2 square matrix of the multivariate characteristics across spatial locations, for $j = 1, 2, \dots, k$ bootstraps.

- Construct $(1 - \alpha) \times 100\%$ bootstrap confidence interval on each determinant of multivariate characteristics $\left| \widehat{\phi}_{li} \right|$ estimated in Step 1:

$$\left| \widehat{\phi}_{li} \right| \mp y_{\alpha} \frac{\hat{\sigma}_{\left| \widehat{\phi}_{li} \right|}}{2} \quad (8)$$

where $\left| \widehat{\phi}_{li} \right|$ is the determinant of the original point estimate from Step 1, and $\hat{\sigma}_{\left| \widehat{\phi}_{li} \right|}$ is its standard error estimate from Step 5.

- Compute the mean/median of the determinant of the multivariate characteristic parameter estimate $\left| \widehat{\phi}_{li} \right|$ in Step 1.
- Reject the null hypothesis that there is constant multivariate characteristic across spatial locations with $(1-\alpha) \times 100\%$ coverage probability if more

than $\alpha\%$ of the constructed intervals fail to contain the mean/median value computed in Step 7.

Using VAR(1) process of the bivariate specification, the data were transformed into cross-sectional for testing the constant multivariate characteristic across time points. The multivariate linear regression model was assumed to be the appropriate model for the cross-sectional data. By extending the univariate model of Guarte and Barrios (2013), the values of $y_i(t)$ were first translated in location and

$$x_{it} = \left(\frac{y'_i(t) - \phi_{1i} y'_i(t-1)}{\beta_{1t}} \right). \quad (9)$$

The above equation was derived from the following relationships:

$$\begin{aligned} y_i(t) &= \phi_{1i} y_i(t-1) + \varepsilon_i(t), \quad \varepsilon_i(t) \sim N(0, I_2) \\ y'_i(t) &= y_i(t) L_2 \sigma + u_{it}, \quad u_{it} = [100 \ 100], \quad i = 1, 2, \dots, N, \quad t = 2, 3, \dots, T, \quad \sigma = 10 \\ y'_i(t) &= \phi_{1i} (y_i(t-1) L_2 \sigma + u_{it}) + \varepsilon_i(t) \\ y'_i(t) &= \phi_{1i} y'_i(t-1) + x_{it} \beta_{1t} \end{aligned}$$

Now, using the covariates (9), the response variable was simulated using the regression model below:

$$y_{it} = x_{it} \beta_{1t} + \varepsilon_{it}, \quad \varepsilon_{it} \sim N(0, I_2 \sigma^2), \quad \begin{matrix} i = 1, 2, \dots, N \\ t = 2, 3, \dots, T \end{matrix}, \quad (10)$$

For this study, $\beta_{1t} = \begin{bmatrix} \beta_{11}^t & \beta_{12}^t \\ \beta_{21}^t & \beta_{22}^t \end{bmatrix}$ for both the response y_{it} and the predictor

x_{it} . Given the series, we test the following hypotheses:

H_0 : $\beta_{11} = \beta_{12} = \dots = \beta_{1T}$, i.e., multivariate regression coefficients have similar characteristics over time, across spatial locations.

H_1 : $\beta_{1i} \neq \beta_{1j}$ for at least one pair of $i \neq j$ i.e., the regression coefficient vary in at least one pair of over time points across spatial locations.

Algorithm 2:

Model (10) is used in the following steps for test constant multivariate characteristics across time points:

1. Estimate the coefficient of model (10), and take the determinant of this, $\left| \widehat{\beta}_{1t} \right|$, for each time point.
2. Generate k bootstrap samples from the pairs of the dependent and independent variables for each time points,

$$b_j = \begin{pmatrix} \mathbf{x}_{it} \\ \mathbf{y}_{it} \end{pmatrix}, \quad (11)$$

$i = j$, implies $k = N$ bootstrap samples. The regression bootstrap procedure above is called case resampling.

3. Estimate the coefficients of model (10) again, and take the determinant, $\left| \widehat{\beta}_{1tj}^* \right|$, using the bootstrap samples obtained in the preceding step for each time point.
4. Compute the standard error of the estimated coefficients $\widehat{\beta}_{1t}$ from Step 1 using the estimated bootstrap determinants, $\left| \widehat{\beta}_{1tj}^* \right|$ from Step 3. Then,

$$\hat{\sigma}_{\left| \widehat{\beta}_{1t} \right|} = \left[\frac{1}{k-1} \sum_{j=1}^k \left(\left| \widehat{\beta}_{1tj}^* \right| - \overline{\left| \widehat{\beta}_{1tj}^* \right|} \right)^2 \right]^{1/2} \quad \overline{\left| \widehat{\beta}_{1tj}^* \right|} = \frac{1}{k} \sum_{j=1}^k \left| \widehat{\beta}_{1tj}^* \right|, \quad (12)$$

where $j = 1, 2, \dots, k$ (bootstrap samples) and $*$ = bootstrap estimates,

$\left| \widehat{\beta}_{1tj}^* \right|$ - determinant of the j bootstrapped estimates, in this case the estimated 2×2 square matrix of the multivariate characteristics across time points, for $j = 1, 2, \dots, k$ bootstraps.

$\overline{\left| \widehat{\beta}_{1tj}^* \right|}$ - mean of the determinant of the j bootstrapped estimates, in this case the estimated 2×2 square matrix of the multivariate characteristics across time points, for $j = 1, 2, \dots, k$ bootstraps.

5. Construct $(1 - \alpha) \times 100\%$ bootstrap confidence interval on each determinant of multivariate characteristics $\left| \widehat{\beta}_{1t} \right|$ estimated in Step 1:

$$\left| \widehat{\beta}_{1t} \right| \mp y_{\alpha/2} \hat{\sigma}_{\left| \widehat{\beta}_{1t} \right|}, \quad (13)$$

where $\left| \widehat{\beta}_{1t} \right|$ is the determinant of the original point estimate from Step 1, and $\hat{\sigma}_{\left| \widehat{\beta}_{1t} \right|}$ is its standard error estimate from Step 5.

6. Compute the mean/median of the determinant of the multivariate characteristic parameter estimate $\left[\widehat{\beta}_{lt}\right]$ in Step 1.
7. Reject the null hypothesis that there is constant multivariate characteristic across time points with $(1-\alpha)\times 100\%$ coverage probability if more than $\alpha\%$ of the constructed intervals fail to contain the mean/median value computed in Step 6.

Algorithm 3:

Power of the test is computed after Step 7 of algorithm 1 and after Step 6 of algorithm 2. The lower and upper limits of the 95% confidence interval were used to compute their corresponding y-scores. That is, for a given confidence interval of the $\left[\widehat{\phi}_{ij}^*\right]$, (LL,UL). The y-score of the LL (lower limit) is,

$$y_{LL} = \frac{LL - \left[\widehat{\phi}_{lt}\right]}{\hat{\sigma}_{\left[\widehat{\phi}_{lt}\right]}} \text{ and/or } y_{LL} = \frac{LL - \left[\widetilde{\phi}_{lt}\right]}{\hat{\sigma}_{\left[\widetilde{\phi}_{lt}\right]}}, \quad (14)$$

$$y_{LL} = \frac{LL - \left[\widehat{\beta}_{lt}\right]}{\hat{\sigma}_{\left[\widehat{\beta}_{lt}\right]}} \text{ and/or } y_{LL} = \frac{LL - \left[\widetilde{\beta}_{lt}\right]}{\hat{\sigma}_{\left[\widetilde{\beta}_{lt}\right]}}, \quad (15)$$

where $\left[\widehat{\phi}_{lt}\right]$ and $\left[\widetilde{\phi}_{lt}\right]$ are the mean and the median value, respectively, of the determinant of the coefficients from the bootstrap samples for multivariate characteristic across spatial locations, and $\left[\widehat{\beta}_{lt}\right]$ and $\left[\widetilde{\beta}_{lt}\right]$ are the mean and the median values, respectively, of the determinant of the coefficients from the bootstrap samples for multivariate characteristics across time points. And the y-scores for the UL (upper limit),

$$y_{UL} = \frac{UL - \left[\widehat{\phi}_{lt}\right]}{\hat{\sigma}_{\left[\widehat{\phi}_{lt}\right]}} \text{ and/or } y_{UL} = \frac{UL - \left[\widetilde{\phi}_{lt}\right]}{\hat{\sigma}_{\left[\widetilde{\phi}_{lt}\right]}}, \quad (16)$$

$$y_{UL} = \frac{UL - \left[\widehat{\beta}_{lt}\right]}{\hat{\sigma}_{\left[\widehat{\beta}_{lt}\right]}} \text{ and/or } y_{UL} = \frac{UL - \left[\widetilde{\beta}_{lt}\right]}{\hat{\sigma}_{\left[\widetilde{\beta}_{lt}\right]}}, \quad (17)$$

Then β is,

$$\beta = P(y_{UL} \leq X \leq y_{LL}). \quad (18)$$

And thus the power of the test is,

$$\begin{aligned}\text{Power} &= 1 - P(y_{UL} \leq X \leq y_{LL}) \\ &= 1 - [P(y \leq y_{UL}) - P(y \leq y_{LL})]\end{aligned}\quad (19)$$

3. Simulation Studies

Four dataset were simulated, of which, two are from balanced datasets $N=T=100$ and $N=T=20$ and the other two sets are from unbalanced datasets $N=70$, $T=50$ and $N=40$, $T=35$.

To evaluate the test of constant multivariate characteristics across spatial locations, consider the stationary VAR(1) with simplest bivariate specification model given below:

$$y_t = \begin{pmatrix} 0.7 & 0.2 \\ 0.2 & 0.7 \end{pmatrix} y_{t-1} + \varepsilon_t \quad \text{with } \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}. \quad (20)$$

Algorithms 1 and 3 were utilized for the needed results. The simulation process has no intercept term for the model.

The same algorithms were applied also to the case of nonconstant multivariate characteristics across spatial locations. Four datasets were generated using (20) model with the first five spatial locations having the same multivariate characteristics for each dataset. In each dataset generated, there were four sets of results and each has either 6 or 7 cases. The first set has a coefficient, Q_{11} that changes with an increment of 0.10 while holding all other coefficients unchanged, the second set, with a coefficient, Q_{12} that changes with also the same increment of 0.10 while all other coefficients were held fixed. The third set with coefficient Q_{21} likewise changes with an increment of 0.10 and the last, with coefficient of Q_{22} also changes with similar increment of 0.10.

The same stationary bivariate VAR(1) mentioned above was used to evaluate the test of constant multivariate characteristics across time points. The datasets were then transformed into cross-sectional and the multivariate linear regression model was assumed to be the appropriate model for the cross-sectional data. Covariates were computed from equation (5) using $\beta_1 = \begin{pmatrix} 0.7 & 0.2 \\ 0.2 & 0.7 \end{pmatrix}$. The response variable was then computed using these covariates and parameter β_j . Algorithms 2 and 3 were applied to obtain the necessary results.

In the case of the nonconstant multivariate characteristics across time points, four datasets were generated with the first 5% time points having the same multivariate characteristics. In each dataset generated, there were four sets of results and each has either 6 or 7 cases. The first set has a coefficient, β_{11} that changes with an increment of 0.10. The second set, with a coefficient, β_{12} changes with also the same increment of 0.10. The third set with coefficient β_{21} changes

with an increment of 0.10, and the last, with coefficient of β_{22} also changes with similar increment of 0.10.

4. Results

From Table 1, of the 100 bootstrap normal confidence intervals that were constructed, only four (4) failed to contain both the mean and median. The four (4) spatial locations that significantly differed from the rest were the third highest multivariate characteristics estimate (0.584), the highest estimate (0.615), the lowest estimate (0.224), and the second highest estimate (0.599). This could mean that even if the population of spatial locations was homogeneous with respect to the multivariate characteristics, sampling variation will induce some spatial locations to be different in the sample with respect to multivariate characteristics. We can conclude that the testing procedure was able to correctly identify the true situation and is properly sized for this dataset. Not rejecting the null hypothesis of constant multivariate characteristics across spatial locations with 95% coverage probability actually captured the fact that not more than five (5) spatial locations differ in multivariate characteristics. There was a correct inference for this large balanced data, $(N, T) = (100, 100)$.

Table 1. Multivariate Characteristics Across Spatial Locations for Stationary Bivariate VAR(1) Model for Large and Small Balanced/Unbalanced Data and 95% Coverage Probability

Sample Size	Criterion	Value	95% Coverage Probability
(N,T) = (100,100)	Mean	0.414	Do not reject H_0 (4)
	Median	0.407	Do not reject H_0 (4)
(N,T) = (70,50)	Mean	0.377	Do not reject H_0 (2)
	Median	0.386	Do not reject H_0 (3)
(N,T) = (40,35)	Mean	0.344	Do not reject H_0 (1)
	Median	0.340	Do not reject H_0 (1)
(N,T) = (20,20)	Mean	0.320	Do not reject H_0 (0)
	Median	0.348	Do not reject H_0 (0)

Note: Figures in parentheses are the number of normal bootstrap confidence intervals that failed to contain the mean/or median.

For large unbalanced dataset, $N=70$, $T=50$, the null hypothesis of constant multivariate characteristics across 70 spatial locations based on both the mean and/or median is not rejected with 95% coverage probability using the bootstrap normal confidence intervals. Of the 70 bootstrap confidence intervals constructed, only two (2) failed to contain the mean and three (3) failed to contain the median. The three (3) spatial locations that significantly differed from the sixty seven (67) based on the median were the highest estimate (0.591), the second lowest estimate

(0.169) and the lowest estimate (0.093). Again, due to sampling variation, this leads to a few spatial locations coming out different with respect to multivariate characteristics in the said sample. We may therefore conclude that the test procedure was able to correctly identify the true situation and is properly sized for this dataset using bootstrap normal confidence intervals. This means not rejecting the null hypothesis of constant multivariate characteristics across spatial locations with 95% coverage probability actually captured the fact that not more than four (4) spatial locations differ in multivariate characteristics effect.

For small unbalanced dataset, $N=40$, $T=35$, the null hypothesis of constant multivariate characteristics effect across spatial locations was not rejected with 95% coverage probability. Of the 40 bootstrap confidence intervals constructed, only one (1) failed to contain both the mean and median estimates. Thus, we can conclude that the test procedure was able to correctly identify the true situation and is properly sized for this dataset. That is, not rejecting the null hypothesis of constant multivariate characteristics across spatial locations with 95% coverage probability actually captured the fact that not more than two (2) spatial locations differ in multivariate characteristics.

For small balanced dataset, $N=20$, $T=20$, the null hypothesis of constant multivariate characteristics across spatial locations was not rejected with 95% coverage probability. Of the 20 confidence intervals constructed, none failed to contain both the mean and median estimates using the bootstrap normal confidence intervals. We may therefore conclude that the testing procedure was able to correctly identify the true situation and is properly sized for this dataset. That is, not rejecting the null hypothesis of constant multivariate characteristics across spatial locations with 95% coverage probability actually captured the fact that not more than one (1) spatial location differ in multivariate characteristics, which in this case, there was none.

The power plots in Figure 1, show the estimated maximum power with 5% different multivariate characteristics across spatial locations for large balanced dataset. The horizontal axis represents the parameter value (Q_{ii}) with an increment of 0.10 and the vertical axis represents the estimated maximum power based on either the mean (denoted by black line) or the median (denoted by red line). The estimated maximum power of the test procedure is high on the diagonal of VAR(1) coefficient matrix (set 1_1 and set 1_4) compared to that of the off-diagonal (set 1_2) and set 1_3). In general, the estimated maximum power is highest for the farthest alternative parameter and less than 0.80 when closer to the true parameter values. The power plots for large unbalanced dataset are shown in Figure 2. As noticed, the power of the test procedure is good (high) in all four sets.

For both large balanced/unbalanced data, it is observed that the estimated maximum power based on the median is slightly higher than based on the mean and the fluctuation of the estimated power plots based on the mean and median are almost identical.

For small unbalanced data, the estimated power plots are shown in Figure 3. Based on that, the maximum power of the test based on the median estimates is manifested clearly higher than that based on the mean estimates. The pattern of the fluctuation of the power plots based on mean and median are almost similar. However, for small balanced dataset, the maximum power of the test procedure is low (below 0.80) in almost all cases (Refer to Figure 4). The pattern of the fluctuation of the estimated power plots based on mean and median are erratic.

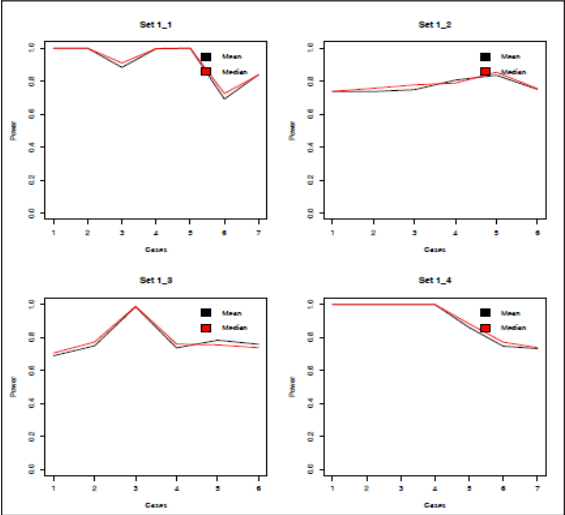


Figure 1. Power Curve with 5% Different Multivariate Characteristics Across Spatial Locations for Large Balanced Dataset

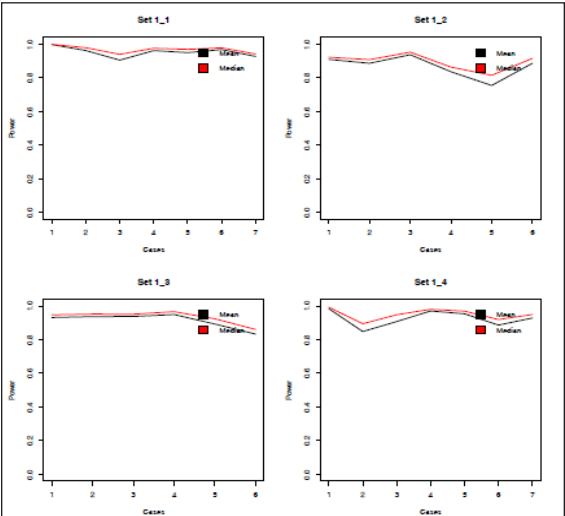


Figure 2. Power Curve with 5% Different Multivariate Characteristics Across Spatial Locations for Large Unbalanced Dataset

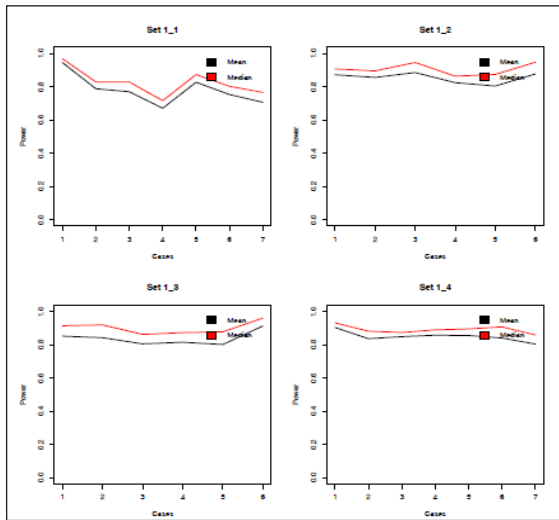


Figure 3. Power Curve with 5% Different Multivariate Characteristics Across Spatial Locations for Small Unbalanced Dataset

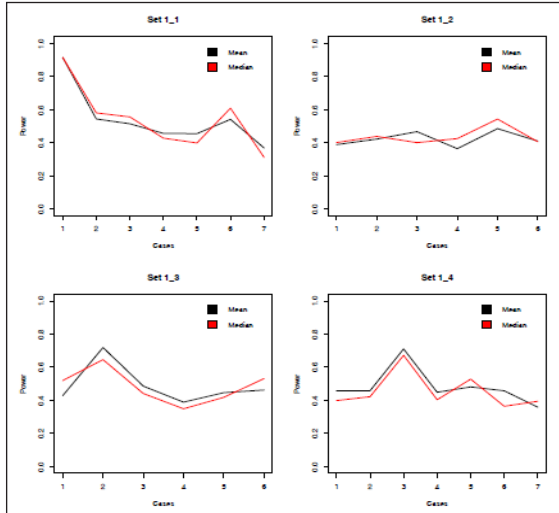


Figure 4. Power Curve with 5% Different Multivariate Characteristics Across Spatial Locations for Small Balanced Dataset

Table 2 presents the results of testing multivariate characteristics across time points. For large balanced data, of the 100 confidence intervals constructed, five (5) failed to contain both the mean and median estimates. For large unbalanced data, of the 50 confidence intervals constructed, one (1) failed to contain the mean estimate and two (2) confidence intervals failed to contain the median estimate. The null hypothesis of constant multivariate characteristics across time points was correctly not rejected with 95% coverage probability for large balanced and/or unbalanced dataset. On the other hand, for small unbalanced dataset set, of the 35 confidence intervals constructed, two (2) failed to contain the mean estimate and only one (1) interval failed to contain the median estimate. In the case of small balanced dataset, of the 20 confidence intervals constructed, only one (1) confidence interval failed to contain both the mean and median estimates. This can be interpreted that the null hypothesis of constant multivariate characteristics across time point was not also rejected with 95% coverage probability.

Table 2. Multivariate Characteristics Across Time Points for Stationary Bivariate VAR(1) Model for Large and Small Balanced/ Unbalanced Data and 95% Coverage Probability

Sample Size	Criterion	Value	95% Coverage Probability
(N,T) = (100,100)	Mean	0.442	Do not reject H_0 (5)
	Median	0.441	Do not reject H_0 (5)
(N,T) = (70,50)	Mean	0.429	Do not reject H_0 (1)
	Median	0.415	Do not reject H_0 (2)
(N,T) = (40,35)	Mean	0.445	Do not reject H_0 (2)
	Median	0.439	Do not reject H_0 (1)
(N,T) = (20,20)	Mean	0.463	Do not reject H_0 (1)
	Median	0.447	Do not reject H_0 (1)

Note: Figures in parentheses are the number of normal bootstrap confidence intervals that failed to contain the mean/or median.

The power plots in Figures 5 to 8, show the estimated maximum power with 5% different multivariate characteristics across time points. The estimated maximum power was high in most of the cases in each set of large balanced/unbalanced data, indicating that the proposed procedure was powerful (see Figures 5 and 6). The pattern of the fluctuation of the estimated power based on mean and median estimates are almost identical with the median estimates which are slightly higher than the mean estimates. However, for small unbalanced/ balanced data, the maximum power of the test procedure was observed to be low in the majority of the cases (refer to Figures 7 and 8).

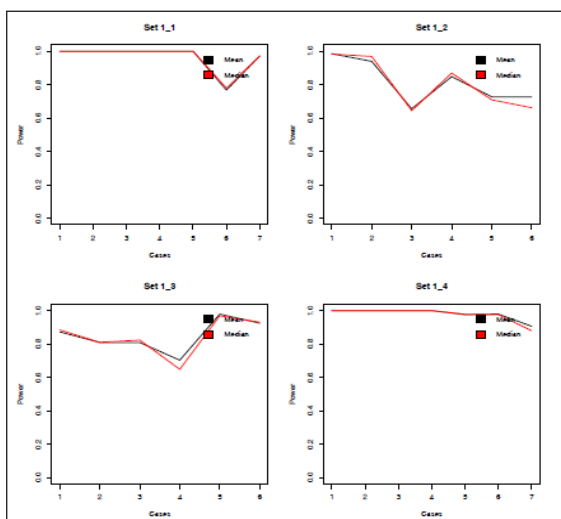


Figure 5. Power Curve with 5% Different Multivariate Characteristics Across Time Points for Large Balanced Dataset

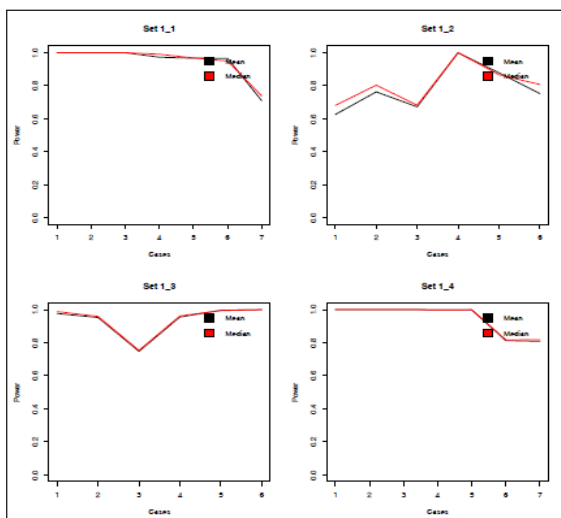


Figure 6. Power Curve with 5% Different Multivariate Characteristics Across Time Points for Large Unbalanced Dataset

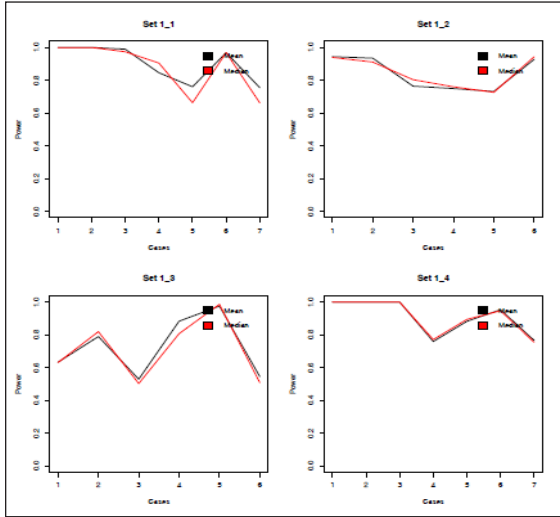


Figure 7. Power Curve with 5% Different Multivariate Characteristics Across Time Points for Small Unbalanced Dataset

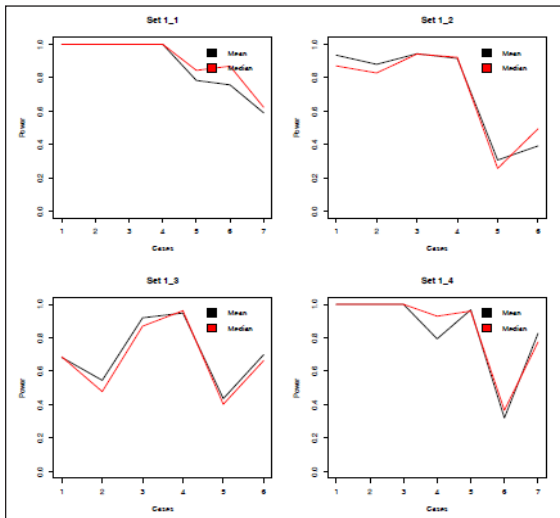


Figure 8. Power Curve with 5% Different Multivariate Characteristics Across Time Points for Small Balanced Dataset

5. Conclusions

The simplest VAR of order 1 with bivariate specification modelling effort is a good foundation for the development of multivariate spatial-temporal models. We have demonstrated the effectiveness of our proposed test procedures on simulated datasets that accurately revealed the true situation. We have also demonstrated that including both spatial and temporal in multivariate model, although difficult, is feasible. We can therefore conclude that the test procedures were able to correctly identify the true situation and are properly sized for large balanced/unbalanced data and are powerful, and so, the tests are robust. However, for small balanced/unbalanced data, the tests were not robust, because of their being less powerful.

References

- ANSELIN, L., BERA, A., 1998, *Spatial Dependence in Linear Regression Models with an Introduction to Spatial Economics*, New York: Marcel Dekker.
- CHERNICK, M., 1999, *Bootstrap Methods: A Practitioner's Guide*. New York: John Wiley & Sons, Inc.
- EFRON, B. and TIBSHIRANI, R., 1993, *An Introduction to the Bootstrap*, New York: Chapman and Hall/CRC.
- GUARTE, J. and BARRIOS, E., 2013, Nonparametric hypothesis testing in a spatial-temporal model: A simulation study, *Communications in Statistics-Simulation and Computation* 42 (1): 153-170.
- HASTIE, T. and TIBSHIRANI, R., 1990, *Generalized Additive Models*, New York: Chapman and Hall/CRC.
- LANDAGAN, O. and BARRIOS, E., 2007, An estimation procedure for a spatial-temporal model, *Stat. and Prob. Letters* 77: 401-406.
- MARTINES, A., 2012, Small area estimation with a multivariate spatial-temporal model, *The Philippine Statistician* 61(2): 1-17.
- SINGH, B., SHUKLA, G. and KUNDU, D., 2005, Spatio-temporal models in small area estimation, *Survey Methodology* 31(2): 183-195.
- TOBLER, W., 1979, Cellular geography, in S. Gale and G. Olsson (eds.) *Philosophy in Geography*, Dordrech: Reidel, pp. 379-38.

Some Zero Inflated Poisson-Based Combined Exponentially Weighted Moving Average Control Charts for Disease Surveillance

Robert Neil F. Leong

Mathematics Department, De La Salle University

Frumencio F. Co

Mathematics Department, De La Salle University

Daniel Stanley Y. Tan

Software Technology Department, De La Salle University

One of the main areas of public health surveillance is infectious disease surveillance. With infectious disease backgrounds usually being more complex, appropriate surveillance schemes must be in order. One such procedure is through the use of control charts. However, with most background processes following a zero-inflated Poisson (ZIP) distribution as brought about by the extra variability due to excess zeros, the control charting procedures must be properly developed to address this issue. Hence in this paper, drawing inspiration from the development of combined control charting procedures for simultaneously monitoring each ZIP parameter individually in the context of statistical process control (SPC), several combined exponentially weighted moving average (EWMA) control charting procedures were proposed (Bernoulli-ZIP and CRL-ZTP EWMA charts). Through an extensive simulation study involving multiple parameter settings and outbreak model considerations (i.e., different shapes, magnitude, and duration), some key results were observed. These include the applicability of performing combined control charting procedures for disease surveillance with a ZIP background using EWMA techniques. For demonstration purposes, application with an actual data, using confirmed measles cases in the National Capital Region (NCR) from January 1, 2010 to January 14 2015, revealed the comparability of the Bernoulli-ZIP EWMA scheme to historical limits method currently in use.

*Keywords: EWMA control charts, disease surveillance,
ZIP distribution, measles*

1. Introduction

Control charts have long been in use in healthcare and public health surveillance. One specific application is on the prospective surveillance of diseases with the aim of early outbreak detections (Unkel et al., 2012; Woodall, 2006). In particular, the earliest temporal outbreak detection algorithms are variants of several control charting procedures. Also, while the health surveillance literature has now advanced to include new methodologies as motivated by regression and time series techniques, control charts are still widely used in most general settings because of their simplicity and ease-of-implementation (Fricker, 2013; Unkel et al., 2012).

However, despite the long history of applications of control charting procedures in public health surveillance, only a few were developed for infectious disease surveillance (Woodall, 2006). The complexity of developing surveillance procedures for infectious diseases is because of the problems associated with on-line outbreak detection: reporting delays, background incidence characterization, disease dynamics, etc. among others (Farrington and Andrews, 2004; Shmueli and Burkom, 2010). Nevertheless, the development of prospective surveillance procedures directly inspired by control charting procedures have garnered interest because of the parallelism between the problems involved – detecting outbreaks temporally is similar to that of detecting out-of-control shifts in industrial manufacturing processes. Particularly, one of the most common variants of statistical process control (SPC) procedures studied for disease surveillance are the Exponentially Weighted Moving Average (EWMA) control charts. EWMA control charts, first studied by Dr. S. W. Roberts in 1959, have gained substantial interest due to their simplicity and their ability to balance the optimal properties and the shortcomings of both Shewhart and Cumulative Sum (CUSUM) charts.

Applications of EWMA charts in prospective disease surveillance have first been loosely used in the development of the numerical derivative-based method by Nobre and Stroup (1994) for monitoring measles in the United States. Extensive studies by Burkom et al. (2007) and Elbert and Burkom (2009) proposed the use of the Holt-Winters technique for generalized exponential smoothing to account for trends and seasonal features. Moreover, Dong et al. (2008) proposed three types of EWMA charts that do not require an assumption of identical distributions of counts to detect positive shifts in incidence rates.

Empirical evaluations of EWMA charts have revealed their comparability with CUSUM variants in terms of sensitivity in detecting small shifts in process mean (Wong and Moore, 2006). Furthermore, Han et al. (2010) showed that for appropriately selected parameters, EWMA charts generally perform well for detecting shifts in Bernoulli observations. Also, Szarka et al. (2011) concluded that EWMA methods using adaptive thresholds generally performed better than most procedures in the Early Aberration Reporting System (EARS) of the Center for Disease Control (CDC). In the context of real-time monitoring of hospital

acquired infections, Morton et al. (2001) concluded that EWMA charts are ideal for monitoring bacteraemia and multi-resistant organism rates.

In spite of a long list of literature in the applications of SPC procedures in on-line disease surveillance, none in the aforementioned list has considered EWMA variants based on a ZIP distribution despite the knowledge that disease counts may contain excess zeros particularly for rare diseases (Ngatchou-Wandji and Paris, 2011). The lack of development of ZIP-based monitoring schemes in infectious disease surveillance may be due to the later developments of ZIP-based control charting procedures. A more recent development in the utilization of ZIP-based chart in disease surveillance is the development of an EWMA chart by Fatahi et al. (2012) to monitor unintentional needle-stick occurrences in Tehran. However, a perceived drawback of such is that it only monitors shifts in the mean process level, as opposed to the knowledge that a ZIP process is characterized by two parameters (π : vulnerability level, and λ : intensity level). Hence, a combination of control charting procedures must be ideal to simultaneously monitor the parameters of a ZIP process individually. The rationale behind simultaneous monitoring is to better identify the nature of the shift in the mean disease incidence level when it occurs. While this problem has never been tackled in any disease surveillance literature, several studies in the context of SPC have considered this before. In particular, as motivated by the Generalized ZIP model of Chen et al. (2008), He et al. (2012) developed the p - λ CUSUM chart, based on the combination of individual CUSUM charts. He et al. (2014) likewise developed the CRL-ZTP CUSUM, which is a combination of two CUSUM charts based on conforming run length to detect shifts in the occurrence of random shocks and on a zero-truncated Poisson distribution to monitor the magnitude of the occurrences of the random shocks.

Inspired by these developments of combined control charting procedures, this paper aims to showcase the usage of combination of control charting procedures to simultaneously detect increases in both parameters of a ZIP process in the context of disease surveillance. In particular, the paper aims to develop a combination of EWMA control charting procedures based on existing individual EWMA charts for a Bernoulli parameter and a Poisson parameter for simultaneously monitoring parameters of a ZIP process. It is desired that the results of this study will contribute to the literature of using control charts in performing disease surveillance, with an aim of developing improved early warning systems for outbreaks.

The next section provides a short discussion of the ZIP distribution followed by the development of two combined EWMA control charting procedures. A simulation study follows in the next section to provide empirical performance evaluations of the proposed charts in the context of disease surveillance. The fourth section then presents a case application of the proposed procedures with actual measles data in comparison with the historical limits procedure which is

currently in use. Lastly, conclusions and recommendations for future works are provided.

2. Conceptual Framework

2.1. Zero-inflated Poisson distribution

A random variable Y is defined to have a zero-inflated Poisson distribution if its probability mass function is given by

$$f(y) = \begin{cases} 1 - \pi + \pi e^{-\lambda}, & y = 0 \\ \pi \frac{e^{-\lambda} \lambda^y}{y!}, & y = 1, 2, 3, \dots \end{cases} \quad (1)$$

where $0 \leq \pi \leq 1$ is the probability that a random shock is observed (where in the context of disease surveillance will be known as the vulnerability level) and $\lambda > 0$ is the Poisson intensity of the shocks as they occur (where in the same context will be known as the intensity level). Clearly, $1 - \pi$ is the excess proportion of zeros causing the extra poisson variation (i.e., overdispersion). Some of its moments are as follows:

$$E(Y) = \lambda\pi \quad (2)$$

$$Var(Y) = \lambda\pi (\lambda + 1 - \lambda\pi). \quad (3)$$

It can be shown that the maximum likelihood estimator of π is given by

$$\hat{\pi}_{mle} = \frac{\sum_{i=1}^n (1 - I_0(y_i)) / n}{1 - e^{-\hat{\lambda}_{mle}}}, \quad (4)$$

where $I_0(y_i)$ is the indicator function for $y_i = 0$. Likewise, it can be shown that the maximum likelihood estimator of λ is the solution to the nonlinear equation

$$\frac{\hat{\lambda}_{mle}}{1 - e^{-\hat{\lambda}_{mle}}} = \frac{\sum_{i=1}^n y_i (1 - I_0(y_i))}{\sum_{i=1}^n (1 - I_0(y_i))}. \quad (5)$$

Note that $\sum_{i=1}^n y_i (1 - I_0(y_i))$ is the sum of all nonzero observations and $\sum_{i=1}^n (1 - I_0(y_i))$ is the number of nonzero observations in the sample.

2.2 Bernoulli-ZIP EWMA chart

An initial attempt in using EWMA control charting procedure for ZIP-distributed data was developed by Fatahi et al. (2012). The control chart was called ZIP-EWMA control chart, specifically developed to monitor rare health-related events. The chart monitors the statistic

$$E_t = \max[0, (1 - \kappa) E_{t-1} + \kappa Y_t], \quad t = 1, 2, 3, \dots, \quad (6)$$

where Y_t is the observed number of cases at time t , $0 \leq \kappa \leq 1$, being the weighting parameter, and $E_0 = E(Y) = \lambda\pi$ as an initialization procedure. The ZIP-EWMA chart signals whenever $E_t > h_{ZIP-EWMA}$, where $h_{ZIP-EWMA}$ is the steady-state threshold taken as

$$h_{ZIP-EWMA} = \lambda\pi + L_{ZIP-EWMA} \sqrt{\left(\frac{\kappa}{2 - \kappa}\right) [\lambda\pi(\lambda + 1 - \lambda\pi)]}, \quad (7)$$

and $L_{ZIP-EWMA}$ is appropriately chosen to achieve the desired average false signal rate.

However, the ZIP-EWMA control charting procedure was developed only to monitor shifts in λ . Thus, to augment its ability to detect shifts in π , it is proposed to be utilized simultaneously with an EWMA chart for Bernoulli data. In this case, the EWMA control chart to be used for Bernoulli data will be called Bernoulli-EWMA control chart, wherein the Bernoulli data are generated through the indicator function $I_t (y \geq 1)$ which takes on the value 1 whenever the observed number of cases y_t is at least one. Note that some of the moments of $I_t (y \geq 1)$ are as follows:

$$E(I_t (y \geq 1)) = \pi(1 - e^{-\lambda}), \quad (8)$$

$$Var(I_t (y \geq 1)) = \pi(1 - e^{-\lambda})[1 - \pi(1 - e^{-\lambda})]. \quad (9)$$

Thus, the Bernoulli-EWMA chart is designed to monitor the statistic

$$F_t = \max [0, (1 - \kappa) F_{t-1} + \kappa I_t (y \geq 1)], \quad t = 1, 2, 3, \dots \quad (10)$$

whereas an initialization procedure, $F_0 = E(I_t (y \geq 1)) = \pi(1 - e^{-\lambda})$, and the Bernoulli-EWMA chart signals whenever $F_t > h_{Ber-EWMA}$, where $h_{Ber-EWMA}$ is the steady-state threshold taken as

$$h_{Ber-EWMA} = \pi(1 - e^{-\lambda}) + L_{Ber-EWMA} \sqrt{\left(\frac{\kappa}{2 - \kappa}\right) \left\{ \pi(1 - e^{-\lambda}) [1 - \pi(1 - e^{-\lambda})] \right\}}, \quad (11)$$

and $h_{Ber-EWMA}$ is appropriately chosen to achieve the desired average false signal rate.

In this paper, a combined EWMA control charting scheme is proposed to simultaneously monitor shifts in λ (based on the ZIP-EWMA chart) and in π (based on the Bernoulli-EWMA chart). This control charting procedure will be called as the Bernoulli-ZIP EWMA chart, and is made to generate a signal whenever one of the individual EWMA chart signals.

2.3 CRL-ZTP EWMA Chart

As an alternative EWMA procedure to monitor upward shifts in λ , a zero-truncated Poisson (ZTP)-based procedure is proposed following that of He et al. (2014) which viewed the ZIP process as a combination of Bernoulli and ZTP processes. That is, for the proposed ZTP-EWMA control chart, the procedure monitors the statistic

$$W_t = \max[0, (1 - \kappa)W_{t-1} + \kappa Y_t], \quad t = 1, 2, 3, \dots, \quad (12)$$

where Y_t is the observed number of cases at time t assumed to follow a ZTP distribution with parameter λ , and $W_0 = E(Y) = \frac{\lambda}{1 - e^{-\lambda}}$ as an initialization procedure. Here, it should be noted that W_t updates only whenever nonzero cases are observed in time t . The ZTP-EWMA chart signals whenever $W_t > h_{ZTP-EWMA}$, where $h_{ZTP-EWMA}$ is the steady-state threshold taken as

$$h_{ZTP-EWMA} = \frac{\lambda}{1 - e^{-\lambda}} + L_{ZTP-EWMA} \sqrt{\left(\frac{\kappa}{2 - \kappa}\right) \left[\frac{\lambda}{1 - e^{-\lambda}} \left(1 - \frac{\lambda e^{-\lambda}}{1 - e^{-\lambda}}\right) \right]}, \quad (13)$$

and $L_{ZTP-EWMA}$ is appropriately chosen to achieve the desired average false signal rate.

To be able to detect shifts in π , the CRL-EWMA chart is proposed to be simultaneously utilized. The CRL-EWMA chart is a Bernoulli-type EWMA chart based on geometric conforming run lengths (CRL) as developed by Mavroudis and Nicolas (2013). The CRL-EWMA chart is an extension of the EWMA chart of Yeh et al. (2008) with increased detection sensitivity. That is, only the accumulation of positive deviations from the target are considered and is based on the CRL, which is assumed to follow a geometric distribution with mean $\mu_{CRL} = 1/p_0$, between nonzero cases. The control charting procedure transforms each CRL count into

$$CRL^- = \min [\mu_{CRL}, CRL] = \mu_{CRL} - \max[0, \mu_{CRL} - CRL]. \quad (14)$$

Mavroudis and Nicolas (2013) have shown that some of the moments of CRL^- are as follows:

$$E(CRL^-) = \frac{1 - 2p_0 - (1 - p_0)^{\frac{1+p_0}{p_0}}}{p_0(1 - p_0)}, \quad (15)$$

$$E\left([CRL^-]^2\right) = \frac{p_0^2 - 4p_0 + 2 + (1 - p_0)^{\frac{1}{p_0}}(-p_0^2 + 5p_0 - 4)}{p_0^2(1 - p_0)}, \quad (16)$$

$$Var(CRL^-) = E\left([CRL^-]^2\right) - [E(CRL^-)]^2. \quad (17)$$

Hence, the chart monitors the statistic

$$G_t = \max\left[0, (1 - \kappa)G_{t-1} + \kappa CRL_t^-\right], \quad t = 1, 2, 3, \dots \quad (18)$$

where CRL_t^- is the transformed conforming run length at time t assuming nonzero cases were observed at time t and $G_0 = E(CRL^-)$ as an initialization procedure. Likewise, note that G_t only updates whenever nonzero cases are observed in time t . The CRL-EWMA chart signals whenever $G_t < h_{CRL-EWMA}$, where $h_{CRL-EWMA}$ is the steady-state threshold taken as

$$h_{CRL-EWMA} = E(CRL^-) - L_{ZTP-EWMA} \sqrt{\left(\frac{\kappa}{2 - \kappa}\right) Var(CRL^-)}, \quad (19)$$

and $L_{CRL-EWMA}$ is appropriately chosen to achieve the desired average false signal rate.

Similar to the other combined charting procedure, this control charting procedure, to be called the *CRL-ZTP EWMA chart* henceforth, will signal whenever one of the individual EWMA chart signals.

3. Simulation Study

3.1. Simulation Framework

To empirically compare the proposed combined EWMA charting procedures in the context of disease surveillance, a simulation study was performed using MATLAB routines. The algorithms are available from the authors upon request. The study was done as follows:

1. From a ZIP distribution as characterized by each π - λ parameter combination (i.e., the disease background) from the sets $\pi: \{0.30, 0.60\}$ (i.e., low vulnerability, high vulnerability) and $\lambda: \{1.0, 4.0\}$ (i.e., low intensity, high intensity), 1,000 data sets, each having a length of 1,000 time points, were simulated. The choice of these parameter sets are based on the recommendations by Sim and Lim (2008), which are typically found in those of biological applications. Each simulated data set was subjected to a χ^2 goodness-of-fit test to ensure that the particular data followed the desired distribution.
2. The first 250 observations of each series were used to determine the parameter settings of each of the control charting procedure, i.e., the Phase I procedure. That is, for the EWMA charts, in-control π_0 was estimated by $\hat{\pi}_{mle}$ (see Equation (4)), while in-control λ_0 was estimated by $\hat{\lambda}_{mle}$ (see Equation (5)). Hence, for the Bernoulli-EWMA chart,

$$E\left(I_t(\widehat{y} \geq 1)\right) = \hat{\pi}_{mle} \left(1 - e^{-\hat{\lambda}_{mle}}\right), \quad (20)$$

which is equivalent to the sample proportion of nonzero cases. For the CRL-EWMA control chart, CRLs were computed based on the notion of time between *signal events*, defined by Fraker et al. (2008) as two consecutive time periods wherein the charting procedure generated signals. Thus, p_0 was estimated as

$$\widehat{p}_0 = \left[\hat{\pi}_{mle} (1 - e^{-\hat{\lambda}_{mle}}) \right]^2, \quad (21)$$

i.e., the square of the sample proportion of nonzero cases. For the weighting parameter κ , two settings were considered, namely $\kappa = 0.25$ (i.e., relatively small weight on most recent value) and $\kappa = 0.45$ (i.e., relatively large weight on most recent value). The former is based on the usual setting done in the application of EWMA charts in infection control (Fatahi et al., 2012), while the latter to only serve as a comparison when κ is large.

3. To ensure comparability of the proposed schemes with regard to their abilities to detect parameter shifts, the average time between false signals (ATFS) was preset at 90 for each combined charts. That is, appropriate thresholds for each chart were determined to achieve the desired ATFS values via simulation. For each data set, the charting schemes were applied for the remaining 750 observations as a Phase II procedure. Following Fricker (2013), the estimated ATFS for each data set is computed as

$$\widehat{ATFS} = \frac{750}{\text{number of signals generated}} \quad (22)$$

For such ATFS formula to be applicable for charting schemes with memory, the statistics for such charts were designed to reset whenever a signal was generated. Based on the desired ATFS level, thresholds for each combined charting procedure was adjusted accordingly. The calibrated threshold parameters which achieved such ATFS are summarized in Tables 1 and 2 for the Bernoulli-ZIP EWMA and the CRL-ZTP EWMA charts, respectively. It is worth noting that for a combined mean ATFS of 90, individual mean ATFS values for both π - and λ - designed charts ranged from 200 to 250.

Table 1. Threshold Parameters for Bernoulli-ZIP EWMA Charts for ATFS = 90

ZIP Parameters		Threshold Parameter Values			
		$\kappa = 0.25$		$\kappa = 0.45$	
π_0	λ_0	L_π	L_λ	L_π	L_λ
0.30	1.0	2.8113	3.1366	2.9037	3.7901
0.30	4.0	2.5483	2.7960	2.4317	3.1301
0.60	1.0	2.3548	2.7885	2.1300	3.2568
0.60	4.0	1.8928	2.4516	1.4616	2.6546

Table 2. Threshold Parameters for CRL-ZTP EWMA Charts for ATFS = 90

ZIP Parameters		Threshold Parameter Values			
		$\kappa = 0.25$		$\kappa = 0.45$	
π_0	λ_0	L_π	L_λ	L_π	L_λ
0.30	1.0	0.4261	6.2749	0.7875	5.4816
0.30	4.0	0.8403	5.0894	1.1953	4.3238
0.60	1.0	0.6801	4.6344	1.0388	4.1749
0.60	4.0	0.0001	3.1460	0.0001	2.8630

- Outbreaks, as determined by shifts in either one of the ZIP parameters or both, were injected in three phases: phase one (time periods 251-500), phase two (time periods 501-750), and phase three (time periods 751-1,000). These outbreaks were determined by shifting the parameter magnitudes upward. Three shapes were considered, namely spike, triangular, and ramp. Two intensity levels (M_0) were determined for each parameter: small ($M_\lambda = +3$; $M_\pi = +25\%$) and large ($M_\lambda = +6$; $M_\pi = +50\%$). Two outbreak durations (D) were considered, namely short (D = 15) and long (D = 30). For each intensity-duration (M-D) combination, random starts for the injected outbreaks for different phases were determined. For each shape-intensity-

length combination, three scenarios were considered, namely (Scenario 1) only λ increased, (Scenario 2) only π increased, and (Scenario 3) both λ and π increased. It is emphasized that the outbreaks are assumed to be stochastic and additive on the mean background disease process. Moreover, the outbreak equations added to the simulated disease background data followed those which were discussed in Fricker (2013).

- 5.) The charting procedures were empirically evaluated using several performance metrics discussed by Sonesson and Bock (2003) and Fricker (2013), with emphasis on their ability to properly detect the injected outbreaks. In particular, charts with lower conditional expected delay (CED) value and higher probability of successful detection (PSD), proportion of outbreaks detected (POD), and probability of true detection (PTD) values are determined to perform better. For each outbreak phase, the metrics were computed as follows:

$$\widehat{PSD} = \begin{cases} 1, & \text{if the chart signaled at least once during the outbreak} \\ 0, & \text{otherwise} \end{cases}, \quad (23)$$

$$\widehat{CED} = \text{time of first true signal} - \text{time of start of outbreak}, \quad (24)$$

$$\widehat{POD} = \frac{\text{total number of signals generated within the outbreak}}{\text{total outbreak duration}}, \quad (25)$$

$$\widehat{PTD} = \frac{\text{total number of signals generated within the outbreak}}{\text{total number of signals generated during the phase}}, \quad (26)$$

As numerical summaries, means were computed. Also, for comparison purposes, 95% confidence intervals for the mean values were estimated using bias-corrected bootstrapping procedure with 1,000 replications.

3.2. Simulation results and discussion

In this section, results of the simulation study are presented and discussed. To simplify the presentation, chart legends are used which are given in Table 3.

Table 3. Chart Legends

Legend	Chart Description
A	Bernoulli-ZIP EWMA Chart ($\kappa = 0.25$)
B	Bernoulli-ZIP EWMA Chart($\kappa = 0.45$)
C	CRL-ZTP EWMA Chart($\kappa = 0.25$)
D	CRL-ZTP EWMA Chart($\kappa = 0.45$)

Table 4 provides the means, as well as the 95% confidence intervals, of performance metrics of the four charts considered for small short spike outbreaks with different ZIP backgrounds. As a summary, all charts were able to soundly identify outbreaks defined by any shift scenarios, except the cases wherein outbreaks were only defined by shifts in π for disease backgrounds with smaller π_0 . Noticeably, mean PSDs were almost 100.00% (i.e., almost perfect detection success) when disease backgrounds had larger π_0 . Meanwhile, CEDs were typically longer (hence, slower detections) for outbreaks defined by shifts in π . Lastly, PODs and PTDs were generally higher when outbreaks were defined by shifts in both λ and π , but were observably lowest when outbreaks were only defined by shifts in π .

Table 4. Performance Metrics for Small Short Spike Outbreaks

Shifts	Chart	PSD		CED		POD		PTD	
		mean	95% CI	mean	95% CI	mean	95% CI	mean	95% CI
$\pi_0 = 0.30$ and $\lambda_0 = 1.0$									
λ only	A	96.87%	(96.27%, 97.40%)	4.4580	(4.3318, 4.5759)	26.49%	(25.85%, 27.06%)	52.58%	(51.74%, 53.46%)
	B	97.03%	(96.43%, 97.60%)	4.2943	(4.1889, 4.4188)	24.65%	(24.16%, 25.14%)	52.54%	(51.78%, 53.40%)
	C	96.80%	(96.17%, 97.37%)	4.8533	(4.7328, 4.9961)	20.12%	(19.69%, 20.60%)	46.41%	(45.54%, 47.20%)
	D	97.03%	(96.40%, 97.67%)	4.4367	(4.3166, 4.5704)	22.43%	(21.95%, 22.92%)	50.38%	(49.62%, 51.26%)
π only	A	66.50%	(64.88%, 68.23%)	7.3037	(7.1317, 7.4963)	11.85%	(11.36%, 12.36%)	31.97%	(30.94%, 32.98%)
	B	65.03%	(63.23%, 66.87%)	7.1910	(7.0112, 7.3968)	9.12%	(8.74%, 9.46%)	28.18%	(27.23%, 29.14%)
	C	63.20%	(61.46%, 64.83%)	7.4262	(7.2208, 7.6172)	9.24%	(8.86%, 9.59%)	27.68%	(26.83%, 28.72%)
	D	63.07%	(61.35%, 64.70%)	7.2894	(7.1042, 7.4930)	9.12%	(8.77%, 9.50%)	28.34%	(27.35%, 29.31%)
both λ and π	A	98.93%	(98.53%, 99.30%)	3.9068	(3.7908, 4.0187)	33.99%	(33.34%, 34.69%)	58.14%	(57.39%, 58.90%)
	B	99.03%	(98.70%, 99.40%)	3.7800	(3.6741, 3.8879)	30.70%	(30.18%, 31.26%)	57.84%	(57.07%, 58.52%)
	C	98.73%	(98.22%, 99.10%)	4.2482	(4.1471, 4.3546)	25.67%	(25.19%, 26.18%)	51.97%	(51.24%, 52.71%)
	D	98.97%	(98.53%, 99.30%)	3.8703	(3.7592, 3.9717)	28.39%	(27.87%, 28.96%)	55.89%	(55.18%, 56.63%)
$\pi_0 = 0.30$ and $\lambda_0 = 4.0$									
λ only	A	72.67%	(70.97%, 74.32%)	6.9577	(6.7816, 7.1342)	12.98%	(12.52%, 13.44%)	34.20%	(33.19%, 35.17%)
	B	72.43%	(70.97%, 73.99%)	6.7061	(6.5306, 6.8928)	11.78%	(11.36%, 12.16%)	32.03%	(31.11%, 33.14%)
	C	69.77%	(68.07%, 71.30%)	7.1824	(7.0002, 7.3531)	9.59%	(9.23%, 9.97%)	28.82%	(27.87%, 29.70%)
	D	72.07%	(70.33%, 73.57%)	6.8442	(6.6612, 7.0257)	9.06%	(8.77%, 9.36%)	29.27%	(28.47%, 30.17%)

π only	A	45.87%	(44.13%, 47.62%)	7.6677	(7.4206, 7.9091)	7.22%	(6.83%, 7.61%)	21.58%	(20.58%, 22.58%)
	B	47.23%	(45.27%, 49.07%)	7.6135	(7.3888, 7.8452)	6.40%	(6.08%, 6.74%)	19.60%	(18.66%, 20.54%)
	C	45.33%	(43.50%, 47.00%)	7.8395	(7.5971, 8.0943)	5.44%	(5.15%, 5.74%)	18.55%	(17.73%, 19.47%)
	D	46.53%	(44.87%, 48.27%)	7.7007	(7.4783, 7.9499)	4.75%	(4.50%, 4.98%)	17.72%	(16.87%, 18.47%)
both λ and π	A	82.23%	(80.90%, 83.60%)	6.6640	(6.5062, 6.8327)	17.70%	(17.12%, 18.26%)	40.90%	(39.88%, 42.08%)
	B	81.87%	(80.24%, 83.17%)	6.4268	(6.2598, 6.6207)	15.78%	(15.23%, 16.30%)	38.30%	(37.36%, 39.27%)
	C	79.83%	(78.37%, 81.23%)	6.9173	(6.7478, 7.0756)	12.74%	(12.33%, 13.19%)	34.68%	(33.83%, 35.57%)
	D	81.03%	(79.59%, 82.39%)	6.7195	(6.5516, 6.8820)	11.72%	(11.41%, 12.12%)	34.52%	(33.70%, 35.29%)

$\pi_0 = 0.60$ and $\lambda_0 = 1.0$									
λ only	A	100%	(44.13%, 47.62%)	6.9577	(7.4206, 7.9091)	12.98%	(6.83%, 7.61%)	34.20%	(20.58%, 22.58%)
	B	100%	(45.27%, 49.07%)	6.7061	(7.3888, 7.8452)	11.78%	(6.08%, 6.74%)	32.03%	(18.66%, 20.54%)
	C	100%	(43.50%, 47.00%)	7.1824	(7.5971, 8.0943)	9.59%	(5.15%, 5.74%)	28.82%	(17.73%, 19.47%)
	D	100%	(44.87%, 48.27%)	6.8442	(7.4783, 7.9499)	9.06%	(4.50%, 4.98%)	29.27%	(16.87%, 18.47%)
π only	A	100%	(100%, 100%)	7.6677	(2.1875, 2.2824)	7.22%	(62.53%, 63.74%)	21.58%	(68.03%, 69.36%)
	B	100%	(100%, 100%)	7.6135	(2.1100, 2.2123)	6.40%	(60.75%, 62.07%)	19.60%	(67.47%, 68.96%)
	C	100%	(100%, 100%)	7.8395	(2.2630, 2.3738)	5.44%	(46.89%, 48.17%)	18.55%	(64.21%, 65.47%)
	D	100%	(100%, 100%)	7.7007	(2.1133, 2.2137)	4.75%	(49.42%, 50.42%)	17.72%	(67.08%, 68.30%)
both λ and π	A	100%	(100%, 100%)	6.6640	(1.4669, 1.5267)	17.70%	(74.85%, 75.90%)	40.90%	(71.65%, 72.92%)
	B	100%	(100%, 100%)	6.4268	(1.4083, 1.4653)	15.78%	(74.62%, 75.62%)	38.30%	(71.47%, 72.88%)
	C	100%	(100%, 100%)	6.9173	(1.4672, 1.5270)	12.74%	(62.25%, 63.38%)	34.68%	(70.24%, 71.52%)
	D	100%	(100%, 100%)	6.7195	(1.4007, 1.4554)	11.72%	(67.79%, 68.69%)	34.52%	(73.61%, 74.65%)

$\pi_0 = 0.60$ and $\lambda_0 = 4.0$									
λ only	A	99.77%	(99.52%, 99.87%)	3.5340	(3.4441, 3.6303)	37.97%	(37.26%, 38.72%)	52.41%	(51.49%, 53.42%)
	B	99.77%	(99.57%, 99.90%)	3.1713	(3.0940, 3.2645)	43.96%	(43.17%, 44.93%)	44.98%	(43.54%, 46.25%)
	C	99.90%	(99.73%, 99.97%)	3.1875	(3.1086, 3.2716)	23.84%	(23.58%, 24.12%)	53.38%	(52.70%, 54.07%)
	D	99.90%	(99.73%, 99.97%)	2.9485	(2.8665, 3.0330)	29.68%	(29.33%, 30.06%)	58.47%	(57.77%, 59.18%)

π only	A	98.87%	(98.50%, 99.23%)	4.8485	(4.7394, 4.9671)	38.36%	(37.48%, 39.20%)	50.90%	(50.05%, 51.81%)
	B	98.70%	(98.17%, 99.07%)	4.2597	(4.1503, 4.3768)	43.70%	(42.46%, 44.75%)	43.51%	(42.35%, 44.57%)
	C	97.27%	(96.63%, 97.80%)	4.9360	(4.8163, 5.0598)	14.43%	(14.20%, 14.68%)	41.01%	(40.32%, 41.70%)
	D	96.77%	(96.10%, 97.37%)	4.6513	(4.5334, 4.7797)	17.34%	(17.02%, 17.69%)	45.15%	(44.35%, 45.94%)
Both λ and π	A	99.97%	(99.80%, 100.00%)	2.8160	(2.7550, 2.8768)	50.99%	(50.22%, 51.77%)	58.27%	(57.35%, 59.18%)
	B	99.97%	(99.83%, 100%)	2.5277	(2.4661, 2.5966)	56.73%	(55.80%, 57.62%)	50.16%	(48.95%, 51.38%)
	C	100%	(100%, 100%)	2.5783	(2.5157, 2.6381)	29.34%	(29.04%, 29.61%)	58.39%	(57.74%, 59.08%)
	D	100%	(100%, 100%)	2.3840	(2.3191, 2.4477)	36.80%	(36.41%, 37.20%)	63.48%	(62.82%, 64.08%)

Table 5 provides the means, as well as the 95% confidence intervals, of performance metrics of the four charts considered for small short triangular outbreaks with different ZIP backgrounds. As a summary, all charts were able to soundly identify outbreaks defined by any shift scenarios, particularly when outbreaks were defined by either shifts in λ only or in both λ and π when disease background has lower π_0 , and when outbreaks were defined by shifts in π only when disease background has higher π_0 . Meanwhile, CEDs were mostly similar in any scenario considered. Lastly, PODs and PTDs were generally higher when outbreaks were defined by shifts in both λ and π , but were observably lowest when outbreaks were only defined by shifts in π whenever the disease background has lower π_0 . However, the converse was true whenever the disease background has higher π_0 .

Table 5. Performance Metrics for Small Short Triangular Outbreaks

Shifts	Chart	PSD		CED		POD		PTD	
		mean	95% CI	mean	95% CI	mean	95% CI	mean	95% CI
$\pi_0 = 0.30$ and $\lambda_0 = 1.0$									
λ only	A	69.67%	(68.00%, 71.26%)	6.6021	(6.3819, 6.8145)	10.78%	(10.34%, 11.18%)	30.44%	(29.48%, 31.40%)
	B	69.97%	(68.40%, 71.67%)	6.5695	(6.3409, 6.8074)	9.56%	(9.25%, 9.89%)	29.42%	(28.47%, 30.25%)
	C	68.07%	(66.32%, 69.63%)	7.0038	(6.7759, 7.2398)	10.05%	(9.70%, 10.42%)	29.36%	(28.53%, 30.27%)
	D	67.93%	(66.23%, 69.49%)	6.7933	(6.5645, 7.0363)	9.39%	(9.09%, 9.74%)	29.01%	(28.10%, 29.82%)

π only	A	62.87%	(60.97%, 64.63%)	7.2641	(7.0825, 7.4493)	10.34%	(9.91%, 10.78%)	29.26%	(28.32%, 30.26%)
	B	61.20%	(59.58%, 62.92%)	7.0940	(6.9008, 7.2895)	8.24%	(7.91%, 8.57%)	26.32%	(25.27%, 27.21%)
	C	60.07%	(58.17%, 61.70%)	7.4921	(7.2938, 7.6907)	8.03%	(7.68%, 8.37%)	25.11%	(24.24%, 26.02%)
	D	61.00%	(59.26%, 62.76%)	7.4678	(7.2686, 7.6780)	8.10%	(7.80%, 8.44%)	26.22%	(25.29%, 27.19%)
both λ and π	A	73.87%	(72.37%, 75.46%)	6.4251	(6.2336, 6.6403)	12.35%	(11.90%, 12.78%)	33.35%	(32.41%, 34.35%)
	B	73.97%	(72.43%, 75.53%)	6.3209	(6.1169, 6.5359)	11.00%	(10.66%, 11.35%)	32.41%	(31.52%, 33.33%)
	C	72.73%	(71.10%, 74.10%)	6.7644	(6.5646, 6.9758)	11.55%	(11.17%, 11.96%)	32.18%	(31.28%, 33.16%)
	D	72.53%	(70.93%, 74.10%)	6.5334	(6.3271, 6.7352)	10.84%	(10.46%, 11.22%)	32.11%	(31.00%, 32.92%)

$\pi_0 = 0.30$ and $\lambda_0 = 4.0$									
λ only	A	86.37%	(85.17%, 87.63%)	5.8687	(5.6822, 6.0477)	16.58%	(16.06%, 17.09%)	39.60%	(38.66%, 40.48%)
	B	87.67%	(86.47%, 88.90%)	5.5153	(5.3527, 5.7203)	16.50%	(16.06%, 16.95%)	39.58%	(38.71%, 40.43%)
	C	85.03%	(83.70%, 86.27%)	6.2239	(6.0366, 6.4057)	13.58%	(13.16%, 13.96%)	36.11%	(35.32%, 37.00%)
	D	87.37%	(86.20%, 88.57%)	5.8422	(5.6620, 6.0172)	12.65%	(12.36%, 12.95%)	36.47%	(35.68%, 37.26%)
π only	A	73.10%	(71.53%, 74.60%)	7.0012	(6.8188, 7.1730)	13.17%	(12.69%, 13.67%)	34.22%	(33.22%, 35.23%)
	B	73.47%	(71.77%, 74.90%)	6.6621	(6.4830, 6.8560)	12.03%	(11.60%, 12.46%)	32.38%	(31.41%, 33.38%)
	C	71.63%	(69.90%, 73.43%)	7.2344	(7.0519, 7.4315)	9.83%	(9.49%, 10.21%)	29.51%	(28.57%, 30.41%)
	D	73.60%	(72.00%, 75.20%)	6.9666	(6.7879, 7.1624)	9.27%	(8.97%, 9.60%)	29.65%	(28.84%, 30.52%)
both λ and π	A	87.17%	(85.93%, 88.40%)	5.6874	(5.5208, 5.8572)	16.97%	(16.53%, 17.49%)	40.31%	(39.48%, 41.20%)
	B	89.17%	(88.00%, 90.20%)	5.5603	(5.3790, 5.7279)	16.71%	(16.28%, 17.13%)	40.06%	(39.24%, 41.03%)
	C	87.03%	(85.70%, 88.17%)	6.3059	(6.1136, 6.4904)	13.84%	(13.47%, 14.26%)	36.70%	(35.91%, 37.46%)
	D	88.57%	(87.33%, 89.63%)	5.9071	(5.7367, 6.1098)	12.85%	(12.53%, 13.14%)	37.10%	(36.29%, 37.88%)

$\pi_0 = 0.60$ and $\lambda_0 = 1.0$									
λ only	A	25.80%	(24.40%, 27.47%)	6.6615	(6.2849, 7.0497)	2.77%	(2.55%, 2.99%)	9.58%	(8.78%, 10.39%)
	B	27.87%	(26.26%, 29.50%)	6.5482	(6.1783, 6.9538)	2.93%	(2.71%, 3.16%)	9.22%	(8.61%, 10.07%)
	C	27.37%	(25.73%, 29.00%)	6.8589	(6.4840, 7.2494)	2.47%	(2.29%, 2.65%)	8.72%	(8.12%, 9.40%)
	D	28.37%	(26.67%, 30.00%)	6.8241	(6.4691, 7.1906)	2.39%	(2.22%, 2.55%)	9.26%	(8.62%, 9.95%)

π only	A	83.70%	(82.27%, 84.83%)	6.8014	(6.6449, 6.9492)	18.37%	(17.75%, 18.96%)	40.14%	(39.11%, 41.17%)
	B	80.13%	(78.63%, 81.53%)	6.6598	(6.4893, 6.8152)	16.55%	(15.92%, 17.16%)	36.85%	(35.85%, 38.03%)
	C	81.40%	(80.07%, 82.80%)	6.8789	(6.6959, 7.0307)	13.26%	(12.77%, 13.74%)	34.12%	(33.25%, 34.98%)
	D	80.63%	(79.23%, 82.11%)	6.7995	(6.6210, 6.9606)	12.14%	(11.75%, 12.52%)	33.97%	(33.11%, 34.81%)
both λ and π	A	25.97%	(24.46%, 27.51%)	6.5317	(6.1599, 6.9128)	2.81%	(2.61%, 3.03%)	9.65%	(8.94%, 10.47%)
	B	28.17%	(26.53%, 29.80%)	6.8700	(6.4717, 7.2602)	2.98%	(2.77%, 3.20%)	9.61%	(8.90%, 10.42%)
	C	29.20%	(27.57%, 30.90%)	7.0600	(6.6986, 7.4108)	2.57%	(2.40%, 2.76%)	9.11%	(8.44%, 9.76%)
	D	29.90%	(28.23%, 31.60%)	7.0976	(6.7435, 7.4850)	2.49%	(2.35%, 2.66%)	9.84%	(9.25%, 10.54%)
$\pi_0 = 0.60$ and $\lambda_0 = 4.0$									
λ only	A	35.60%	(33.83%, 37.40%)	6.3201	(5.9810, 6.6761)	4.39%	(4.10%, 4.73%)	11.66%	(10.88%, 12.67%)
	B	48.77%	(46.55%, 50.70%)	6.1421	(5.8308, 6.4630)	7.42%	(7.00%, 7.91%)	11.34%	(10.48%, 12.27%)
	C	40.20%	(38.50%, 42.30%)	7.0474	(6.6965, 7.4023)	3.00%	(2.86%, 3.15%)	12.00%	(11.39%, 12.63%)
	D	43.23%	(41.51%, 45.03%)	7.0572	(6.6952, 7.3958)	3.40%	(3.24%, 3.56%)	13.79%	(13.14%, 14.41%)
π only	A	95.47%	(94.67%, 96.17%)	5.5207	(5.3954, 5.6443)	27.76%	(26.92%, 28.55%)	44.19%	(43.11%, 45.13%)
	B	95.70%	(94.83%, 96.33%)	4.8818	(4.7510, 5.0069)	33.11%	(32.18%, 34.17%)	37.67%	(36.62%, 39.00%)
	C	94.27%	(93.43%, 95.13%)	5.3640	(5.2241, 5.4854)	12.42%	(12.20%, 12.65%)	37.49%	(36.71%, 38.26%)
	D	94.10%	(93.20%, 94.93%)	4.9977	(4.8694, 5.1323)	14.82%	(14.51%, 15.12%)	41.50%	(40.77%, 42.33%)
both λ and π	A	35.53%	(33.77%, 37.30%)	6.4587	(6.0988, 6.8245)	4.37%	(4.06%, 4.70%)	11.59%	(10.77%, 12.46%)
	B	49.17%	(47.30%, 51.17%)	6.0970	(5.7786, 6.3881)	7.41%	(6.98%, 7.88%)	11.33%	(10.52%, 12.33%)
	C	41.17%	(39.41%, 43.02%)	7.0133	(6.7023, 7.3658)	3.11%	(2.95%, 3.27%)	12.55%	(11.94%, 13.28%)
	D	44.53%	(42.77%, 46.43%)	6.8932	(6.5029, 7.2135)	3.55%	(3.41%, 3.74%)	14.28%	(13.65%, 14.99%)

Table 6 provides the means, as well as the 95% confidence intervals, of performance metrics of the four charts considered for small short ramp outbreaks with different ZIP backgrounds. As a summary, observed patterns were largely similar to those of the cases considered for small short triangular outbreaks. The only exception was that CEDs were shortest when outbreaks were defined by shifts only in π whenever the disease background has higher π_0 .

Table 6. Performance Metrics for Small Short Ramp Outbreaks

Shifts	Chart	PSD		CED		POD		PTD	
		mean	95% CI	mean	95% CI	mean	95% CI	mean	95% CI
$\pi_0 = 0.30$ and $\lambda_0 = 1.0$									
λ only	A	86.27%	(85.07%, 87.47%)	6.4167	(6.2644, 6.5695)	20.64%	(20.05%, 21.29%)	45.09%	(44.15%, 45.96%)
	B	83.53%	(82.33%, 84.95%)	6.3819	(6.2251, 6.5587)	15.83%	(15.39%, 16.23%)	40.94%	(40.04%, 41.88%)
	C	82.47%	(81.07%, 83.83%)	6.7249	(6.5638, 6.8994)	16.69%	(16.24%, 17.16%)	40.45%	(39.49%, 41.35%)
	D	82.90%	(81.50%, 84.37%)	6.5679	(6.4001, 6.7014)	16.42%	(15.91%, 16.86%)	41.50%	(40.61%, 42.41%)
π only	A	63.70%	(61.90%, 65.47%)	7.2701	(7.0565, 7.4680)	10.52%	(10.09%, 10.99%)	29.49%	(28.42%, 30.53%)
	B	61.90%	(60.07%, 63.67%)	7.0936	(6.9018, 7.3155)	8.36%	(8.03%, 8.71%)	26.48%	(25.55%, 27.37%)
	C	59.87%	(58.07%, 61.70%)	7.3736	(7.1606, 7.5847)	8.19%	(7.86%, 8.55%)	25.46%	(24.52%, 26.38%)
	D	60.13%	(58.23%, 61.83%)	7.2368	(7.0373, 7.4475)	8.14%	(7.84%, 8.49%)	26.23%	(25.37%, 27.15%)
both λ and π	A	89.97%	(88.85%, 91.02%)	6.1556	(6.0112, 6.2959)	23.39%	(22.69%, 24.09%)	48.25%	(47.34%, 49.17%)
	B	87.80%	(86.67%, 89.05%)	6.0199	(5.8789, 6.1849)	17.97%	(17.50%, 18.48%)	44.04%	(43.22%, 44.96%)
	C	86.93%	(85.77%, 88.17%)	6.4415	(6.2913, 6.5830)	18.94%	(18.45%, 19.49%)	43.67%	(42.79%, 44.43%)
	D	87.50%	(86.33%, 88.70%)	6.3385	(6.1827, 6.4848)	18.78%	(18.29%, 19.24%)	44.92%	(43.98%, 45.77%)
$\pi_0 = 0.30$ and $\lambda_0 = 4.0$									
λ only	A	97.83%	(97.29%, 98.30%)	5.3138	(5.1915, 5.4239)	33.25%	(32.59%, 33.87%)	56.11%	(55.32%, 56.93%)
	B	97.23%	(96.57%, 97.77%)	5.0823	(4.9707, 5.1910)	29.59%	(28.97%, 30.16%)	54.18%	(53.36%, 54.98%)
	C	97.60%	(96.97%, 98.10%)	5.7182	(5.5929, 5.8340)	24.59%	(24.08%, 25.12%)	50.60%	(49.89%, 51.31%)
	D	97.50%	(96.90%, 98.07%)	5.3493	(5.2381, 5.4738)	21.19%	(20.81%, 21.58%)	48.96%	(48.24%, 49.68%)
π only	A	73.00%	(71.36%, 74.57%)	6.9622	(6.7905, 7.1622)	13.28%	(12.78%, 13.80%)	34.40%	(33.32%, 35.47%)
	B	73.63%	(72.03%, 75.16%)	6.7343	(6.5583, 6.9135)	12.02%	(11.60%, 12.47%)	32.16%	(31.23%, 33.14%)
	C	70.57%	(68.93%, 72.23%)	7.0836	(6.8921, 7.2601)	9.86%	(9.47%, 10.23%)	29.30%	(28.37%, 30.17%)
	D	73.77%	(72.13%, 75.40%)	6.8600	(6.6730, 7.0615)	9.53%	(9.23%, 9.85%)	30.11%	(29.29%, 31.05%)
both λ and π	A	97.97%	(97.43%, 98.43%)	5.2302	(5.1136, 5.3421)	33.75%	(32.98%, 34.46%)	56.38%	(55.59%, 57.14%)
	B	97.33%	(96.63%, 97.83%)	5.0488	(4.9322, 5.1609)	30.00%	(29.40%, 30.65%)	54.44%	(53.54%, 55.24%)
	C	97.77%	(97.23%, 98.27%)	5.6067	(5.4931, 5.7285)	24.92%	(24.34%, 25.49%)	50.82%	(50.15%, 51.53%)
	D	97.53%	(96.93%, 98.03%)	5.3120	(5.2057, 5.4387)	21.61%	(21.25%, 22.00%)	49.56%	(48.84%, 50.31%)

$\pi_0 = 0.60$ and $\lambda_0 = 1.0$									
λ only	A	26.73%	(25.10%, 28.27%)	6.3671	(6.0084, 6.7365)	2.91%	(2.69%, 3.15%)	10.07%	(9.31%, 10.91%)
	B	28.67%	(26.95%, 30.37%)	6.4879	(6.1465, 6.8835)	3.04%	(2.82%, 3.30%)	9.82%	(9.08%, 10.62%)
	C	29.57%	(28.00%, 31.43%)	6.8969	(6.5725, 7.2832)	2.64%	(2.45%, 2.82%)	9.70%	(9.06%, 10.45%)
	D	30.23%	(28.63%, 31.83%)	6.8179	(6.4013, 7.2016)	2.56%	(2.40%, 2.71%)	10.11%	(9.43%, 10.86%)
π only	A	83.83%	(82.50%, 85.10%)	6.6583	(6.5067, 6.8053)	18.68%	(18.03%, 19.27%)	40.69%	(39.70%, 41.60%)
	B	81.07%	(79.57%, 82.53%)	6.5051	(6.3494, 6.6638)	16.93%	(16.34%, 17.53%)	37.22%	(36.21%, 38.31%)
	C	81.30%	(79.81%, 82.67%)	6.8588	(6.7014, 7.0255)	13.38%	(12.90%, 13.86%)	34.30%	(33.49%, 35.16%)
	D	81.40%	(80.00%, 82.80%)	6.6858	(6.5438, 6.8456)	12.31%	(11.98%, 12.69%)	34.27%	(33.49%, 35.25%)
both λ and π	A	28.20%	(26.63%, 29.90%)	6.6360	(6.2451, 6.9932)	3.01%	(2.79%, 3.23%)	10.29%	(9.59%, 11.18%)
	B	30.40%	(28.53%, 32.07%)	6.6260	(6.2221, 7.0071)	3.18%	(2.96%, 3.41%)	10.07%	(9.37%, 10.79%)
	C	30.83%	(29.12%, 32.53%)	7.0906	(6.7383, 7.4784)	2.72%	(2.55%, 2.92%)	9.66%	(9.02%, 10.31%)
	D	32.20%	(30.56%, 33.93%)	7.0815	(6.7327, 7.4898)	2.71%	(2.55%, 2.89%)	10.42%	(9.78%, 11.11%)

$\pi_0 = 0.60$ and $\lambda_0 = 4.0$									
λ only	A	37.23%	(35.53%, 39.18%)	6.5174	(6.1689, 6.9199)	4.62%	(4.32%, 4.98%)	12.30%	(11.42%, 13.22%)
	B	50.57%	(48.53%, 52.42%)	6.2088	(5.9072, 6.5339)	7.79%	(7.31%, 8.29%)	12.18%	(11.33%, 13.14%)
	C	43.23%	(41.63%, 45.20%)	7.1837	(6.8084, 7.5188)	3.26%	(3.11%, 3.40%)	13.21%	(12.57%, 13.84%)
	D	46.30%	(44.50%, 48.03%)	6.9779	(6.6418, 7.3625)	3.72%	(3.57%, 3.89%)	14.96%	(14.32%, 15.65%)
π only	A	95.77%	(95.05%, 96.50%)	5.5022	(5.3707, 5.6328)	27.46%	(26.64%, 28.29%)	43.94%	(42.82%, 44.86%)
	B	96.40%	(95.60%, 97.00%)	4.9072	(4.7759, 5.0304)	33.23%	(32.21%, 34.15%)	37.83%	(36.60%, 38.98%)
	C	95.47%	(94.67%, 96.20%)	5.4215	(5.2895, 5.5522)	12.61%	(12.37%, 12.83%)	37.90%	(37.20%, 38.55%)
	D	95.57%	(94.77%, 96.37%)	5.1413	(5.0031, 5.2890)	15.00%	(14.70%, 15.28%)	41.69%	(40.90%, 42.50%)
both λ and π	A	37.37%	(35.49%, 39.19%)	6.4357	(6.1012, 6.7668)	4.56%	(4.25%, 4.90%)	12.18%	(11.35%, 13.04%)
	B	50.63%	(48.77%, 52.68%)	6.1160	(5.8223, 6.4872)	7.72%	(7.27%, 8.18%)	11.77%	(10.98%, 12.77%)
	C	43.67%	(41.87%, 45.57%)	7.1542	(6.8193, 7.5104)	3.29%	(3.14%, 3.45%)	13.39%	(12.69%, 14.03%)
	D	46.77%	(44.94%, 48.53%)	6.9035	(6.5870, 7.2489)	3.79%	(3.62%, 3.95%)	15.24%	(14.53%, 15.90%)

Mostly similar results and patterns were observed for any other simulation scenarios considered (i.e., different outbreak magnitudes and lengths). However, they were not presented anymore due to space limitations. Nevertheless, all results and associated tables are available from the authors upon request.

In summary, most outbreaks were detected by both EWMA charts, especially when the outbreaks were determined by shifts in λ for scenarios with small disease backgrounds, and when the outbreaks were determined by shifts in π for scenarios with large disease backgrounds. In particular, with PSDs, both proposed EWMA charts performed considerably well, with detection rates nearly comparable with one another. In particular, Bernoulli-ZIP EWMA charts mostly had higher PSD rates when π_0 is smaller, while the same statement is true for CRL-ZTP EWMA charts when π_0 is higher. Regarding CEDs, several situations likewise saw both EWMA charts having comparable values, wherein detection lags were mostly shorter for shifts in λ only or in both λ and π when π_0 is smaller, and for shifts in π only when π_0 is larger. Concerning PODs, the proposed Bernoulli-ZIP EWMA charts were found to perform best in almost all situations considered. In particular, it was observed that a smaller weighting parameter κ must be desired if the disease background is smaller, and vice versa. Likewise with PTDs, Bernoulli-ZIP EWMA charts mostly had better performances than the CRL-ZTP EWMA charts, except the situations when the disease background had higher π_0 and λ_0 . Also, quite noticeable was that regardless of the disease background or the outbreak situation, smaller κ must be desired for Bernoulli-ZIP EWMA charts, while larger κ must be desired for CRL-ZTP EWMA charts.

Concerning types of outbreaks, all charts performed best when spike outbreaks were considered; worst when triangular outbreaks were considered. In particular, PTDs were typically higher and CEDs were generally shorter by almost two to three days with spike outbreaks as against similar situations but with triangular or ramp shapes. These findings can be explained by that spike outbreaks had sudden shifts; hence, were easier to detect. On the other hand, triangular outbreaks gradually shift and only peak once; hence, were harder to detect particularly by those charts with memory. With ramp outbreaks, while shifts were also gradual, the peaks lasted for a considerable amount of time; hence, were detected better than triangular outbreaks. Nevertheless, all charts were able to make the detection mostly during the first half of the outbreak period, which presented an opportunity to possibly introduce an intervention to control the infection spread before the outbreak is at its peak.

4. Case Application

4.1 The measles dataset

To demonstrate how the proposed combined EWMA schemes may be used with real data, all four previously considered charts were applied with actual

measles data consisting of both laboratory confirmed and epidemiologically-linked measles daily cases in the National Capital Region (NCR) spanning the dates January 1, 2010 to January 14, 2015, obtained from the Philippine Department of Health - National Epidemiology Center, NCR.

A plot of the data is provided in Figure 1, albeit truncated at 100. It should be noted that recorded cases from December 17, 2013 up to February 12, 2014 exceeded 100, hence are not shown in the graph. Three significant shifts can be observed; which upon confirmation from news reports, they were identified as outbreaks (Aragones, 2013; Agreda, 2013; RSJ, GMA News, 2013; Santos, 2013). The identified outbreak periods are the following:

Outbreak 1: January 1, 2010 to May 31, 2010 (151 days)

Outbreak 2: January 1, 2011 to May 31, 2011 (151 days)

Outbreak 3: September 1, 2013 to April 15, 2014 (227 days)

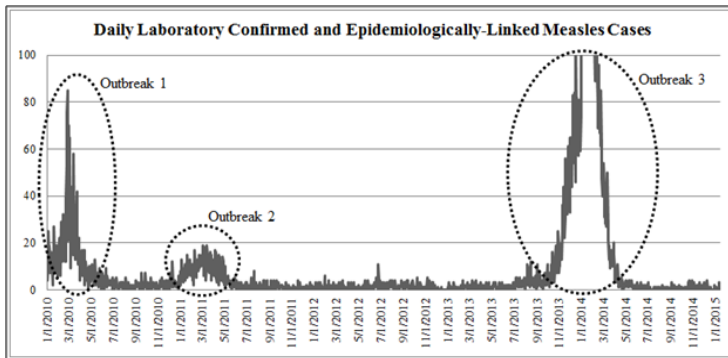


Figure 1. Time-series plot of daily laboratory confirmed and epidemiologically-linked measles cases in NCR from January 1, 2010 to January 14, 2015.

4.2. Phase I procedure

For the Phase I procedure, daily measles counts from January 1, 2010 up to December 31, 2012 were used to identify the disease background characteristics. In particular, the ZIP parameters π and λ were estimated. Prior to the estimation, pre-processing of the data was done. In which case, this procedure involved the deletion of the identified outbreak periods (January 1, 2010 to May 31, 2010 and January 1, 2011 to May 31, 2011) in the background. The resulting data consisted of 794 observations.

It was observed that the mean daily count was 1.3438, while the variance was 2.3116; accounting for a variance-to-mean dispersion index of 72.02%. Furthermore, out of the 794 observations, 35.26% are zeros. These findings may have provided hints that the underlying disease background follows that of the ZIP distribution. Hence, based on the assumption of a ZIP background process,

estimates of the parameters were provided by their corresponding MLEs, in which case were computed to be $\hat{\pi}_{mle} = 0.7930$ and $\hat{\lambda}_{mle} = 1.6946$. To formally determine whether the background indeed followed a ZIP process, the ZIP-Score Test by Xie et al. (2001) was employed. In which case, the test statistic was computed to be

$$S = \frac{(n_0 - np_0)^2}{np_0(1 - p_0) - n\bar{y}p_0^2} = 216.73, \tag{27}$$

where $n = 794$ is the total number of observations, $n_0 = 280$ is the number of zeros, $\bar{y} = 1.3438$ is the sample mean of the observations, and $p_0 = e^{-\hat{\lambda}_{mle}} = e^{-1.6946}$. Asymptotically, $S \sim \chi^2_{(1)}$; hence, the observed test statistic $S = 216.73$ had a p -value < 0.0001 . Clearly, the p-value provided overwhelming evidence against H_0 , in which case the decision to reject H_0 meant that the overdispersion index was significant. Thus, the underlying distribution was indeed that of the ZIP distribution.

For the chart thresholds, since the estimated ZIP parameters were closest to the simulation situation wherein $\pi = 0.60$ and $\lambda = 1.0$, then the threshold parameters for all the combined control charting procedures were correspondingly set to be equivalent to those of the aforementioned situation. In addition to the simultaneous charting procedures, the historical limits procedure was also applied in order to demonstrate the comparability of the proposed charting schemes with the current system in use (Catindig and Niñal, 2008). It can be recalled that the historical limits procedure is simply a variant of the standard upper Shewhart chart utilizing a 2σ -limit. Thus, the historical limits threshold was computed as follows:

$$h_{HL} = \hat{\mu} + 2\hat{\sigma} = 4.3846, \tag{28}$$

i.e., the chart was made to generate a signal whenever the observed count for the particular day is at least five. A summary of the computed thresholds for each chart is provided in Table 7.

Table 7. Chart Thresholds Used as Applied in Measles Data

Chart	h_{λ}	h_{π}
Historical Limits	4.3846	*
Bernoulli-ZIP EWMA Chart ($\kappa=0.25$)	2.7638	0.9860
Bernoulli-ZIP EWMA Chart ($\kappa=0.45$)	3.7081	1.0840
CRL-ZTP EWMA Chart ($\kappa=0.25$)	4.0610	2.0537
CRL-ZTP EWMA Chart ($\kappa=0.45$)	4.0646	1.4280

**The historical limits procedure only used one threshold (i.e., a single-charting scheme), which simply monitors upward shifts in λ .*

4.3. Phase II procedure

A summary of all the charts considered regarding their performances in Phase II are provided in Table 8. The charts were assessed similarly with the framework provided in the simulations, wherein their abilities to detect the September 1, 2013 – April 15, 2014 outbreak (lasting for 227 days) were given emphasis. In addition, ATFS measures were also provided to assess their regularity of generating false signals during non-outbreak periods (January 1, 2013 – August 31, 2013 and April 16, 2014 – January 14, 2015; lasting for 517 days). Visualizations of these charts are provided in the corresponding figures identified in the second column of Table 8.

Regarding ATFS, it can be observed that the least specific were the CRL-ZTP EWMA variants, which generate false signals almost at least once in a span of less than two weeks. The Bernoulli-ZIP EWMA charts, on the other hand, had widely varying ATFSs as evidenced by a value of 32.31 for $\kappa = 0.25$ and a value of 129.25 for $\kappa = 0.45$. Lastly, the historical limits procedure was seen to generate false signals almost once a month, with ATFS value of 34.47. In which case, only the Bernoulli-ZIP EWMA chart with $\kappa = 0.25$ was found to have ATFS performances which were comparable to the current system in use.

Table 8. Performance Assessment of the Different Charts as Applied in Measles Data

Chart	Figure	ATFS	PSD*	CED**	POD	PTD
Historical Limits	Figure 2	34.47	✓	Sept. 2, 2013	84.58%	92.75%
Bernoulli-ZIP EWMA Chart ($\kappa = 0.25$)	Figure 3	32.31	✓	Sept. 2, 2013	82.38%	92.12%
Bernoulli-ZIP EWMA Chart ($\kappa = 0.45$)	Figure 4	129.25	✓	Sept. 2, 2013	79.30%	97.83%
CRL-ZTP EWMA Chart ($\kappa = 0.25$)	Figure 5	9.40	✓	Sept. 2, 2013	84.58%	77.73%
CRL-ZTP EWMA Chart ($\kappa = 0.45$)	Figure 6	5.07	✓	Sept. 1, 2013	85.02%	65.42%

*The PSD in this table is a pseudo-measurement, in which case only identified if the Sept. 1, 2013 - Apr. 15, 2014 outbreak was detected.

**The CED in this table is a pseudo-measurement, in which case is the first day that the chart signaled during the identified outbreak period.

Since it is impossible to measure PSD using only a single dataset, then a pseudo-measurement is provided instead. To that end, the PSD was taken as simply either the outbreak was detected or not (i.e., the control charting procedure signaled at least once during the entire outbreak duration). Clearly, all charts were able to detect the outbreak. In fact, all were able to detect the outbreak as soon as the second day (September 2, 2013), except for the CRL-ZTP EWMA chart with

$\kappa = 0.45$, wherein it was able to detect sooner (September 1, 2013). The reason for the detection right on the second day was quite trivial because 11 confirmed cases were recorded on that particular day. Regarding the earlier detection of CRL-ZTP EWMA chart with $\kappa = 0.45$, the reason was identified to be because it was the least specific (which also translates to being the most sensitive); hence, an observed value of three cases on September 1, 2013 was already enough to generate a signal. The first day that these charts signaled during the entire outbreak duration (i.e., the time of outbreak detection) was used as a pseudo-measurement of CED due to the same reasons cited above.

Further discussing their abilities to soundly detect the outbreak, POD and PTD metrics were also computed. Likewise, seen in the table was that CRL-ZTP EWMA charts, in general, were found to be the most sensitive (i.e., with typically higher PODs). The historical limits procedure, on the other hand, was observed to have a POD of 84.58%. Lastly, concerning PTDs, the CRL-ZTP EWMA charts were observed to generate the least true positives (as evidenced by typically lower PTDs). With the Bernoulli-ZIP EWMA charts applied in the given dataset, higher κ was observed to have higher PTD (97.83%), and vice versa (92.12%). Regarding the performance of the historical limits procedure, a PTD of 92.75% was observed. In which case, for both POD and PTD, only the Bernoulli-ZIP EWMA charts were found to have comparable performances with the current system in use.

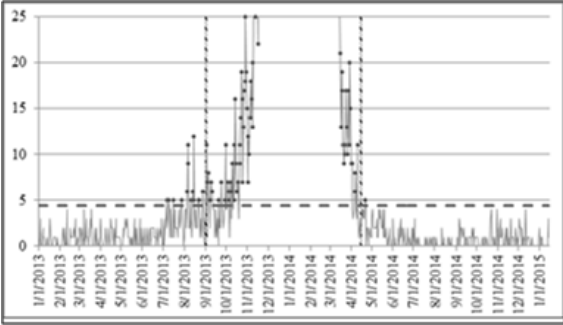


Figure 2. Visualization of the historical limits control charting procedure. The threshold is given by the broken horizontal line. Bold markers denote the instances the chart signaled. The Sept. 1, 2013 to Apr. 15, 2014 outbreak period is enclosed within the black dashed vertical lines.

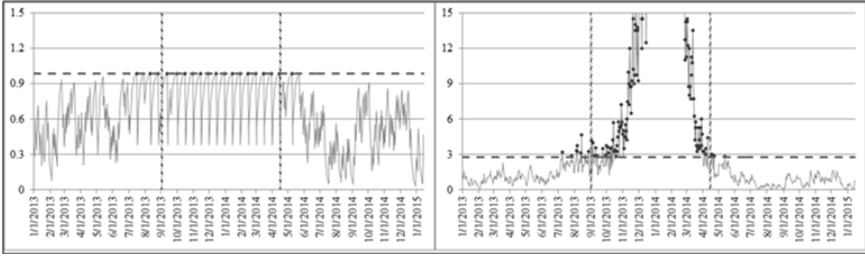


Figure 3. Visualization of the Bernoulli-ZIP EWMA charting procedures ($\kappa = 0.25$). [Left: Bernoulli-EWMA Chart, Right: ZIP-EWMA Chart] The thresholds are given by the broken horizontal lines. Bold markers denote the instances the chart signaled. The Sept. 1, 2013 to Apr. 15, 2014 outbreak period is enclosed within the black dashed vertical lines.

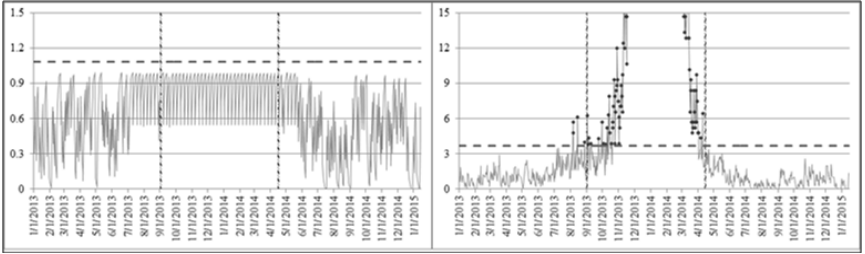


Figure 4. Visualization of the Bernoulli-ZIP EWMA charting procedures ($\kappa = 0.45$). [Left: Bernoulli-EWMA Chart, Right: ZIP-EWMA Chart] The thresholds are given by the broken horizontal lines. Bold markers denote the instances the chart signaled. The Sept. 1, 2013 to Apr. 15, 2014 outbreak period is enclosed within the black dashed vertical lines.

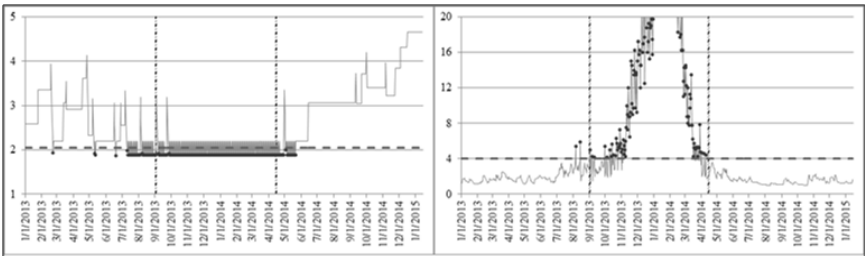


Figure 5. Visualization of the CRL-ZTP EWMA charting procedures ($\kappa = 0.25$). [Left: CRL-EWMA Chart, Right: ZTP-EWMA Chart] The thresholds are given by the broken horizontal lines. Bold markers denote the instances the chart signaled. The Sept. 1, 2013 to Apr. 15, 2014 outbreak period is enclosed within the black dashed vertical lines.

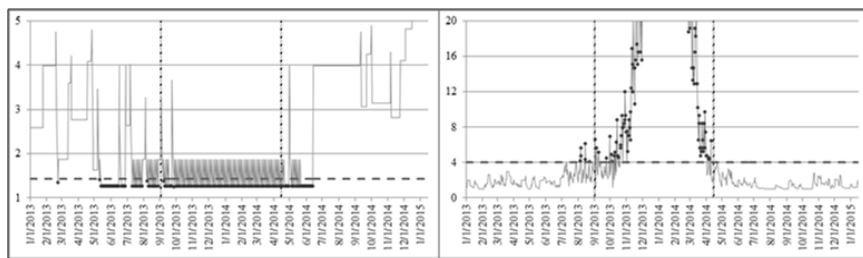


Figure 6. Visualization of the CRL-ZTP EWMA charting procedures ($\kappa = 0.45$). [Left: CRL-EWMA Chart, Right: ZTP-EWMA Chart] The thresholds are given by the broken horizontal lines. Bold markers denote the instances the chart signaled. The Sept. 1, 2013 to Apr. 15, 2014 outbreak period is enclosed within the black dashed vertical lines.

5. Conclusion

This paper has discussed and demonstrated the possibility of using a combined charting procedure for disease surveillance, wherein disease background follows a ZIP distribution. The rationale behind a combined charting procedure is that the ZIP process is characterized by two parameters. Hence, simultaneous individual charting procedure for each parameter is proposed in an effort to improve shift detection (i.e., outbreak detection in the context of disease surveillance). While this idea has already been explored in the context of SPC, two EWMA charts were developed, namely Bernoulli-ZIP EWMA and CRL-ZTP EWMA charts.

Based on an extensive simulation study, the proposed Bernoulli-ZIP EWMA charting procedure mostly had better performances than the CRL-ZTP EWMA chart regardless of outbreak scenario considered. Furthermore, it was empirically shown that a higher weighting parameter must be desired if the disease background has large parameters, and vice versa. These results were mostly consistent with the demonstration on actual measles data. For demonstration purposes, the current system in place (the historical limits method) was set as the reference chart. In which case, it was identified that the Bernoulli-ZIP EWMA chart had mostly equivalent performances to those of the historical limits procedure.

6. Recommendations

While this paper has offered some perspectives on the application of combined charting procedures for disease surveillance, several research directions can still be explored to enhance the proposed surveillance schemes. First, a fast initial response (FIR) feature may be added to improve sensitivity without compromising specificity. In this case, FIR is defined to be an initialization enhancement wherein the control chart starts at some value closer to the threshold (which, in most cases, is taken to be halfway between the threshold and the original initialization value). Furthermore, when the disease background is dynamic, an adaptive threshold scheme may be used to account for the gradual (or even seasonal) changes in

the background process. The most common procedure is a moving average and a moving standard deviation with a pre-identified window. Lastly, additional considerations in the simulation framework may be in order to further improve the evaluation of the proposed control charting procedures. These may include consideration of a more complex outbreak model, consideration of a few more ZIP parameters, and an extensive comparison with the modified Shewhart, CUSUM, and historical limits procedures.

References

- AGREDA, J.M., 2013, May 24, Measles cases soar high by 157%, Retrieved from Sun Star: <http://www.sunstar.com.ph/baguio/local-news/2013/05/24/measles-cases-soar-157-283866>
- ARAGONES, S., 2011, April 5, Over 2000 measles cases in Jan-March, Retrieved from ABS-CBN News: <http://www.abs-cbnnews.com/lifestyle/04/05/11/over-2000-measles-cases-jan-march>
- BURKOM, H.S., MURPHY, S.P., and SHMUELI, G., 2007, Automated time series forecasting for biosurveillance. *Statistics in Medicine* 26 (22): 4202-4218.
- CATINDIG, N. T. and NIÑAL, M. O., 2008, Manual of Procedures for the Philippine Integrated Disease Surveillance and Response. National Epidemiology Center of the Department of Health, Philippines.
- CHEN, N., ZHOU, S., CHANG, T.S., and HUANG, H., 2008, Attribute control charts using generalized zero-inflated Poisson distribution, *Quality and Reliability Engineering International* 24 (7): 793-806.
- DONG, Y., HEDAYAT, A.S., and SINHA, B.K., 2008, Surveillance strategies for detecting changepoint in incidence rate based on exponentially weighted moving average methods, *Journal of the American Statistical Association* 103 (482): 843-853.
- ELBERT, Y. and BURKOM, H.S., 2009, Development and evaluation of a data-adaptive alerting algorithm for univariate temporal biosurveillance data, *Statistics in Medicine* 28(26): 3226-3248.
- FARRINGTON, P. and ANDREWS, N., 2004, Outbreak detection: Application to infectious disease surveillance. In Brookmeyer, R. and Stroup, D.F., (editors) *Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance*, New York: Oxford University Press, Inc. Pp. 203-231.
- FATAHI, A.A., NOOROSSANA, R., DOKOUHAKI, P., and MOGHADDAM, B.F., 2012, Zero-inflated Poisson EWMA control chart for monitoring rare health-related events, *Journal of Mechanics in Medicine and Biology* 12(4): 881-895.
- FRAKER, S.E., WOODALL, W.H., and MOUSAVI, S., 2008, Performance metrics for surveillance schemes, *Quality Engineering* 20(4): 451-464.
- FRICKER, R.D. 2013, *Introduction to Statistical Methods for Biosurveillance: With an Emphasis on Syndromic Surveillance*, New York: Cambridge University Press.
- HAN, S.W., TSUI, K.L., ARIYAJUNYA, B., and KIM, S.B., 2010, A comparison of CUSUM, EWMA, and temporal scan statistics for detection of increases in Poisson rates, *Quality and Reliability Engineering International* 26 (3): 279-289.
- HE, S., HUANG, W., and WOODALL, W.H., 2012, CUSUM charts for monitoring a zero-inflated Poisson process. *Quality and Reliability Engineering International* 28 (2): 181-192.

- HE, S., LI, S., and HE, Z., 2014, A combination of CUSUM charts for monitoring a zero-inflated Poisson process, *Communications in Statistics – Simulation and Computation* 43 (10): 2482-2497.
- MAVROUDIS, E. and NICOLAS, F., 2013, EWMA control charts for monitoring high yield processes, *Communications in Statistics – Theory and Methods* 42 (20): 3639-3654.
- MORTON, A. P., WHITBY, M., MCLAWS, M.L., DOBSON, A., MCELWAIN, S., LOOKE, D., STACKELROTH, J., and SARTOR, A., 2001, The application of statistical process control charts to the detection and monitoring of hospital-acquired infections, *Journal of Quality in Clinical Practice* 21 (4): 112-117.
- NGATCHOU-WANDJI, J. and PARIS, C., 2011, On the zero-inflated count models with application to modelling annual trends in incidences of some occupational allergic diseases in France, *Journal of Data Science* 9 (4): 639-659.
- NOBRE, F.F. and STROUP, D.F., 1994, A monitoring system to detect changes in public health surveillance data, *International Journal of Epidemiology* 23 (2): 408-418.
- ROBERTS, S.W., 1959, Control chart tests based on geometric moving averages, *Technometrics* 42 (1): 97-101.
- RSJ, GMA NEWS, 2013, DOH issues measles alert in NCR as cases rise. Retrieved from GMA Network News: <http://www.gmanetwork.com/news/story/339346/news/metromanila/doh-issues-measles-alert-in-ncr-as-cases-rise>, December 11.
- SANTOS, T. G., 2013, Measles cases up in Metro, says DOH. Retrieved from *The Philippine Daily Inquirer*: <http://newsinfo.inquirer.net/544665/measles-cases-up-in-metro-says-doh>, December 11.
- SHMUELI, G. and BURKOM, H.S., 2010, Statistical challenges facing early outbreak detection in biosurveillance, *Technometrics* 52 (1): 39-51.
- SIM, C.H. and LIM, M.H., 2008, Attribute charts for zero-inflated Poisson processes. *Journal of Communications in Statistics – Simulation and Computation* 37 (7): 1440-1452.
- SONESSON, C. and BOCK, D., 2003, A review and discussion of prospective statistical surveillance in public health, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 166 (1): 5-21.
- SZARKA, J. L., GAN, L., and WOODALL, W., 2011, Comparison of the early aberration reporting system (EARS) W2 methods to an adaptive threshold method, *Statistics in Medicine* 30 (5): 489-504.
- UNKEL, S., FARRINGTON, P.C., GARTHWAITE, P.H., ROBERTSON, C., and ANDREWS, N., 2012, Statistical methods for the prospective detection of infectious disease outbreaks: A review, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 175 (1): 49-82.
- WONG, W.K. and MOORE, A.W., 2006, Classical time-series method for biosurveillance. In M.M. Wagner, A.W. Moore, and R.M. Aryel (Eds.), *Handbook of Biosurveillance*, Burlington: Elsevier Academic Press, Pp.217-234.
- WOODALL, W.H., 2006, The use of control charts in health-care and public-health surveillance, *Journal of Quality Technology* 38 (2): 89-104.
- XIE, M., HE, B., and GOH, T.N. (2001). Zero-inflated Poisson model in statistical process control, *Computational Statistics and Data Analysis* 38 (2): 191-201.
- YEH, A. B., MCGRATH, R.N., SEMBOWER, M.A., and SHEN, Q., 2008, EWMA control charts for monitoring high-yield processes based on nontransformed observation. *International Journal of Production Research* 46 (20): 5679-5699.

Modelling Rice Yield in the Philippines using Dynamic Spatio-Temporal Models

Stephen Jun V. Villejo

School of Statistics

University of the Philippines Diliman

Recent unpredictable and extreme weather episodes and infestation are some of the realistic occurrences of structural change in the agricultural system which produce outliers and extreme values in our data, and consequently pose problems when building statistical models. An estimation procedure which is robust to structural change is therefore necessary. Three spatial-temporal models with varying dynamic characteristics of the parameters are postulated each with a different estimation procedure for the agricultural yield in irrigated areas of the Philippines. One of which is a robust estimation procedure using forward search algorithm with bootstrap in a backfitting algorithm. The other two algorithms also used the backfitting algorithm but infused with the Cochranne-Orcutt procedure. The robust estimation procedure and the other one which considers varying parameter across space gave competitive predictive abilities and are better than the ordinary linear model. Simulation studies show the superiority of the robust estimation procedure over the Cochranne-Orcutt procedure and ordinary linear model in the presence of structural change.

Keywords: Spatial-temporal model; Backfitting algorithm; robust estimation; additive model

1. Introduction

The main stimulus of the development of this spatial-temporal model is in modelling agricultural production which accounts for spatial and temporal dependencies. In an agricultural system, there are dependencies among spatial units and the error term will most likely contain autocorrelation. Classical methods that assume independent observations are unarguably violated in data sets indexed by space and time, hence, classical regression models may yield misleading results.

Spatial and time interaction is very evident in the context of agricultural production (Landagan and Barrios, 2007). Similar crops are planted in the same location due to geographical, climatic, and soil conditions. Adjacent locations are most likely similar or resembles almost the same characteristics in terms of soil quality, availability and quality of irrigation systems, amount of rainfall, terrain and climate or weather condition. In addition to the spatial dependence, there could also exist dependence among values adjacent in time due to seasonal weather patterns. Moreover, yield of crops is affected by physical and geographical conditions, indicator/quantitative variables that represent information within the area with similar conditions, and another term which accounts for production shocks. These components in the model will vary through space and time. Systems with such complexities which exhibit spatial and temporal dependence is best facilitated by dynamic spatio-temporal models. In this way, there can be a better understanding of the dynamic relationships in the system and to give accurate forecasts for policy-making purposes.

Moreover, recent unpredictable and extreme weather episodes and infestation are some of the realistic occurrences of structural change in the system. These episodes are coined as “structural change.” According to Hackl and Westlund (1989), an essential element of “structural change” is nonconstancy of the parameters and could be evident in time-series or cross-sectional data. In other words, there is a change in the system and in the behavior of the random variables which occurs in a certain period of time or in particular neighborhoods. Structural change will be evident in many cases and these occurrences are just temporary since after a certain period of time, the series will go back to its original and regular behavior. However, they have adverse effects for instance in the production of agricultural crops and will be a problem in the estimation of the statistical model. An estimation procedure which is robust to structural change is therefore necessary.

Three estimation procedures are proposed in this study, which are based on three postulated models describing the dynamic structure of the system. One of the proposed estimation procedures is based on the backfitting algorithm imbedded with bootstrap and forward search algorithm for robust estimation. The other two algorithms are modifications of the Landagan and Barrios (2007) model through the assumption of varying covariate, spatial and temporal parameters. Results are compared to the Cochranne-Orcutt procedure and ordinary regression. Simulation studies are also conducted to assess the quality of the proposed model under specified conditions and assumptions.

2. Related Literature

Landagan and Barrios (2007) proposed an estimation procedure of a spatial-temporal model using the backfitting algorithm infused with the Cochran-Orcutt procedure. The main motivation of the paper is in understanding agricultural

production in the Philippines particularly the yield of certain crops. Yield of crops is affected by geophysical conditions like the area harvested, by spatial parameters which gives information on characteristics within areas of similar conditions, and an error term which captures other information and shocks. The following model is postulated:

$$Y_{it} = \beta X_{it} + \delta W_{it} + \varepsilon_{it}, \text{ where } \varepsilon_{it} = \mu_i + v_{it} \quad (1)$$

where Y_{it} is the response variable from location i and time t , X_{it} the set of covariates from location i at time t , W_{it} the set of variables in the neighborhood system of location i at time t , and ε_{it} the error component that would take the disturbances. The error component is postulated to follow an auto-regressive behavior. An important assumption of the model is constant covariate, spatial and temporal effect across space and across time. The idea is to alternately and iteratively estimate the parameters with the covariate and spatial effects first then the temporal effect. The covariate considered in the paper is area harvested. Several neighborhood variables are considered, namely count of adjacent units with yield greater than national yield, average production of neighboring units, average yield of neighboring units, count of units within a region with production greater than the national median production, count of units within a region with yield greater than the national median yield, regional mean production and regional mean yield. The postulated spatial-temporal model is found to be superior to commonly used models like the ordinary linear model, linear model with auto-correlated errors and mixed linear model. Among the seven neighborhood variables, regional average yield gave the smallest mean absolute prediction error. The proposed method by Landagan and Barrios is generally capable of summarizing relationships among observations related in space and time as long as appropriate neighborhood variables are used. The model is also superior over the ordinary linear model, linear model with auto-correlated errors, and mixed linear model.

Dumanjug, Barrios, Lansangan (2010) introduced bootstrap methods into the algorithm of Landagan and Barrios (2007). Two bootstraps were introduced: the first method using the ordinary bootstrap and the second method using time blocks of consecutive observations. The confidence intervals of the covariate and temporal effects were narrower from using the second method. The second method gave more efficient and reliable bootstrap estimates. Furthermore, the simulation study showed that the model gave very low prediction errors.

Generalized additive models

Hastie and Tibshirani (1986) introduced the class of *generalized additive models* which is similar to the normal linear regression model and the linear logistic model where the covariates assume a linear or other parametric form. They also proposed a method for its estimation. The linear form $\sum \beta_j X_j$ is replaced by a sum of smooth functions $\sum s_j X_j$, hence the term *generalized additive models*.

The additive model generalizes the linear regression model and it is therefore assumed that

$$E(Y | X_1, X_2, \dots, X_p) = s_0 + \sum_{j=1}^p s_j(X_j). \quad (2)$$

The $s_j(\cdot)$'s are smooth functions where $E[s_j(X_j)] = 0$.

The $s_j(\cdot)$'s are estimated in an iterative manner using the procedure called the *local scoring algorithm* which uses scatter plot smoothers to generalize the usual Fisher scoring procedure for computing maximum likelihood estimates. The smooth functions produced by the local scoring can be used as a data description, for prediction, or to suggest covariate transformations. Generalized additive models provide a flexible method for identifying nonlinear covariate effects in exponential family models and other likelihood-based regression models.

Backfitting algorithm

Hastie and Tibshirani (1986) introduced the class of generalized additive models which is similar to the normal linear regression model and the linear logistic model where the covariates assume a linear or other parametric form. The linear form $\sum \beta_j X_j$ is replaced by a sum of smooth functions $\sum f_j X_j$, hence the term generalized additive models. The iterative process used by Hastie and Tibshirani to estimate the additive regression model is the backfitting algorithm (Friedman and Stuetzle, 1981). The backfitting algorithm enables the fitting of an additive model where the predictor effects can be examined separately. The estimation can be done using any regression-type mechanism and can be parametric, nonparametric, or semi-parametric in nature. The additive model is defined and characterized as follows:

$$Y = \alpha + \sum_{i=1}^p s_j(X_j) + \varepsilon \quad (3)$$

where the X 's are the predictors, the error term ε is independent of the X 's, $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$. The s_j 's are arbitrary univariate functions,

- i. Initialize: $\alpha = \text{ave}(y_i), f_j = f_j^0, j = 1, \dots, p$
- ii. Cycle: $j = 1, \dots, p$

$$f_j = S_j \left[\left(y - \alpha - \sum_{k \neq j} f_k \right) | x_j \right]$$

- iii. Continue (ii) until the individual functions do not change where S_j denotes a smoothing of the response y against the predictor X_j .

Another way of viewing the process is given by Hastie and Tibshirani (1986).

Suppose the model is given by $Y = s_0 + \sum_{i=1}^p s_j(X_j) + \varepsilon$ and the partial

residual is defined by $R_j = Y - s_0 - \sum_{k \neq j} s_k(X_k)$, then $E[R_j | X_j] = s_j(X_j)$ and minimizes $E\left[Y - s_0 - \sum_{k=1}^p s_k(X_k)\right]^2$. This provides a way for estimating each $\hat{s}_i(\cdot)$ given estimates $\{\hat{s}_i(\cdot), i \neq j\}$.

The algorithm starts with the initialization:

$$s_0 = E(Y), s_1^1(\cdot) = s_2^2(\cdot) = \dots = s_p^p(\cdot) = 0, m = 0.$$

Then iterate: $m = m + 1$ for $j = 1$ to p and do:

$$R_j = Y - s_0 - \sum_{k=1}^{j-1} s_k^m(X_k) - \sum_{k=j+1}^p s_k^{n-1}(X_k)$$

$$s_j^m(X_j) = E(R_j | X_j).$$

The iteration will continue until:

$$RSS = E\left[Y - s_0 - \sum_{j=1}^p s_j^m(X_j)\right]^2 \text{ fails to decrease.}$$

The RSS does not increase at any step of the algorithm which assures convergence.

Buja et al. (1989) proved convergence and consistency of the backfitting in an additive model with a general smoother. Chen and Tsay (1993) has shown that the backfitting procedure will work in the time-series context if the dependence structure is not quite strong. The convergence and asymptotic property of the estimators using backfitting algorithm are studied by Opsomer (2000).

Forward search algorithm

The forward search algorithm is a powerful method to detect multiple masked outliers, to determine their effects on models fitted to the data, and to detect systematic model inadequacy. (Atkinson and Riani, 2007). The algorithm starts by choosing an initial outlier-free subset of m observations from the n observations and obtaining robust estimates from the subset. The algorithm will move forward to a larger subset by getting the squared residuals from the n observations using the least squares fit to the subset of m observations and using the $m + 1$ observations as the new larger subset. Usually, only one observation is added at each step, but sometimes two or more is added in instances when at least one leave which is often and indication of the introduction of some of a cluster of outliers. In the algorithm, a series of parameter estimates will be obtained starting from very robust estimates to least squares at the end. The observations which

are far from the fitted model which are usually outliers, unidentified subset, or an indication of systematic failure of the model, will enter at the end of the search. Quantities indicative of model quality and inadequacy will be monitored.

Bootstrap

The bootstrap method is a general method used to estimate the sampling distribution of some random variable on the basis of some observed data (Efron, 1979). Let $X = (X_1, X_2,..., X_n)$ denote the random sample from a completely unspecified probability distribution F and $x = (x_1, x_2,..., x_n)$ be the realization of the random variable. The bootstrap method is accomplished by drawing a random sample with replacement of size n wherein all the $x_1, x_2,..., x_n$ have the same chances of selection. The bootstrap distribution can be accomplished by direct theoretical calculation, Monte Carlo approximation, or Taylor Series expansion which has similar form as the Jackknife method.

3. The Spatial-Temporal Models

Similar to Landagan and Barrios (2007), this paper is motivated to characterize and understand the dynamics of agricultural production. Adjacent locations or spatial units in the same neighborhood are homogenous with respect to soil quality, availability of irrigation systems, amount of rainfall, terrain, and climate or weather conditions. Yield for certain crops can be affected by factors such as physical and geophysical conditions, indicator variables which represent information within the area with similar conditions, and an error term that can account for production shocks such as irregular weather, infestation, etc.

Three models are postulated as follows:

Model 1: $Y_{it} = \beta_t X_{it} + \delta_i W_{it} + \epsilon_{it}$ (4)

Model 2: $Y_{it} = \beta_t X_{it} + \delta_i W_{it} + \epsilon_{it}$ (5)

Model 3 : $Y_{it} = \beta_t X_{it} + \delta_i W_{it} + \epsilon_{it}$ (6)

where Y_{it} is the response variable from location i at time t , X_{it} the set of covariates from location i at time t , W_{it} the set of variables in the neighborhood system of location i at time t , and ϵ_{it} the error component that would take the disturbances. In the spatial-temporal model explored in this paper, the response variable Y_{it} is the yield of irrigated rice at province i and time t , the covariate X_{it} is the area harvest in province i at time t , the only spatial variable explored is the regional mean yield which gave a competitive predictive performance for lag 1 models according to Landagan and Barrios (2007). The error term is assumed to follow some autoregressive behavior and is postulated as

$$\epsilon_{it} = \rho_i \epsilon_{i,t-1} + u_i \tag{7}$$

The disturbances follow a one-way error component with $u_i \sim IID(0, \sigma_u^2)$ and the remaining disturbances following AR(1).

For the first postulated model in (4), the covariate (β_i) effect and spatial effect (δ_i) varies across time. For the second postulated model in (5), both the covariate and spatial effect varies across locations. Lastly, for the third postulated model in (6), the covariate effect varies across space while the spatial effect varies across time. For all the three postulated models, the temporal effect varies across space.

4. Estimation of the Model

Algorithm 1: Estimation of model 1

The backfitting algorithm is used to estimate the parameters of the spatial-temporal model where the parameters of the model will be estimated iteratively. Forward search algorithm and bootstrap strategy is incorporated in the backfitting procedure to give robust estimates. The algorithm is given in the following steps:

Step 1. Using the realizations $\{y_{it}\}$, $i=1, \dots, N$, $t=1, \dots, T$, where i is the index for location and t the index for time, β_i and δ_i will be simultaneously estimated using ordinary least squares regression for each of the T time points. Forward search algorithm will be incorporated in each of the ordinary least squares estimation through the following steps:

- i. A subset of size n , $n < N$ from N observations for each time point will be chosen. The n observations are ideal and outlier-free. The ordinary least squares will be fitted on the full data set. The choice of the n observations corresponds to the smallest n residuals.
- ii. Fit the model $Y_{it} = \beta_i X_{it} + \delta_i W_{it} + \epsilon_{it}$ to the selected n observations and generate the parameter estimates $\hat{\beta}_i$ and $\hat{\delta}_i$.
- iii. The fitted values will be computed to the $N-n$ left-out observations and the residuals will be obtained.
- iv. The observation corresponding to the smallest residual from the $N-n$ residuals will be included in the subset of observations from (i).
- v. Fit the model $Y_{it} = \beta_i X_{it} + \delta_i W_{it} + \epsilon_{it}$ on the $n+1$ observations in (iv).
- vi. The previous procedure will be repeated iteratively adding one observation at a time until all N locations have been included in the model or until the model is behaving wildly based on the Cook's D.

The Cook's D is said to be influential if its value exceeds $\frac{4}{n}$ where n is the number of observations. The algorithm then stops if the Cook's D is no longer influential to the model based on this threshold.

The forward search is used to obtain robust estimates of the spatial and covariate parameters. Observations which behaves wildly based on the postulated model will not be included in the estimation of the parameters.

Step 2. After obtaining the final subset from the forward search algorithm for each time point, the following is performed for each time point:

- i. Draw a simple random sample of size m with replacement.
- ii. Fit the model $Y_{it} = \beta_t X_{it} + \delta_t W_{it} + \epsilon_{it}$. This gives a value of the statistic computed using the bootstrap sample in (i).
- iii. Repeat (i) and (ii) B times, yielding $B \hat{\beta}_t$ and $\hat{\delta}_t$. The Monte Carlo estimates for the mean and standard deviation of δ_t will be obtained.

$$\hat{\delta}_t^* = \frac{1}{B} \sum_{k=1}^B \hat{\delta}_{tk}^* \quad (8)$$

$$S.E.(\hat{\delta}_t^*) = \frac{1}{B-1} \sum_{k=1}^B (\hat{\delta}_{tk}^* - \hat{\delta}_t^*)^2 \quad (9)$$

$$\hat{\beta}_t^* = \frac{1}{B} \sum_{k=1}^B \hat{\beta}_{tk}^* \quad (10)$$

$$S.E.(\hat{\beta}_t^*) = \frac{1}{B-1} \sum_{k=1}^B (\hat{\beta}_{tk}^* - \hat{\beta}_t^*)^2 \quad (11)$$

From the notation, $\hat{\delta}_{tk}^*$ is the k th bootstrap estimate of δ_t while $\hat{\beta}_{tk}^*$ is the k th bootstrap estimate of β_t .

From this step, the set of parameter estimates $\{\hat{\delta}_1^*, \hat{\delta}_2^*, \dots, \hat{\delta}_T^*, \hat{\beta}_1^*, \hat{\beta}_2^*, \dots, \hat{\beta}_T^*\}$ with their standard errors is obtained. Optimality is achieved since the covariate and spatial parameter are simultaneously estimated for each time point.

Step 3. For each Y_{it} , generate the residuals $e_{it} = Y_{it} - \hat{\beta}_t^* X_{it} - \hat{\delta}_t^* W_{it}$. The computed residuals will contain information on the true error and the temporal parameter. For each location, AR(1) estimation will be performed to get the temporal parameter estimates.

Step 4. In this step, we have the set of parameter estimates $\{\hat{\delta}_1^*, \hat{\delta}_2^*, \dots, \hat{\delta}_T^*, \hat{\beta}_1^*, \hat{\beta}_2^*, \dots, \hat{\beta}_T^*, \hat{\rho}_1^*, \hat{\rho}_2^*, \dots, \hat{\rho}_N^*\}$. A new dependent variable will then be computed adjusting for the temporal component, given by $Y_{it}^{new} = Y_{it} - \hat{\rho}_t \epsilon_{i,t-1}$.

After Step 4, the algorithm will go back to *Step 1* using the new values of the dependent variable. After removing the temporal effect, the focus will be on the covariate-spatial effect when the ordinary least squares is once again fitted in *Step 1*. In updating the estimates of the error term in *Step 3*, the original values

of Y will be used. After updating the error terms and performing *Step 3*, the new dependent variable will once again be computed using the original values of the dependent variable and the updated estimates of the error terms from *Step 3*. The iteration converges when there are minimal changes in the values of the parameter estimates.

An alternative model postulate is assuming constant covariate effect across time points. As an example, if the X_{it} is the area harvested at time t , this means that if area harvested is high at $t = 0$, the area harvested will remain to be high at $t = 1, 2, \dots, T$. In this framework, the estimation procedure will change particularly in the estimation of the constant covariate effect β . The Monte Carlo Estimate of β will be computed as follows:

$$\hat{\beta}^* = \frac{1}{BT} \sum_{i=1}^T \sum_{k=1}^B \widehat{\beta}_{ik} \quad (12)$$

The index i is for the time point while the index k is for the bootstrap sample.

Thus, $\widehat{\beta}_{ik}$ is the k^{th} bootstrap estimate of β from time point i . The standard error of the estimate is then computed as follows:

$$S.E.(\hat{\beta}^*) = \frac{1}{BT-1} \sum_{i=1}^T \sum_{k=1}^B \left(\widehat{\beta}_{ik} - \hat{\beta}^* \right)^2 \quad (13)$$

Algorithm 2: Estimation of model 2

For the postulated model (2), a modification of Landagan and Barrios (2007) is implemented. The estimation is as follows:

Step 1: For each spatial unit, the parameters β_i and ρ_i are estimated simultaneously using the Cocchrane-Orcutt procedure. The residuals, which contains information on the spatial component, will be generated.

Step 2: Another regression will be performed on the residuals from Step 1. Since the spatial effect is assumed to be varying across locations, the estimation of the spatial effect δ_i will also use the regression model with autocorrelated errors using the Cochranne-orcutt procedure. The estimates for the ρ will be disregarded in this step. This will be performed per spatial unit.

Step 3: A new dependent variable will be computed adjusting for the spatial component, given by $Y_{it}^{new} = Y_{it} - \hat{\delta}_i W_{it}$. Another regression model with autocorrelated errors will be fitted in Step 1. The iteration converges when there are minimal changes in the parameter estimates. When the backfitting algorithm have converged, another regression model will be fitted but this time all the parameter will be estimated simultaneously.

Algorithm 3: Estimation of model 3

The estimation of model (3) is basically similar to the estimation procedure proposed by Landagan and Barrios (2007) but with varying parameters. That is, in every step of the estimation procedure the estimate for the parameters will not be averaged.

5. Results and Discussion

In modelling yield of rice in irrigated areas in the Philippines, the covariate considered is the total area harvested, while the neighborhood variable is the regional mean yield. The total area harvested is shown to be a good proximate indicator of agricultural efficiency under specified conditions (Landagan and Barrios, 2007). The spatial variable considered is the regional mean yield which showed to have a competitive mean absolute percentage error from the model using the proposed estimation procedure from the same paper. Logarithmic transformations were made on the variables for modelling purposes. Autoregressive structure of order 1 was explored. The data used is a quarterly data from 1994 to 2014 second quarter with 76 provinces and 16 regions. Thus, the length of the time series is 82 with 76 spatial units. In such a long series, autocorrelation and spatial correlation would most likely be evident.

Other than the 3 proposed algorithms, other estimation procedures were run for comparison purposes, wherein most of them are just variants and modifications of the proposed algorithm in the estimation of model 1. The modifications are the following: constant covariate effect; without bootstrap; forward search algorithm only, no backfitting, without bootstrap, but all parameters will be estimated; and forward search only, no bootstrap, no backfitting, no auto-regression. The other models are ordinary linear model; and regression with auto-correlated errors estimated using the Cochran-Orcutt procedure.

In the Landagan and Barrios (2007) proposed algorithm, changes in the parameter estimates for the backfitting algorithm usually becomes negligible after second or third iteration. In this paper's proposed estimation procedure on the postulated model 1 with varying parameters, the convergence would take relatively longer. For the case when the covariate effect is assumed constant, changes in the parameter estimates would stabilize after 4 or 5 iterations. For the cases when all the parameters are assumed to be varying, the backfitting algorithm would converge longer than usual. The changes in the estimates of the spatial effect and temporal effect would be negligible after 4 iterations and 5 iterations respectively, but not for the estimate of the covariate effect whose parameter changes in the iterations are unstable. The change in the estimate of the covariate effect would be acceptable after 15 iterations. For the estimation of model 2, the estimation converged after three iterations. The third proposed algorithms took 8 iterations before convergence.

Parameter estimates

Model 1: β and δ varies across time

Since total area harvested is a proximate indicator of agricultural efficiency, the parameter β_i accounting for the covariate effect is expected to be positive. From the models estimated, the models using the proposed algorithm with varying parameters had 2 negative β_i estimates while the model using the proposed algorithm without bootstrap had 1 negative β_i estimate. For all the other models which are modifications of the proposed algorithm, the covariate effect estimates for all time points are positive. Despite some anomalies discovered for 2 out of 82 time points, it can still be concluded that area is a proximate indicator of agricultural efficiency. As harvested area increases, efficiency in the use of farm technologies and other inputs can be maximized leading to higher productivity.

Even when the spatial and temporal parameters are introduced, the proposed estimation procedure did not lead to major adjustments in the signs of the β_i estimates. In the proposed estimation procedure of Landagan and Barrios (2007), the estimate of the constant covariate effect β is negative. It is hypothesized that the simultaneous estimation of β and ρ (the constant temporal effect) caused the automatic adjustments in the parameter resulting to the negative sign of β . In this paper's proposed algorithm, the varying covariate effect and spatial effect are simultaneously estimated first before estimating the temporal effect. The different order in the estimation may have avoided the shift in the signs of the covariate and spatial parameter estimates.

The spatial parameter γ_i estimate is also expected to be positive. In all the models estimated, the sign of the effect of γ_i for all time points is consistently positive. Also, from Landagan and Barrios (2007), this is consistently estimated to be positive for both cases when only the spatial component is considered and when both spatial and temporal components are considered.

Model 2: β and δ varies across space

The parameter estimates obtained from the second proposed estimation procedure is almost similar to those obtained from the usual Cochrane-Orcutt procedure. The effect of the spatial variables are positive for all locations. The effect of the covariate is not consistently positive for all spatial units. Nonetheless, majority of the spatial units has a positive estimate of the covariate effect.

Model 3: β varies across space and δ varies across time

The estimates of the covariate effects obtained from implementing the third proposed estimation procedure are positive for all spatial units. The estimates for the spatial parameters for all time points are also positive. The result obtained from here is similar with model (1).

Predictive ability of models

The proposed algorithm 1 will be superior over the other models when the assumption of varying covariate and spatial effect across time and varying temporal effect across locations is satisfied. Moreover, the advantage of the algorithm which uses robust estimation procedures will be evident under structural change. Below are some historical plots of the covariate and neighborhood variable for four sample provinces. Figure 1 shows the historical plot of regional yield for four provinces. For the earlier time points, there is a possible occurrence of structural change in which province _32 behaved differently from the other 3 provinces and hence could be a possible outlier in the dataset. There is also a possible occurrence of structural change in the middle of the series in which the same province has the different behavior. In Figure 2, a historical plot on area harvested for another set of 4 provinces is shown. It is very noticeable that province _30 has a different behavior from the rest somewhere in the earlier time point. This outlying behavior is most likely influential and has a serious effect in the estimation procedure and modelling stage. This possible structural change is crucial and are best facilitated by robust estimation procedures such as the proposed algorithm.

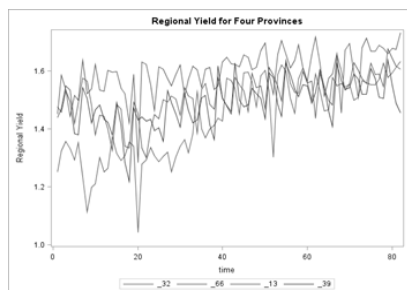


Figure 1. Historical Time Plot of Regional Mean Yield for Four Provinces in the Philippines

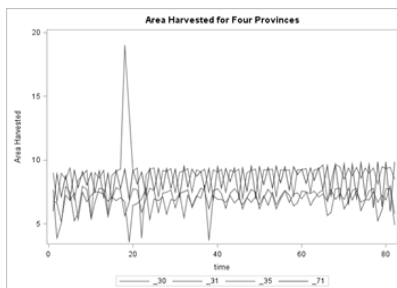


Figure 2. Historical Time Plot of Area Harvested for Four Provinces in the Philippines

The predictive power of all the explored estimation procedures are assessed using the in-sample mean absolute prediction error (MAPE). The MAPE for the different models are presented in Table 1. The proposed algorithm assuming constant covariate effect gave a MAPE of 7% which is relatively higher than the other estimation procedures. This is an evidence that the covariate effect is not constant, and would vary across time or across space. There is an improvement in the predictive ability of the model from the ordinary linear model to the proposed algorithm assuming varying parameters. Ordinary linear model gave a MAPE of 5.8789%. After introducing the forward search algorithm, there is 0.23% reduction

in the error over ordinary regression model. Since forward search algorithm is a robust procedure, it is expected to give better prediction ability especially when there are outliers in the dataset. A slight improvement in the MAPE could be an evidence that there are no extremely influential values in the yield for all spatial units. This could also be an effect of the logarithmic transformation done on the variables. Furthermore, estimating the temporal effects through auto-regression also lead to a reduction in the predictive error from 5.87% to 5.32%. Temporal components are important for the agricultural system and including them in the model will make the predictions more accurate. For the procedure using the proposed algorithm but without bootstrap, the MAPE is 5.29%. There is a slight improvement from the previous model which performs forward search and autoregression only. The backfitting algorithm eliminates some of the biases in the estimation which lead to more accurate estimation of the parameters. This could be the reason for the improvement of the predictive ability of the model. The proposed algorithm 1 gave the most competitive predictive ability among all the procedures from the modifications of the proposed algorithm. This is contributed by the robust procedure through the forward search, the bootstrap, and backfitting algorithm.

Table 1. Mean Absolute Percentage Errors (MAPE) for different models

Estimation Procedure	MAPE
Proposed Algorithm 1: β and δ varies across time	5.2914
Proposed Algorithm 1 but β is constant	6.9995
Proposed Algorithm 1 but without Bootstrap	5.2926
Forward Search only, No Bootstrap, No Backfitting	5.3155
Forward Search only, No Autoregression, No Backfitting	5.6422
Ordinary Least Squares: β and δ varies across time	5.8769
Cochranne-Orcutt Procedure: β and δ varies across space	3.8158
Proposed Algorithm 2: β and δ varies across space	3.8597
Proposed Algorithm 3: β varies across space and δ varies across time	9.8148

The MAPE obtained from the third proposed algorithm gave the poorest MAPE of 9.82%. The assumptions from the postulated model 3 therefore does not hold on the data resulting to inferior predictive ability. The MAPE obtained from the second proposed algorithm gave the smallest from the 3 proposed estimation procedures. Among the three postulated models, the assumptions for the second model best characterizes the spatio-temporal relationship of interest. The obtained MAPE 3.86% is slightly larger by 0.04% than the one obtained from the fitted regression model with autocorrelated errors. The MAPE of the models are comparable. This could be an evidence that the spatio-temporal model of yield

in the Philippines is more represented by varying parameters across locations and space.

Occurrence of structural change is very likely to be present in the dataset. However, algorithm 1 which was expected to give the best model fit and smallest prediction error did not give the highest predictive ability and is inferior over algorithm 2 and the Cochran-Orcutt procedure. A strong reason for this is that the assumptions of the proposed algorithm do not hold for the data. Nonetheless, the fit of algorithm 1 still gives a competitive predictive performance. For datasets which will follow the postulated model 1 and in the presence of structural change or contamination, algorithm 1 will be superior and will give the best predictive ability. algorithm 2 will give the smallest MAPE for cases when the covariate effect and spatial effect varies across locations while algorithm 3 will have a comparative advantage when the covariate effect varies across locations and the spatial effect varies across time. However, it should be noted that a consequence of doing robust estimation through forward search algorithm is that not all points will be included in the computation of the estimates which might have negative consequences on the predictions especially on points which are considered outlier or extreme values.

6. Simulation Study

The first postulated model along with the estimated procedure will be evaluated using a simulated data in the balanced ($N = T$) and unbalanced ($N > T$) scenarios. The true model is given below:

$$Y_{it} = \beta X_{it} + \delta_i W_{it} + \epsilon_{it} \quad (13)$$

$$\epsilon_{it} = \rho_i \epsilon_{it-1} + a_{it}, \quad |\rho_i| < 1, \quad a_{it} \sim IID(0, \sigma^2 a) \quad (14)$$

The covariate X_{it} was generated from Normal (100, 10) while the spatial variable was generated from the Poisson distribution with varying parameters. There are two scenarios on the number of clusters or neighborhoods: 5 and 10. In the case where there are 5 clusters, the parameters of the Poisson distribution are 10, 20, 30, 40 and 50. In the case where there are 10 clusters, the parameters are 10, 20, 30, ..., 80, 90, 100. For the balanced data case, the number of time points and spatial units is 20. For the unbalanced case, there will be 30 time points and 50 spatial units. The simulation study aims to recreate the reality of the agricultural behavior with structural change. The structural change is nested on the following features: contamination in the recent period of the time series vs. contamination in the middle of the time series, structural change in all clusters in the time points affected by the contamination versus structural change in 20% of the clusters (if 5 clusters, only 1 is affected; if 10 clusters, 2 clusters are affected), and the change in the parameter in the contaminated spatial units to be 10% and 20%. The number of clusters will show the performance of the estimation procedure when the

population is divided to small number of neighborhood systems versus otherwise. The scope of contamination will be facilitated by making all clusters affected by the contamination versus having only a subset of clusters contaminated. The setting where only 20% of the clusters is contaminated may be viewed as the case wherein the structural change occurred within specific locations. The presence of the contamination is either at the middle time points or at the recent time points of the series. The percentage of time points affected by the contamination is set to be 10%.

Simulation results

Table 2 shows the in-sample MAPE comparing the proposed algorithm, Cochranne-Orcutt procedure, and ordinary linear model under different scenarios.

Table 2. Table on the MAPEs for the different simulation scenarios

					Contamination in Recent Periods of the Time Series			Contamination in Middle Periods of the Time Series		
					MAPE					
	T	N	Cluster	Cont	Proposed Algo- rithm	Co- chranne Orcutt	OLS	Proposed Algo- rithm	Co- chranne Orcutt	OLS
Contamination in all clusters	20	20	5	0.1	4.91	26.35	26.61	4.95	26.66	26.98
	20	20	5	0.2	5.18	26.67	26.9	5.18	27.09	27.45
	20	20	10	0.1	4.93	26.54	26.54	4.98	26.85	27.19
	20	20	10	0.2	5.21	26.88	26.88	5.22	27.3	27.67
	30	50	5	0.1	4.72	29.21	29.21	4.75	29.23	29.66
	30	50	5	0.2	4.99	29.75	29.87	5.04	29.48	29.41
	30	50	10	0.1	4.75	29.25	29.41	4.78	29.27	29.73
	30	50	10	0.2	5.02	29.8	29.95	5.07	29.54	29.47
Contamination in 20% of clusters	20	20	5	0.1	4.94	26.38	26.74	4.93	26.41	26.77
	20	20	5	0.2	5.17	26.4	26.77	5.19	26.47	26.73
	20	20	10	0.1	4.93	26.57	26.94	4.95	26.6	26.83
	20	20	10	0.2	5.19	26.6	26.98	5.15	26.66	26.97
	30	50	5	0.1	4.74	28.98	29.20	4.78	29.01	27.04
	30	50	5	0.2	4.99	29.06	29.27	5.01	29.08	29.20
	30	50	10	0.1	4.74	29.02	29.27	4.78	29.05	29.27
	30	50	10	0.2	4.98	29.1	29.35	5	29.12	29.35

The proposed algorithm shows superior predictive ability under the different scenarios considered. Therefore, if the assumption of varying (or constant) covariate effect and spatial effects across locations is satisfied, the proposed algorithm is superior over the commonly used estimation procedures. Another advantage of the estimation procedure is its robustness in the presence of structural change.

7. Conclusions

Modelling yield of rice in irrigated areas of the Philippines is best represented by a model with varying covariate, spatial, and temporal effects across locations and with an estimation procedure which will consider that dynamic aspect of the system. That is the reason why the proposed algorithm 2 gave the best predictive performance among the 3 proposed algorithms. Moreover, the advantage of optimality from the simultaneous estimation of the covariate and spatial parameter also led to better forecasts. From the modifications of the proposed algorithm 1, the contribution of the robust estimation through forward search and bootstrap is evident in the reduction of the prediction error. Even if the reduction in the prediction error seems to be small, this could be an effect of the logarithmic transformations done on the dataset. The proposed algorithms 1 and 2 are generally capable of summarizing relationships in a spatio-temporal model. Since the neighborhood variable used in the models is regional mean yield which is selected based on the results from Landagan and Barrios (2007), the consistency of the quality of results from the estimation procedures could be affected by the appropriateness of the neighborhood indicators used.

The prediction performance of the estimation procedure will depend on the assumptions that will hold in the dataset. Based on the simulation study, the results of estimating the postulated spatio-temporal model A using the proposed algorithm 1, is superior compared to the ordinary linear model estimated using ordinary least squares and regression model with autocorrelated errors using Cochranne-Orcutt procedure. Extending the simulation studies to more settings and scenarios can be explored. Also, a more extensive simulation study of the two proposed algorithms will be the next step.

References

- LANDAGAN, O. and BARRIOS, E., 2007, An estimation procedure for spatio-temporal model, *Statistics and Probability Letters* 77: 401-406.
- DUMANJUG, C., BARRIOS, E. and LANSANGAN, J., 2010, Bootstrap procedures in spatiotemporal model, *Journal of Statistical Computation and Simulation* 80: 809-822.
- HASTIE, T. and TIBSHIRANI, R. (1986). Generalized additive models, *Statistical Science* 1(3): 297-318.
- FRIEDMAN, J. H. and STUETTZE, W., 1981, Project pursuit regression, *Journal of the American Statistical Association* 76: 817-823.
- ATKINSON, A. and RIANI, M., 2002, Forward search added-variable t-tests and the effect of masked outliers on model selection, *Biometrika* 89: 939-946.
- EFRON, B., 1979, Bootstrap methods: Another look at the jackknife, *The Annals of Statistics* 7(1):1-26.
- HACKL, P. and WESTLUND A. H., 1989, Statistical analysis of "structural change": An annotated bibliography, *Empirical Economics* 14(2): 167-192

Statistical Evaluation of In Vivo Micronucleus Assays in Toxicology

John Closter F. Olivo
Central Luzon State University

Several alternative statistical procedures have been suggested and published to statistically analyze the incidence of micronucleated polychromatic erythrocytes (MNPCs) among treatment groups, but no standard procedure has been singled out and exclusively recommended. In this study, the potential of T02 to induce chromosomal damage is tested using both Poisson and quasi-Poisson models for the statistical evaluation of *in vivo* micronucleus (MN) assay. The genotoxic activity of T02 is assessed in the rodent bone marrow micronucleus test using male mice. Results show that MN frequencies are significantly elevated in mice exposed to any dose level of T02 administered orally in a single frequency of dose. Moreover, results indicate that T02 is tested to be a positive compound under the anticipated condition of the tests used.

Keywords: in vivo, micronucleus, MNPC, Poisson model, quasi-Poisson model, T02

1. Introduction

Genotoxicity testing is a vital element of a product safety assessment which is highly recommended by the regulatory agencies around the globe. It is designed to assess and detect chemicals that induce genetic damage. Thus, the Fourth International Conference on Harmonization (ICH4) of Genotoxicity Guidelines recommended the micronucleus assay (MNC) as one of the standard three-test batteries for genotoxicity testing of pharmaceuticals. The two most well-established tests as screening methods of new chemical entities with a widespread acceptance in industry and authorities are the *in vivo* and *in vitro* micronucleus assays.

As a measure for chromosomal aberrations *in vivo*, data on the frequency of micronucleated erythrocytes (MN) per a certain number of scored polychromatic erythrocytes (PCE) per animal are usually determined and analyzed using an

appropriate statistical test. The attention of this paper was restricted only on the statistical procedure for the analysis of micronuclei particularly on the application of methodology suggested by Hothorn and Gerhard (2008).

HG considered different aspects in the statistical evaluation of the *in vivo* micronucleus assay. One aspect concerns the choice of the experimental unit and the potential presence of overdispersion often ignored in traditional analyses. In *in vivo*, biologically speaking, the standard protocol design is basically a randomized one-way layout including a negative control, several doses or treatment groups and, optionally, a positive control. Five to ten animals are randomly assigned to each treatment group. For each animal, a number of polychromatic erythrocytes (PCEs) are scored and evaluated for the presence of small micronuclei.

As proposed by Kim et al. (2000), a formal statistical approach is to assume a binomial distribution of the number of micronuclei observed in 2,000 PCEs. This means pooling the number of MN over the animals to have one estimate of the proportion for each treatment group. These proportions are evaluated using a Dunnett-type procedure or the Cochran-Armitage trend test. However, such approach is not recommended by HG because pooling the number of MN over all animals/cultures ignores the variability between the animals as experimental units. They propose to model this between-animal variability using extra-Poisson or extra-binomial models.

The second statistical aspect is the contradiction between statistical significance and biological relevance. Traditionally, scientists rely on the p-values from their reporting systems to draw conclusions regarding the safety of a compound. P-values represent a probability of falsification but do not provide interpretation in terms of biological relevance. HG suggest the use of confidence intervals which allow interpretations of both statistical significance and biological relevance at the same time.

The third aspect is the type of inference. HG dwell upon the type of inference that is of relevance in trying to identify an increasing dose-related trend, possibly with downturn effects at high doses (Bretz and Hothorn, 2002).

Finally, HG propose a proof of safety approach for a possible claim that a compound is not genotoxic.

The purpose of this study is to implement, apply and discuss the recommendations of HG using the genotoxicity database of the test compound T02. The code name T02 is used in this study for confidentiality purposes. Furthermore, the results are compared with the widespread “traditional” ANOVA-like approaches that completely ignore the between-animal variability, focus on p-values and linear trend tests only, and are set up in terms of proof of hazard.

2. Motivating Example

The data considered are from an independent *in vivo* assay with compound T02. Five male mice are randomly allocated in each different dose levels of

testing groups – low, medium, high and a concurrent negative control/vehicle control (VC) group. In total, 20 male mice were sacrificed. Each mouse in the vehicle control is given demineralised water while each mouse in the remaining groups is treated with a certain dose of T02. Next, blood samples are obtained for each mouse based on the harvesting time (27h) after dosing which is specified in the protocol. Micronucleus frequencies are determined for each animal by scoring 20,000 polychromatic erythrocytes (PCEs) and the micronucleus occurrence per 20,000 PCEs was recorded.

3. Statistical Methodology

The number of micronucleated erythrocytes (MN) constitute the primary endpoint. The micronucleus frequencies are determined by analyzing the number of micronuclei from at least 20,000 PCEs per animal.

3.1 Poisson model

The statistical evaluation of the number of micronuclei is primarily focused on multiple contrast tests for comparisons versus the vehicle control. This number of micronuclei as counts are analyzed using the classical approach Poisson model. This means that for each treatment group only one count is estimated by pooling the number of MN over animals. According to Parodi and Bottarelli (2006), the Poisson regression model is often applied to study the occurrence of small number of counts or events as a function of a set of predictor variables.

Let Y_{ij} be the number of micronuclei observed in the i^{th} animal at dose d_j , $j = 0, 1, 2, \dots, m$. If Y_{ij} can be assumed to be distributed independently with

$$Y_{ij} \sim \text{Poisson}(\mu(x_j)) \quad (1)$$

then, the log-linear Poisson regression can be applied such that

$$\mu(x_j) = \exp(\alpha + \beta^T x_j) \quad (2)$$

where α is the log of mean of the reference group, i.e., vehicle control, β^T is a vector of the parameter estimates and x_j is a vector of the covariates. In this case, the covariate is the indicator variable for the dose group.

Accordingly, to observe any significant differences amongst the dosage sets in assessing the genotoxicity effects, the Dunnett's-type procedure was used to determine if any differed from the vehicle control.

The key assumption of this model is that mean and variance are equal, i.e., $V(\mu) = \mu$. This approach, however, does not recognize animals as experimental units and is therefore not recommended since it results in too liberal decisions (HG, 2009).

3.2 Quasi-Poisson model

HG highlight the importance of taking into account the between-animal variability since the individual animal is the experimental unit. Without accounting for extra-variability, the simple Poisson approach becomes more liberal with increasing overdispersion. Its implications include underestimation of standard errors and thus, wrongly inflating the level of significance (Lee et al., 2012). They recommend fitting a quasi-Poisson model for counts. A quasi-Poisson model is a type of generalized linear models where instead of the maximization of the Poisson likelihood, a more relaxed relationship of the mean-to-variance dependency is assumed. Ramon et al. (2002) suggested that, with generalized linear models, overdispersion can be accounted for by fitting the Poisson model but adjust the standard errors and test statistics. This can be done by introducing an additional parameter to the model, representing the deviation from the Poisson variance assumption. To account for extra variation, the quasi-likelihood methods can be fitted under the assumption $V(\mu) = \rho\mu$ where a value of ρ larger than one indicates overdispersion (Hothorn and Gerhard, 2009).

3.3 The trend test

Traditionally, to evaluate the number of micronuclei, it is suggested by Margolin and Risko (1988), cited by Hayashi et al. (1989), to use the Cochran-Armitage (CA) trend test (Cochran, 1954; Armitage, 1955) to verify the dose-response trends of MNPCEs. This tests the null hypothesis of no trend, i.e., the number of MNPCEs is the same for all dose levels versus the alternative that there is a linear trend across increasing levels of dosage. However, the Cochran-Armitage trend test should not be recommended because it is underpowered when the true trend is not linear, i.e., when the dose and response exhibit a convex and concave relationship (Bretz and Hothorn, 2002). Moreover, this should not be recommended because of its ignorance of the between-animal variability (HG, 2009). For this reason HG propose to use a Williams-type procedure for trend test over Cochran-Armitage (CA) trend test. The CA trend test is particularly sensitive for near-linear shape, whereas, the Williams trend test is sensitive to several shapes. This procedure tests the null hypothesis of no difference among the counts against the alternative that the counts are increasing with increasing dosage compared to the control group. To achieve this comparison, higher concentration groups are successively pooled and compared to the control (Herberich and Hothorn, 2012).

3.4 Proof of hazard vs. proof of safety

The purpose of toxicology testing is to assess the safety of a new test substance relative to a control whether it is harmless up to a specified dose, or harmful. Based on this aim, statistical test of the classical null hypothesis of no difference are usually performed. Failing to reject the hypothesis, i.e., either a nonsignificant p-value or when the point-one hypothesized value is greater than

the lower bound confidence limit of relative risk, often leads to the conclusion that the compound has no harmful effect. This is the traditionally used criterion for harmlessness which demonstrates proof of hazard. The major drawback of this approach is the fact that what is controlled by a pre-specified level is the probability of erroneously concluding hazard. In fact the primary control of the false decision rate, i.e., confidence in negative results, should be preferred in toxicology (Herberich and Hothorn, 2012). Summing up, proof of hazard is an indirect approach and often leads to the problem that statistical significance does not necessarily mean toxicological relevance and statistical nonsignificance does not necessarily mean toxicological irrelevance (Hauschke et al., 1999). In short, be confident with negative results (Kirkland, 2000). Thus, the so-called proof of hazard is inappropriate simply because absence of evidence is not evidence of absence (Altman and Bland, 1995). For this reason, HG recommend proof of safety to demonstrate harmlessness where the probability of erroneously concluding safety is directly controlled. The differences to the proof of hazard with the control of the familywise error rate are (1) the estimation of the upper confidence limits instead of the lower limits, and (2) interpreting interval inclusion instead of superiority interpretation by means of point estimator and confidence limit (Hothorn and Gerhard, 2009).

In this study, for proof of hazard, the common decision is to conclude harmlessness if the p-value of the test for any dose versus control is nonsignificant (p-value > 0.05), otherwise harmfulness is concluded. This classical test problem is formulated as follows:

$$\begin{aligned} H_0: \pi_{dose} / \pi_{control} &\leq \text{harmless} \\ H_a: \pi_{dose} / \pi_{control} &> \text{harmful} \end{aligned}$$

For the proof of safety approach, the specified direction of harmlessness is defined, i.e., only an increasing number of micronuclei are considered to be harmful. Thus, the noninferiority test, a one-sided hypotheses test, is used for proof of safety to demonstrate the possible harmfulness of a certain dose assuming a three-fold threshold of tolerability. Harmlessness is declared for at least one dose if the upper limit of the relative risk is below the three-fold threshold.

The following hypotheses are evaluated:

$$\begin{aligned} H_0: \pi_{dose} / \pi_{control} &\geq 3 \text{ harmful} \\ H_a: \pi_{dose} / \pi_{control} &< 3 \text{ harmless} \end{aligned}$$

4. Results and Discussion

Figure 1 illustrates the results of the mice bone marrow erythrocyte micronucleus assay. The assay with T02 posted a positive assay. An increase in the frequency of micronucleated polychromatic erythrocytes (MNPCEs) in

treated mice is an indication of induced chromosome damages (Krishna and Hayashi, 2000).

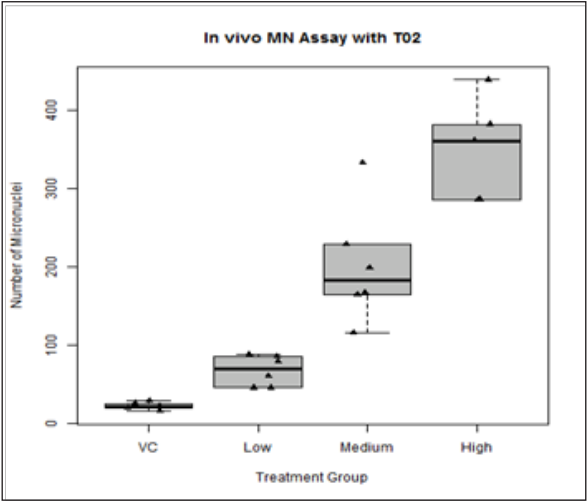


Figure 1. Boxplot for T02

Traditional approach: Poisson model

Table 1 summarizes the results of the in vivo micronucleus test for each dosage group per sacrifice interval. Given in the tables are the relative risk (RR) and the corresponding lower confidence limits estimated by the Poisson model (lower RR).

Genotoxicity activity is indicated by statistically significant dose-related incidence of MNPCEs in the treatment group. A compound is traditionally considered mutagenic in the test system if, at any of the preparation intervals, a statistically significant increase in the number of MNPCEs is found in comparison to the negative control (Shahrim et al., 2006).

Table 1 shows a significant increase in MNPCE observed in all level of doses of T02 over the vehicle control since the point one hypothesized value is less than the estimated lower limit of relative risk. This compound had an indication to induce clastogenic effects and therefore was evaluated as positive at any level of dose administered.

To further explain the genotoxic properties of T02, it is very important to assess the dose-response relation to confirm the toxicity of chemicals (Hayashi et al., 1989). The Cochran-Armitage trend test supports a significant trend (p-value < 0.000) which implies that MN occurrence increases with increasing dose score. The trend test p-values, however, did not provide any information about the biological relevance of the results.

Table 1. Dunnett-type contrasts for relative risks (RR) on MN for T02

Comparison	RR	Poisson		Quasi-Poisson	
		Lower RR	Pvalue	Lower RR	Pvalue
Low-Vehicle	3.13	2.58	0.00*	1.61	0.00*
Med-Vehicle	9.42	7.89	0.00*	5.13	0.00*
High-Vehicle	16.43	13.80	0.00*	9.04	0.00*

* significant at 5% level of significance

Hothorn-Gerhard recommended approach: Quasi-Poisson model

To properly account for the between-animal variations, a quasi-Poisson model was fitted. The estimated dispersion parameter was 12.0 which indicate the occurrence of extra-variability.

Table 1 provides the estimated relative risk alongside with the one-sided lower confidence limit. Apparently, this model yielded similar conclusions as what was previously discussed from using the Poisson model. There was a significant increase in the micronucleus frequency compared with the vehicle control.

Furthermore, HG propose Williams-type contrasts for the relative risk as an alternative to Cochran-Armitage test since the main objective was to demonstrate a possible dose-related trend. This test is sensitive to several shapes as compared to CA test which is sensitive for near linear shape. The direction of interest, i.e., increasing micronuclei induces chromosomal damage, and only the lower limit is required for a conclusion regarding the trend. Therefore we are only after in the performance of the lower bound confidence limits. The multiple comparisons were done by successively pooling the higher dose groups and compared to the control group. In Table 2 the relative risk estimates and their lower simultaneous confidence limits for Williams-type contrasts for the quasi-Poisson model are given. Since all three comparisons were significantly larger than 1, it was concluded that there is a significant increase in the number of MN with increasing dosage wherein the High dose group led to the most pronounced change in the number of MN compared to the control group.

Table 2. Williams-type contrasts for relative risks (RR) on MN for T02

Comparison	RR	Lower RR
C1: Vehicle vs. High	16.43	9.41
C2: Vehicle vs. Medium and High	12.13	6.99
C3: Vehicle vs. All doses	7.52	4.32

Proof of hazard vs. Proof of safety

Using the traditional approach, the evaluation of the number of micronuclei using proof of hazard revealed that none of the doses of T02 found to be harmless, i.e., significant p-values was observed in all tests (see Table 1). In this study, noninferiority can be determined by one-sided upper confidence limits because increasing micronuclei is of interest. Assuming a-priori definition of an acceptance threshold θ equals to three for proof of safety, a noninferiority could not be claimed in all dosage groups of T02 (see Table 3). This coincides with the previous findings that T02 is a positive compound.

Table 3. Proof of Safety: One-sided Confidence Limit on Noninferiority (T02)

Comparison	RR	Upper RR
Low-Vehicle	3.13	6.08
Med-Vehicle	9.42	17.29
High-Vehicle	16.43	29.87

5. Conclusions and Recommendations

The ultimate goal of this study is to implement, apply and discuss the recommendations of HG to explain the induced chromosomal damage potential of T02 using *in vivo* micronucleus assay. The two different statistical approaches, i.e., Poisson and quasi-Poisson models lead to the same conclusion – a positive result for T02. However, the crucial part in the analysis is the consequence of failing to account for overdispersion, as in the case of fitting the Poisson model, which may lead to incorrect inferences. Evidently, the comparison of incidences of micronucleated polychromatic erythrocytes (MNPCE) using a quasi-Poisson model is shown to be more appropriate than a similar procedure using a Poisson model. The latter statistical approach becomes more liberal with increasing overdispersion, giving confidence limits for significance that are too narrow.

In terms of claiming a possible genotoxicity of the given compounds, a three-fold threshold of tolerability for the noninferiority test is assumed. However, this chosen threshold indicates the difficulty of claiming harmlessness with sample size of five. HG discuss that for a primary claim of being not genotoxic, more animals than five are needed. In toxicological studies, the sample size is often determined on the basis of regulatory guidelines. Moreover, aside from using non-inferiority test for proof of safety, an alternative approach such as the step-up estimation of the maximal safe dose can also be performed (Hothorn and Gerhard, 2009).

In terms of the interpretation of the results in *in vivo*, there are several criteria for determining a positive response, one of which is a statistically significant dose-

related increase in the number of micronucleated polychromatic erythrocytes. Another criterion may be based upon detection of a reproducible and statistically significant positive response for at least one of the test substance concentrations. If none of the criteria are satisfied, then the test substance is considered to be nonmutagenic in this system. Sofuni et al. (1990) considered the dose response to be (strong) positive, if it had two significant doses out of three dose groups and decided it to be weakly positive if it had only one significant dose and there was a significant trend. However, both biological and statistical significance should be considered together in an evaluation. Lastly, it is highly recommended to confer with toxicologist for the biological relevance of the findings.

References

- ALTMAN, D.G. and BLAND, J.M., 1995, Absence of Evidence is Not Evidence of Absence, *British Medical J.*, 311(7003): 485.
- ARMITAGE, P., 1955, Test for Linear Trends in Proportions and Frequencies, *Biometrics* 11(3): 375-386.
- BRETZ, F. and HOTHORN, L.A., 2002, Detecting dose-response using contrasts: Asymptotic power and sample size determination for Poisson data, *Stat Med.* 21(22): 3325-3335.
- COCHRAN, W.G., 1954, Some methods for strengthening the common χ^2 tests. *Biometrics* 10: 417-451.
- HAYASHI, M. et. al., 1989, A procedure for data analysis of the rodent micronucleus test involving historical control, *Environmental and Molecular Mutagenesis* 13: 347-356.
- HERBERICH, E. and HOTHORN, L., 2012, Statistical evaluation of mortality in long-term carcinogenicity bioassays using a Williams-type procedure, *Regulatory Toxicology and Pharmacology* 64: 26-34.
- HAUSCHKE, D. et.al., 1999, Proof of safety in toxicology based on the ratio of two means for normally distributed data, *Biometrical Journal* 41(3): 295-304.
- HOTHORN, L.A. and GERHARD, D., 2008, Statistical evaluation of the in vitro micronucleus assay, Reports of the Institute of Biostatistics No. 09/2008.
- HOTHORN, L.A. and GERHARD, D., 2009, Statistical evaluation of the in vivo micronucleus assay, *Arch Toxicol* 83: 625-634.
- KIM, B.S. et. al., 2000, Statistical analysis of in vivo rodent micronucleus assay, *Mutation Research* 469: 233-241.
- KIRKLAND, D.J., 2000, *Statistical Evaluation of Mutagenicity Test Data*, Cambridge University Press.
- KRISHNA, G. and HAYASHI, M., 2000, In vivo rodent micronucleus assay: Protocol, conduct and data interpretation, *Mutation Research* 455(1-2): 155-166.
- LEE, J. et. al., 2012, Analysis of overdispersed count data: Application to the human Papillomavirus Infection in Men (HIM) study, *Epidemiol Infect* 1087-1094.

- MARGOLIN, B.H. and RISKO, R.J., 1988, The statistical analysis of in vivo genotoxicity data: Case studies of the rat hepatocyte UDS and mouse bC (eds): "Evaluation of Short-Term Test for Carcinogens. Report of the International Program on Chemical Safety's Collaborative Study on In Vivo Assays." Cambridge: Cambridge University Press vol. 1 pp. 129-142.
- PARODI, S. and BOTTARELLI, E., 2006, Poisson regression model in epidemiology - An introduction, *Medic. Vet.* 25-44.
- SHAHIM, Z. et. al., 2006, The in vivo rodent micronucleus assay of Kacip Fatimah (*Labisia pumila*) extract, *Tropical Biomedicine* 23(2): 214-219.
- SOFUNI, T. et. al., 1990, A comparison of chromosome aberration induction by 25 compounds tested by two Chinese hamster cells (CHO and CHL) systems in culture, *Mutation Research* 241:175-213.

Bootstrapping Penalty Analysis in Sensory Evaluation of Pizza Products

Catherine Estiaga
School of Statistics
University of the Philippines Diliman

Penalty analysis is a popular method used to evaluate data from sensory evaluation using the Just About Right Scale and the Hedonic Scale. Although the test estimates the mean drops for the "Too Little" and "Too Much" categories of product attributes, penalty analysis does not provide information that can be used to test the effect of each attribute on the overall liking score. Bootstrap resampling method when used together with penalty analysis estimates the standard error of the mean drops and allows to test for the significance. This method is used in product testing of pizza products.

Key Words: bootstrap method, penalty analysis, just about right scale, hedonic scale, mean drops

1. Introduction

Penalty or Mean Drop Analysis is a method used extensively in sensory data analysis to identify potential directions for the improvement of products. It determines what attributes can lead to an increase in overall liking (Paczkowski, 2009) by showing how many points you lose for having a product "too much" or "too little" for a consumer. Penalty analysis can aid product developers in understanding differences in product preference. It can also help in understanding the basis for consumer segmentation and product substitutability (Rothman, 2007).

Product attributes used in penalty analysis are measured using the just-about-right (JAR) scales (Plaehn, 2009). JAR scale is widely used to measure the suitability of a specific attribute and to provide the optimum levels of attributes in a product. These scales are categorical variables, usually consist of five points, which assess whether there is too little, too much or just-about-right level of a product attribute (Lawless and Heyman, 1999).

Although penalty analysis has been extensively used in the food industry, there is no statistical procedure to determine the significance of the effect of each attribute on the overall liking of a product. Because the categories below and above the JAR levels are collapsed, there are not enough observations to estimate the standard error of the mean drop and to compute for the significance (Rothman, 2007). In this connection, an alternative method was developed by Xiong et al. (2007) called bootstrapping penalty analysis. Bootstrapping penalty analysis allows performing statistical testing on the results of a penalty analysis, using a technique called bootstrap to estimate variance (Xiong et al., 2007).

2. Review of Related Literature

Studies on sensory evaluation, penalty analysis, bootstrap method and bootstrapping penalty analysis are summarized below.

2.1 Sensory evaluation

Sensory Evaluation is defined by The Institute of Food Technologists (IFT) Sensory Evaluation Division USA as “a scientific discipline used to evoke, measure, analyze and interpret sensations as they are perceived by the senses of sight, smell, taste, touch and hearing” (Prell, 1976). Sensory evaluation was initially used as a service provider supplying data. However, its role has changed significantly over the years. Now, it provides insights to help guide product development, product matching and product improvement (Kemp et al., 2009). Aside from these, sensory evaluation is now widely used in process change, cost reduction, quality control, consumer acceptance, consumer preference, monitoring competition, product sensory specification, raw materials specifications, storage stability, panel selection/training and advertising claims (Stone and Sidel, 2004; Mason and Nottingham, 2002). Sensory scientists working on food industry did most of the development on sensory evaluation. It is also now being implemented by varied industries such as in personal care, paint, household cleaners, hospitality management and in many others (Chambers and Wolf, 1996).

The most widely used scale for measuring food acceptability in sensory evaluation is the hedonic scale. It was developed by David Peryam and his colleagues to assess acceptability of food items for soldiers (Peryam and Pilgrim, 1957). The hedonic scale that has been most commonly used is the 9-point scale (Table 1) in which the consumer rates their preference for food, ranging from “dislike extremely” to “like extremely” with the midpoint of 5 being “neither like nor dislike”. Study showed that longer scales such as the 9-point scale tend to be more discriminating compared to a shorter one such as the 7-point scale (Jones et al., 1955).

Table 1. 9-Point Hedonic Scale

9	Like extremely
8	Like very much
7	Like moderately
6	Like slightly
5	Neither like nor dislike
4	Dislike slightly
3	Dislike moderately
2	Dislike very much
1	Dislike extremely

The JAR scale is another scale in food industries and market researchers (Lawless and Heymann, 1999) extensively used to assist them in identifying possible shortcomings of the products evaluated (Xiong et al., 2007). JAR is a dichotomous scale that measures whether a specific attribute is present in optimal levels in a product. It is usually five categories wide as shown in Table 2 and is anchored in the middle by “just right,” to the top with “too much” and to the bottom with “too little” of an attribute (Stone and Sidel, 2004). Although JAR scale could have as few as three categories, this should be used carefully since rating scales usually should not have less than 5 categories (Chambers and Wolf, 1996). Aside from this, people generally tend to rate categories in the middle of the scale that is why a 5-point JAR scale is preferred over the 3-point scale (Whiston, 2009). Data gathered in this fashion are typically summarized by dividing responses to the “just right” scale into three categories: “too little (TL),” “just right (JAR),” and “too much (TM)” (Anon, 2003 and Rothman, 2007).

JAR scale provides a quick indication of attribute intensity direction. JAR data that are normally distributed around the center are indicative of an optimized level of a specific product attribute. However, information should be considered with caution when using this scale. Consumers are not generally familiar with very specific attributes so use of JAR should be limited to very simple attributes like sweetness or saltiness (Gatchalian and Brannan, 2009).

Table 2. 5-Point JAR Scale

5	Much too much
4	Too much
3	Just about right
2	Too little
1	Much too little

2.2 Penalty analysis

Penalty analysis is a test used to identify potential directions for product improvement. It is used by market researchers, sensory analysts and product developers to determine what product attributes affect most the overall liking score, purchase intent and other product-related measure (Plaehn, 2009). It assists in identifying attributes that cause an increase or decrease in hedonic scale associated with sensory attributes not at optimal levels in a product (Paczkowski, 2009) allowing the product developer to decide on what sensory properties should be improved or adjusted.

Penalty analysis itself is premised on the idea that the maximum hedonic score/overall liking will occur at the JAR point (Plaehn, 2009). Thus, it uses the data collected on the 5-point JAR scale and the liking scores on a 9-point hedonic scale.

The principle for calculating mean drops is given in Tables 3 and 4. The 5-point JAR scale for each attribute is collapsed into 3 categories – “too little,” “just about right” and “too much” (Anon, 2003 and Rothman, 2007). The mean overall liking score from the 9-point hedonic scale and the percentage of respondents represented in each of the three categories are calculated. The mean drops in the overall liking score for the “too little” and “too much” from the JAR level are likewise determined (Paczkowski, 2009). Mean drops are calculated by subtracting the mean overall liking score for the JAR group to the mean overall liking score of the “too much” or “too little” categories.

Table 3. Penalty Analysis Calculation Principles (Meullenet et al., 2007)

Hedonic Scores (X)	JAR scores
7	3
6	4
7	5
8	3
9	3
6	1
5	2
2	4
4	5
6	4
6	1

Table 4. Mean Drops and % of Respondents on a Collapsed 3-point JAR Scale

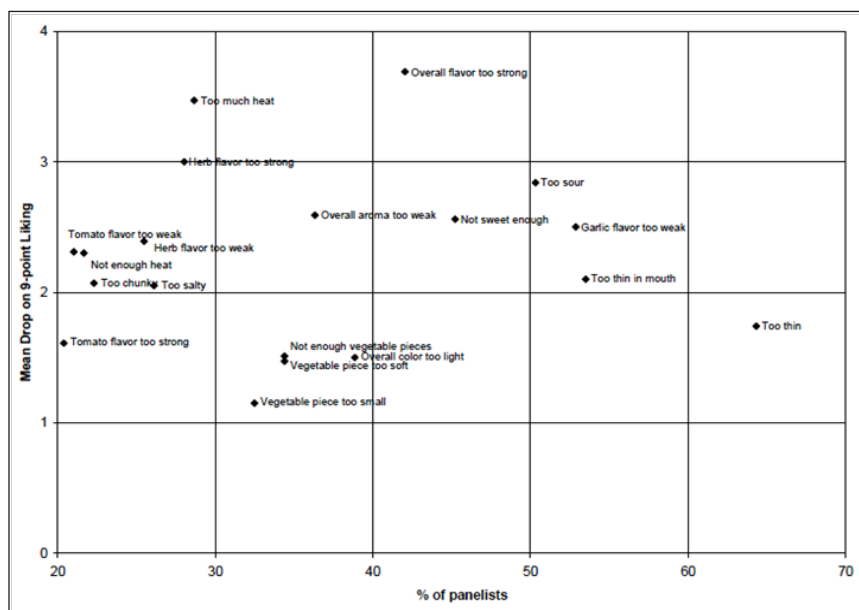
$\bar{X}_{<JAR} = 5.7$	Mean Drop $<JAR = 8.0 - 5.7 = 2.3$	% respondents $<JAR = 3/11 = 27\%$
$\bar{X}_{JAR} = 8.0$		% respondents $JAR = 28\%$
$\bar{X}_{>JAR} = 5.0$	Mean Drop $>JAR = 8.0 - 5.0 = 3.0$	% respondents $>JAR = 5/11 = 45\%$

A minimum of 70% of the responses is usually expected in the JAR categories to conclude that a specific attribute is at its optimal level (Rothman, 2007 and Chambers et al., 1996). A skew cut-off percentage for the “too little” or “too much” categories is also advisable. This percentage depends upon the product category, sample size of the study, comfort level of the researcher and the stage of product research being conducted. Larger skews may be acceptable in earlier stages of research while a smaller skew is recommended during the later part of the study (Rothman, 2007). Since the typical skew cut-off percentage being used in the industry is 20%, mean drops are not calculated if the proportion of respondents who rated a certain attribute is less than this. Responses below 20% is too small to be considered and they might not be reliable enough (Rothman, 2007 and Meullenet et al., 2007).

Mean drops are usually represented graphically for individual products using scatter plot of the percentage of respondents in a non-JAR group on the x-axis and the mean drops on y-axis (Xiong and Meullenet, 2009) as shown in Figure 1. High negative mean drops that are associated with large proportions of respondents (upper right quadrant) are assigned greater importance than low negative or positive mean drops associated with small numbers of respondents (lower left quadrant) (Plaehn, 2009). In the figure below, the most troublesome attributes are overall flavor (too strong), garlic flavor (too weak), sweetness (not enough), overall aroma (too weak), heat (too much), herb flavor (too strong) and thickness (too thin in the mouth) (Xiong and Meullenet, 2009).

Penalty analysis helps the product developer or sensory analyst in determining what product attributes should be improved first (Meullenet et al., 2007). However, it should be noted that adjusting the JAR skews is not a guarantee that the hedonic or liking score will increase by the amount of the mean drops (Rothman, 2007).

Some advantages of penalty analysis are that it is easy to perform, is product-specific and is easily interpretable. One major limitation of this test is the fact that the categories below and above the JAR level are collapsed resulting in a loss of information. There is also no test to determine the significance of the mean drops (Xiong et al., 2007).



Source: Xiong and Meullenet (2009)

Figure 1. Example of Penalty Analysis Plot

2.3 Bootstrap method

The properties of bootstrap and its connection to the other resampling methods were not realized until Efron and Tibshirani (1993). Since then, the bootstrap has provided a powerful set of solutions for statisticians and a source of theoretical and methodological problems for statistics (Davison et al., 1986).

The bootstrap method is a well-established computer-intensive Monte Carlo technique (Xiong and Meullenet, 2009). It replicates the original data to simulate a larger population, thus allowing many samples to be drawn and statistical tests such as bootstrap variances, distributions and confidence intervals to be calculated (Chernick, 1999 and Stine, 1989).

Bootstrap can be useful in areas where it is difficult to obtain large samples (Fan, 1994). This method also does not require the assumption that the standard errors be randomly and normally distributed. Bootstrap method likewise provides a user-friendly alternative to cross-validation and jackknife to augment statistical significance testing.

The bootstrap approach consists of drawing many independent random samples by simple random sampling with replacement, evaluating the sample statistics of the corresponding bootstrap replications, and estimating the standard error of the empirical probability distribution of the data ($\hat{F}$) by the empirical standard deviations of the replications (Efron and Tibshirani, 1993; Chernick, 1999; and Diaconis and Efron, 1983).

Given $x = (x_1, x_2, \dots, x_n)$, a random sample from x with unknown distribution F (Chernick, 1999). Let $\hat{F}$ be the empirical probability distribution of the data. An equal probability of $1/n$ is given on each x_i and let $x_1^*, x_2^*, \dots, x_n^*$ be a random sample from $\hat{F}$ (Efron and Gong, 1983),

$$x_1^*, x_2^*, \dots, x_n^* \sim \hat{F} \quad (1)$$

x_i^* is not the actual data set, but rather a randomized or resampled version of x (Efron and Tibshirani, 1993). Each x_i^* is drawn independently with replacement from the original set $\{x_1, x_2, \dots, x_n\}$ (Efron and Gong, 1983). The number of bootstrap replications required to create an ideal bootstrap sampling distribution ($\hat{F}$) is n^n where n is the original sample size. However, Diaconis and Efron (1983) and Stine (1989) have demonstrated that an ideal $\hat{F}$ can be as low as 100 replications. But for more complicated procedures for construction of confidence intervals, it has been suggested that as many as 1000 replications may be needed.

2.4 Bootstrapping penalty analysis

Bootstrapping penalty analysis allows performing statistical testing on the results of a penalty analysis using the bootstrap method to estimate variance (Xiong et al., 2007). The original data set contains the overall liking score and the JAR scale for each attribute. Mean drops for the TL and TM levels are calculated for each attribute according to the penalty analysis methodology. The bootstrap estimate of variability is obtained through resampling of the data pairs an enormous number of times (Xiong and Meullenet, 2007). The bootstrap method resamples the original data with replacement (Plachn and Horne, 2008). A data pair is sampled and then returned to the dataset and has $1/n$ chance of being drawn again. All bootstrap samples have the same size as the original data. A particular pair can also appear in several of the bootstrap samples and may appear more than once in a particular sample (Xiong et al., 2007). The process of estimating the standard error of a mean drop (Meullenet et al., 2007) for a single attribute is illustrated in Figure 2.

Mean drops for the TL and TM levels are calculated for all the bootstrap samples for each attribute and are averaged to obtain the final mean drops for the two categories. The bootstrap mean and standard error of the mean are then computed based on the bootstrap replications of the mean drops (Xiong et al., 2007). Afterwards, the bias in the bootstrap estimate of the mean is removed to obtain the adjusted bootstrap mean. T-test is finally computed to determine significance of the mean drops for each attribute (Figure 2).

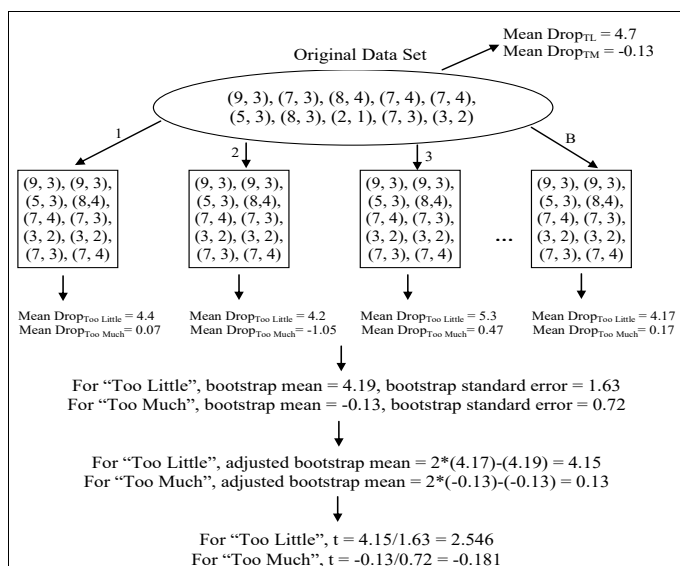


Figure 2. Scheme for Bootstrapping Penalty Analysis

3. Methodology

Sensory evaluation of two Hawaiian Pizza was done. Data gathered from the test was analyzed using bootstrapping penalty analysis. Detailed methodology is described below.

3.1 Sensory evaluation

The data used in this study was obtained from the results of in-house sensory evaluation of two Hawaiian Pizza. The two pizza products were subjected to sensory evaluation protocol by Gatchalian (1989) using 30 respondents for each pizza. Based on the studies conducted by Gatchalian and Brannan (2009) and Chambers and Wolf (1996), the number of respondents for in-house sensory evaluation can be as few as 16-20. However, the usual practice is to require at least 30 respondents. A sample size of at least 30 is considered adequate although larger sample size is much better (Levine and Stephan, 2010). The sensory evaluation conducted was only at the product development stage which maybe different when you do similar test in a pre-commercialization stage.

Respondents were pre-recruited from different areas in Metro Manila and were invited to the Product Development Center. Recruitment of respondents was carried out in advance to ensure that no seats were left unfilled and that respondents arrived on time. Respondents were from the Broad C (C1/C2) households with ages ranging from 18-35 years old. Hawaiian Pizza A has 47% female and 53% male respondents. Hawaiian Pizza B, on the other hand, has 50% female and 50%

male respondents. All respondents should have consumed Hawaiian Pizza for at least once in the past four weeks.

Hawaiian Pizza A was made of pizza sauce, cheese, 1.33 in² spiced ham and glazed pineapple tidbits. Hawaiian Pizza B, on the other hand, was made of pizza sauce, cheese, 1in² spiced ham and fresh pineapple tidbits. The two pizzas were prepared using a family (12 inches) thick crust. After cooking, the baked pizza was sliced into eight pieces. Each respondent was given a slice of pizza for evaluation.

A proto-monadic design was used for each pizza. In the proto-monadic design, each respondent evaluates both products. Specifically, each respondent rates the first product with a monadic test followed by a preference test. The order of the products evaluated is rotated to reduce bias (Baosheng, 2005). In this study, Hawaiian Pizza A was evaluated first with a monadic test. The products were assessed in terms of overall liking using the 9-point hedonic scale, with 9 = like extremely, 1 = dislike extremely (Amerine et al., 1965). A JAR scale was also used to evaluate size of the toppings, amount of cheese, amount of meat, amount of pineapple, saltiness, cheesiness, meaty taste, pineapple taste and overall flavor blend (Table 5). The JAR scale is a 5-pointscale with the center point being just about right, the left anchor being too little of the attribute, and the right anchor being too much of the attribute (Meilgaard et al., 1999). After the monadic test, Hawaiian Pizza A and Hawaiian B were evaluated sequentially by the respondents. They were then asked which of the two pizzas they prefer more. During the test, respondents were provided with mineral water and unsalted crackers for palate cleansing. They were provided afterwards with gift certificates for their participation. The same procedure was done for Hawaiian Pizza B using 30 different respondents.

Table 5. Scales Used in the Sensory Evaluation of Hawaiian Pizza

Overall Acceptability Like extremely Like very much Like moderately Like slightly Neither like nor dislike Dislike slightly Dislike moderately Dislike very much Dislike extremely	Saltiness Definitely too salty Somewhat salty Just right Somewhat lacking in saltiness Not salty at all
Size of the Toppings Definitely too big Somewhat big Just the right size Somewhat small Definitely too small	Cheesiness Definitely too cheesy Somewhat cheesy Just right Somewhat lacking in cheesiness Not cheesy at all
Amount of Cheese Definitely too much Somewhat too much Just the right amount Somewhat too little Definitely too much	Meaty Taste Definitely too strong Somewhat strong Just right Somewhat weak Definitely too weak

Amount of Meat Definitely too much Somewhat too much Just the right amount Somewhat too little Definitely too much	Pineapple Taste Definitely too strong Somewhat strong Just right Somewhat weak Definitely too weak
Amount of Pineapple Definitely too much Somewhat too much Just the right amount Somewhat too little Definitely too much	Overall Flavor Blend Definitely too strong Somewhat strong Just right Somewhat weak Definitely too weak

In analyzing the results of sensory evaluation, the percentage of the top three categories in the 9-point hedonic scale was obtained for each pizza. At least 80% is needed to conclude that a product is acceptable. The result of the preference test was also counted. The minimum number of agreeing judgments is 21 to conclude that a product is significantly preferred over the other at 5% level of significance (Stone and Sidel, 2004).

3.2 Bootstrapping penalty analysis

The responses of the respondents in the 5-point JAR scale were first grouped into either too much, JAR or too little categories. The bottom two-box was categorized as “too little” of the attribute, the middle box as “JAR” and the top two-box as “too much” of the attribute. Collapsing the categories below and above the JAR level is necessary because the number of responses in the TL or TM category is often not large enough (Meullenet et al., 2007). The frequencies and percentage of respondents corresponding to the mean liking scores for the three categories were then calculated for each attribute. Mean drops for the TM and TL categories were afterwards determined by using the formula below (Xiong et al., 2007):

$$\text{Mean Drops}_{\text{Too Little}} = \bar{X}_{JAR} - \bar{X}_{\text{Too Little}} \quad (2)$$

$$\text{Mean Drops}_{\text{Too Much}} = \bar{X}_{JAR} - \bar{X}_{\text{Too Much}} \quad (3)$$

where X = overall liking score and $\bar{X}$ = mean overall liking score.

Mean drops are not computed if the percentage of respondents is less than 20% because it is too small to be considered and they might not be reliable enough. 20% is the skew cut-off percentage used in this study since it is the most commonly used in the industry (Rothman, 2007 and Meullenet et al., 2007).

A random number generator to create bootstrap samples was then done for each attribute. For each replication, a sample was selected from the original data set randomly. A data pair was obtained and then returned to the dataset and has 1/n chance of being drawn again (Xiong and Meullenet, 2007).

100 bootstrap samples were simulated from the original data. Based on the study conducted by Tibshirani (1985) and Schmidheiny (2010), $B = 100$ is usually adequate for estimating variances, biases and standard errors (Stine, 1989). For each attribute, mean drops and percentage of respondents for the TL and TM categories were then calculated for each of the 100 bootstrap replicates (Xiong et al., 2007) using the formulas given below.

$$S_{iTL}^* = \bar{X}_{iJAR}^* - \bar{X}_{iTL}^* \quad (4)$$

$$S_{iTM}^* = \bar{X}_{iJAR}^* - \bar{X}_{iTM}^* \quad (5)$$

where

$i = 1, \dots, B$ where B is the number of bootstrap replicates

$x_i^* = (x_1^*, x_2^*, \dots, x_B^*)$ is a bootstrap sample drawn from the original data set

$\bar{X}_{iJAR}^*$ = mean overall liking calculated from X_i^* for the JAR Level

$\bar{X}_{iTM}^*$ = mean overall liking calculated from X_i^* for the TM Level

$\bar{X}_{iTL}^*$ = mean overall liking calculated from X_i^* for the TL Level

S_{iTL}^* = mean drops calculated from X_i^* for the Too Little Level

S_{iTM}^* = mean drops calculated from X_i^* for the Too Much Level

Mean drops for each of the bootstrap replicates were afterwards averaged to obtain the mean drops for the TL and TM levels for each attribute. The mean and standard error of the mean drops were computed using the formulas below (Xiong et al., 2007). By generating a large number of bootstrap samples, each observation will likely to contribute to the final variance (Plaehn and Horne, 2008).

$$\bar{s}_{bTL}^* = \frac{\sum_{i=1}^B S_{iTL}^*}{B} \quad (6)$$

$$\bar{s}_{bTM}^* = \frac{\sum_{i=1}^B S_{iTM}^*}{B} \quad (7)$$

$$\hat{s}e_{bTL} = \sqrt{\frac{\sum_{i=1}^B (s_{iTL}^* - \bar{s}_{bTL}^*)^2}{B-1}} \quad (8)$$

$$\hat{se}_{bTM} = \sqrt{\frac{\sum_{i=1}^B (s_{iTM}^* - \bar{s}_{bTM}^*)^2}{B-1}} \quad (9)$$

where

B = number of bootstrap samples

$\bar{s}_{bTL}^*$ = bootstrap estimates of the true mean drops for the TL level

$\bar{s}_{bTM}^*$ = bootstrap estimates of the true mean drops for the TM level

$\hat{se}_{bTL}$ = bootstrap estimates of the standard errors of the mean drops for the TL level

$\hat{se}_{bTM}$ = bootstrap estimates of the standard errors of the mean drops for the TM level

To obtain the adjusted bootstrap mean, the bias in the bootstrap estimate of the mean was removed using the following (Meullenet et al., 2007):

$$\bar{s}_{bTL} = 2S_{nTL} - \bar{s}_{bTL}^* \quad (10)$$

$$\bar{s}_{bTM} = 2S_{nTM} - \bar{s}_{bTM}^* \quad (11)$$

where

S_{nTL} = original mean drops obtained from the original data set for the TL level

S_{nTM} = original mean drops obtained from the original data set for the TM level

$\bar{s}_{bTL}$ = adjusted bootstrap estimates of the means for the TL level

$\bar{s}_{bTM}$ = adjusted bootstrap estimates of the means for the TM level

After computing the standard error and the adjusted bootstrap mean, t-test was done to determine the significance of the mean drops for each attribute (Horne and Plaehn, 2008) and to determine which attributes significantly affect the overall liking of the product.

Mean drop testing was done if the number of responses on a certain attribute is greater than 20%. Penalty test was performed to test the following hypothesis:

$$H_0: \mu_{JAR} - \mu_{TM} = 0$$

$$H_0: \mu_{JAR} - \mu_{TL} = 0$$

$$H_a: \mu_{JAR} - \mu_{TM} \neq 0$$

$$H_a: \mu_{JAR} - \mu_{TL} \neq 0$$

where μ is the population mean of the TL, JAR and TM categories.

If H_0 is not rejected, the mean drops on overall liking are not significant for a specific product attribute at 5% level of significance. This means that the current formulation is ready for commercialization and launching. On the other hand, if

Ho is rejected, it means that the mean drops are significant and the formulation still needs to be adjusted to obtain acceptable product.

Results of bootstrapping penalty analysis were represented graphically by plotting the mean drops on the y-axis against the percentage of respondents on the x-axis. Attributes on the upper right quadrant are the ones that need to be prioritized for improvement (Xiong et al., 2007).The formulation that will be commercialized and launched will be based not only on the results of bootstrapping penalty analysis but also on the results of overall acceptability percentage and preference test.

4. Results and Discussion

Sensory evaluation of two Hawaiian Pizza was done. Hawaiian Pizza A was prepared using pizza sauce, cheese, 1.33 in² spiced ham and glazed pineapple tidbits. Hawaiian Pizza B, on the other hand, was prepared using combination of pizza sauce, cheese, 1.0 in² spiced ham and fresh pineapple tidbits. Data from the sensory evaluation were analyzed using ordinary penalty analysis and bootstrapping penalty analysis. Results were discussed below.

4.1 Sensory evaluation

The results of sensory evaluation of Hawaiian Pizza A and Hawaiian Pizza B are shown below in Figure 3. Hawaiian Pizza B obtained a higher overall acceptability rating compared to Hawaiian Pizza A. In terms of overall acceptability, Hawaiian Pizza A and Hawaiian B were rated by most of the panelists as “like moderately” and “like very much” respectively. Overall flavor blend, on the other hand, was perceived as “somewhat too strong” for both pizza products. All other attributes were rated as “just about right” for both products.

The percentages on the top three categories of the 9-point hedonic scale was computed to obtain the overall acceptability percentage. Hawaiian Pizza A and Hawaiian Pizza B obtained an overall acceptability percentage of 86.21% and 90.32% respectively. Both are above the passing rate which is 80%. Results of preference test also showed that there is no significant preference between Hawaiian Pizza A and Hawaiian Pizza B at 5% level of significance.

Table 6. Overall Acceptability Percentage and Preference Test of Hawaiian Pizza A and Hawaiian Pizza B

	Overall Acceptability Percentage (Top three categories)	Preference Test
Hawaiian Pizza A	86.21%	13
Hawaiian Pizza B	90.32%	17

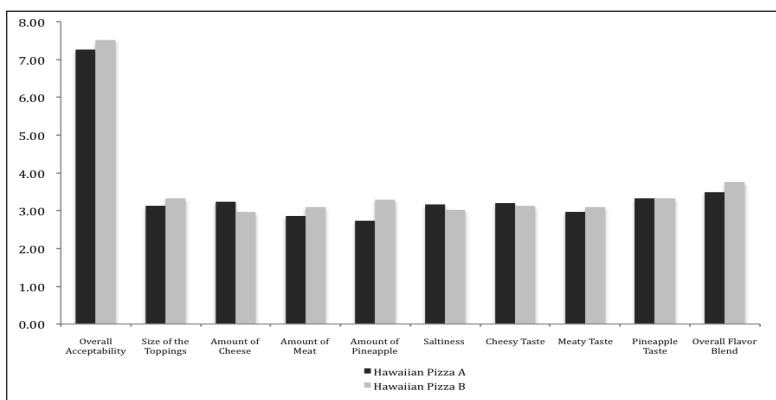


Figure 3. Results of Sensory Evaluation of Hawaiian Pizza A and Hawaiian Pizza B

4.2 Ordinary penalty analysis

Results of sensory evaluation of Hawaiian Pizza A and Hawaiian Pizza B were analyzed using ordinary penalty analysis. Figures 4 and 5 show the percentage of respondents for the too little, just about right and too much levels for the size of the toppings, amount of cheese, amount of meat, amount of pineapple, saltiness, cheesiness, meaty taste, pineapple taste and overall flavor blend of the Hawaiian Pizza A and Hawaiian Pizza B respectively.

Most of the respondents rated Hawaiian Pizza A (Figure 4) as being in the just about right level for the size of the toppings (79%), amount of cheese (54%), amount of meat (46%), amount of pineapple (64%), saltiness (82%), cheesiness (54%), meaty taste (57%), pineapple taste (61%) and overall flavor blend (43%). However, at least 70% of the responses should be at the JAR category to conclude that a specific attribute is at its optimal level. (Xiong et al., 2007). For Hawaiian Pizza A, only two attributes namely size of the toppings and saltiness are at its optimal level.

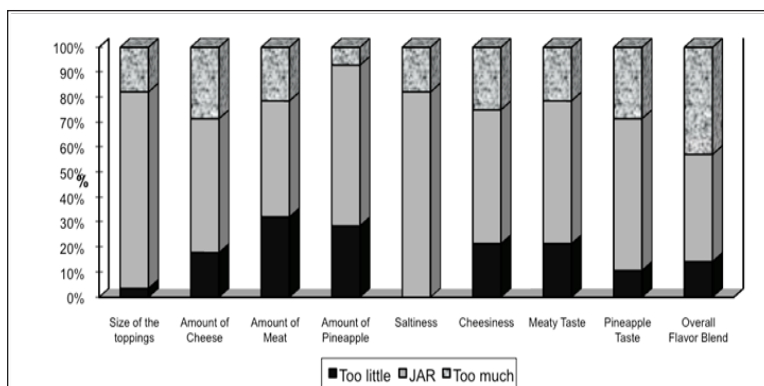


Figure 4. Percentage of Respondents for the Too Little, Just About Right and Too Much Levels of Hawaiian Pizza A

Hawaiian Pizza B (Figure 5) was also rated by most of the respondents as being in the JAR level for the size of the toppings (53%), amount of cheese (47%), amount of meat (50%), amount of pineapple (63%), saltiness (73%), cheesiness (63%), meaty taste (73%) and pineapple taste (53%). Overall flavor blend, on the other hand, was perceived by 60% of the respondents as being in the “too much” level. Saltiness and meaty taste are the only attributes of Hawaiian Pizza B that are at its optimal level.

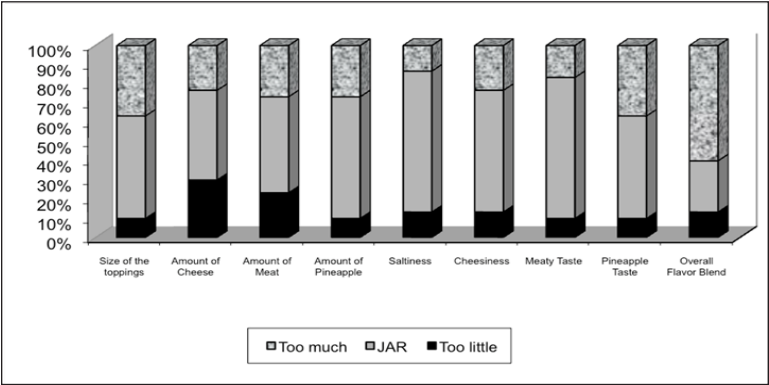


Figure 5. Percentage of Respondents for the Too Little, Just About Right and Too Much Levels of Hawaiian Pizza B

Figures 6 and 7 show the penalty analysis plot for Hawaiian Pizza A and Hawaiian Pizza B respectively. Mean drops are represented by plotting the mean drops on the y-axis against the percentage of respondents in a non-JAR group on the x-axis. Attributes located in the upper right quadrant would be selected as those needing to be improved (Meullenet et al., 2007). Based on Figure 6, the most “troublesome” attribute is the “too little” amount of meat. 32% of the respondents considered Hawaiian Pizza A to be having too little amount of meat with a mean drop of 1.33 from the overall liking score. 29% of the respondents also considered the product to have “too little” amount of pineapple with a mean drop of 1.26. Other attributes that caused negative mean drops from the overall liking score are the “too weak” cheesiness, “too weak” meaty taste, “too much” amount of meat and “too strong” pineapple taste. Overall flavor blend and meaty taste, on the other hand, obtained positive mean drops indicating that the respondents who rated Hawaiian Pizza A as having too strong overall flavor blend and too strong meaty taste liked it more than the respondents who rated the product as “just about right.”

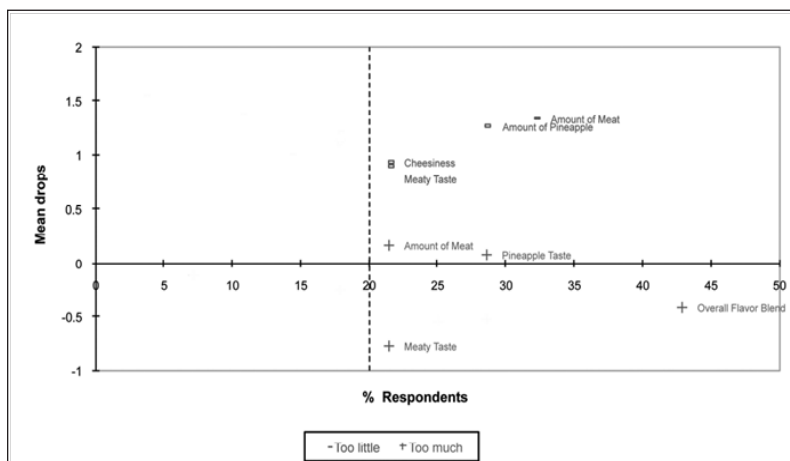


Figure 6. Penalty Analysis Plot for Hawaiian Pizza A

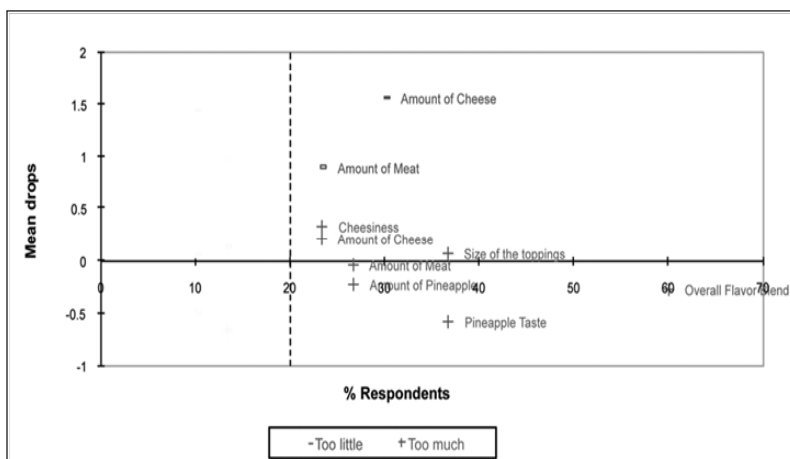


Figure 7. Penalty Analysis Plot for Hawaiian Pizza B

Figure 7 shows that the amount of meat, amount of pineapple, overall flavor blend and pineapple taste were rated as too strong by more than 20% of the respondents causing a positive mean drop on the overall liking score of Hawaiian Pizza B. These respondents liked more the said attributes than the respondents who rated those at the JAR level. Based on Figure 7, the most troublesome attribute is the amount of cheese. It was found to be “too little” by 30% of the respondents and was responsible for a large drop of 1.56 points in the overall liking score. Other attributes that are responsible for negative mean drops are too little amount of meat, too strong cheesiness, too much amount of cheese and too big size of the toppings.

4.3 Bootstrapping penalty analysis

Results of sensory evaluation of Hawaiian Pizza A and Hawaiian Pizza B were subjected to bootstrapping penalty analysis using 100 samples. Figure 8 and 9 show the percentage of respondents for the too little, JAR and too much levels for the size of the toppings, amount of cheese, amount of meat, amount of pineapple, saltiness, cheesiness, meaty taste, pineapple taste and overall flavor blend of the Hawaiian Pizza A and Hawaiian Pizza B respectively. For Figure 8, it can be observed that most of the respondents perceived the product to be in the JAR level for all attributes. However, only saltiness and size of the toppings are at its optimal level with 73% of the respondents rating the product to be in the JAR level. Most of the respondents also perceived Hawaiian Pizza B to be in the JAR level for all attributes except for overall flavor blend which was perceived as being in the “too much” level by 61% of the respondents (Figure 9). Attributes of Hawaiian Pizza B that are at its optimal level are the meaty taste and saltiness.

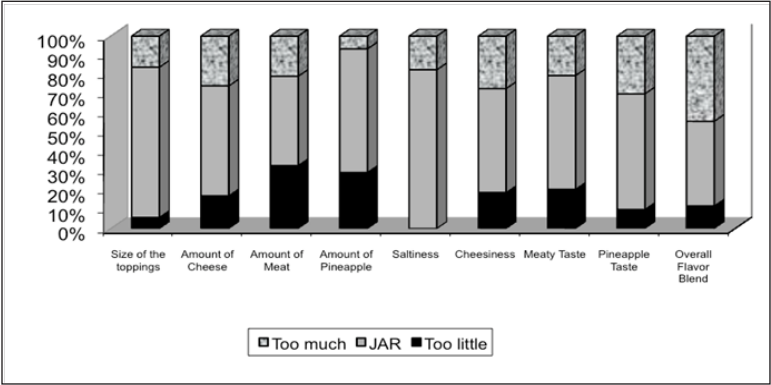


Figure 8. Percentages of Respondents for the Too Little, JAR and Too Much Levels of Hawaiian Pizza A

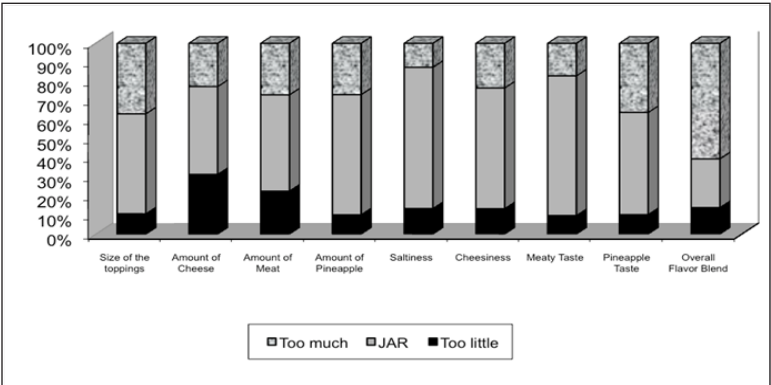


Figure 9. Percentages of Respondents for the Too Little, JAR and Too Much Levels of Hawaiian Pizza B

Figure 10 shows the penalty analysis plot for Hawaiian Pizza A. It can be observed that overall flavor blend, cheesiness, amount of cheese and meaty taste obtained positive mean drops in the overall liking score. This means that more than 20% of the respondents liked more the product attributes stated above for being in the too strong level compared to the respondents who rated those attributes in the JAR level.

Based on the Figure 11, the most “troublesome” attribute is the “too little” amount of meat. 33% of the respondents considered Hawaiian Pizza A to be having too little amount of meat with a mean drop of 1.23 from the overall liking score. Other attributes that are responsible for negative mean drops are the too little amount of pineapple, too weak meaty taste, too much amount of meat and too strong pineapple taste.

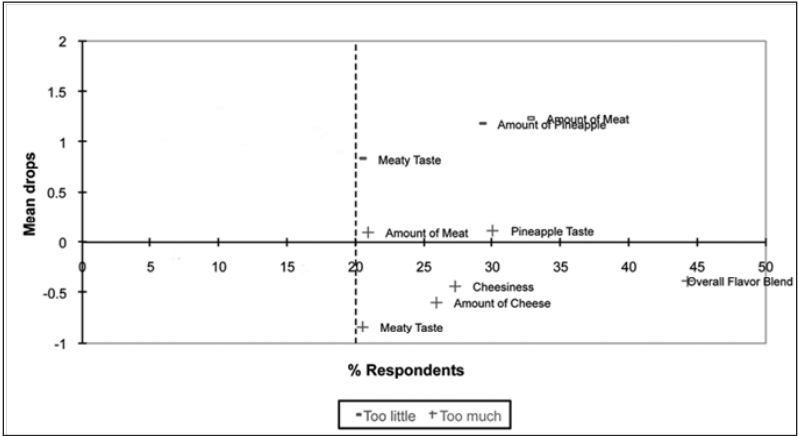


Figure 10. Penalty Analysis Plot for Hawaiian Pizza A

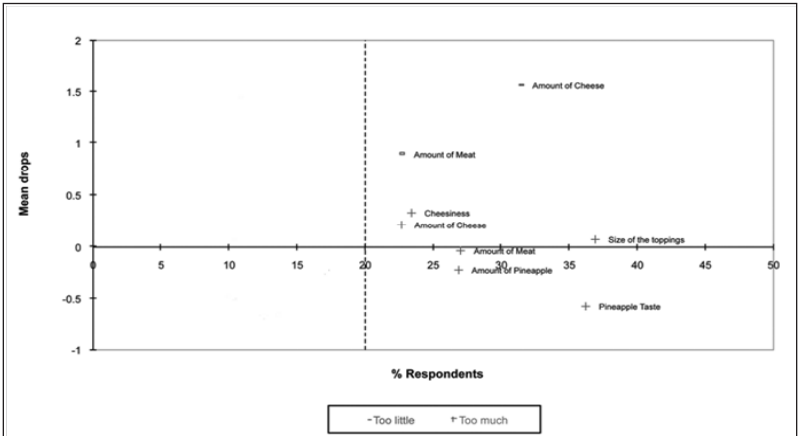


Figure 11. Penalty Analysis Plot for Hawaiian Pizza B

Figure 11 shows the penalty analysis plot for Hawaiian Pizza B. More than 20% of the respondents rated the product as having too strong pineapple taste, too much amount of meat and too much amount of pineapple. These attributes are the ones responsible for the positive mean drops on the overall liking score. The “too little” amount of cheese, on the other hand, was found to be the most troublesome attribute by 31% of the respondents and was responsible for a mean drop of 1.56. Other attributes causing negative mean drops are the too little amount of meat, too strong cheesiness, too much amount of cheese and too big size of the toppings.

T-test was done to determine the significance of the mean drops for each attribute. Tables 7 and 8 summarized the results of bootstrapping penalty analysis for Hawaiian Pizza A and Hawaiian Pizza B respectively. The results include the % of respondents for the non-JAR level, bootstrap mean drop, adjusted bootstrap mean drop, bootstrap standard error and calculated *t*.

Table 7. Results of Bootstrapping Penalty Analysis for Hawaiian Pizza A

Attributes	% Respondents	Mean Drop (Bootstrap)	Adjusted Mean Drop (Bootstrap)	Standard Error (Bootstrap)	t
Size of the Toppings Definitely too small Definitely too big	5.77 16.13				
Amount of Cheese Definitely too little Definitely too much	25.90	0.06	-1.25	0.41	1.31
Amount of Meat Definitely too little Definitely too much	32.67 20.90	0.01 -0.01	2.25 0.20	0.41 0.50	-2.44* -0.20
Amount of Pineapple Definitely too little Definitely too much	29.10 6.67	-0.03	2.38	0.35	-2.41*
Saltiness Not salty at all Definitely too salty	0 17.50				
Cheesiness Definitely too weak Definitely too strong	18.67 27.27	-0.02	-0.86	0.41	0.84
Meaty Taste Definitely too weak Definitely too strong	20.37 20.50	0.01 -0.02	1.66 -1.65	0.51 0.53	-1.65 1.63
Pineapple Taste Definitely too weak Definitely too strong	9.87 30.00	0.07	0.15	0.39	0.08
Overall Flavor Blend Definitely too weak Definitely too strong	11.70 44.27	0.01	-0.78	0.37	0.79

*significant at $\alpha = 0.05$

Based on Table 7, the respondents strongly penalized Hawaiian Pizza A when they perceived the product as having definitely too little amount of meat and definitely too little amount of pineapple. The tests are significantly different at 5% level of significance and the overall liking score is reduced by 2.25 for the amount of meat and 2.38 for the amount of pineapple. Increasing the amount of meat and the amount of pineapple may cause an increase in overall liking of the product. The penalty test is not significant when the respondents perceived Hawaiian Pizza A as having definitely too much amount of cheese, definitely too much meat, definitely too much pineapple, definitely too strong cheesiness, definitely too strong pineapple taste and definitely too strong overall flavor blend. For all the other attributes, the penalty test was not computed since the percentages of respondents who rated the product in the non-JAR level is less than 20% (Xiong et al., 2007).

Table 8. Results of Bootstrapping Penalty Analysis for Hawaiian Pizza B

Attributes	% Respondents	Mean Drop (Bootstrap)	Adjusted Mean Drop (Bootstrap)	Standard Error (Bootstrap)	t
Size of the Toppings					
Definitely too small	10.73				
Definitely too big	36.90	-0.07	0.22	0.42	0.64
Amount of Cheese					
Definitely too little	31.33	-0.02	3.14	0.44	-2.21*
Definitely too much	22.67	-0.06	0.49	0.53	1.31
Amount of Meat					
Definitely too little	22.60	-0.08	1.87	0.55	-2.44*
Definitely too much	27.00	-0.04	-0.03	0.49	-0.20
Amount of Pineapple					
Definitely too little	10.23				
Definitely too much	26.87	-0.01	-0.42	0.45	0.14
Saltiness					
Not salty at all	13.47				
Definitely too salty	12.57				
Cheesiness					
Definitely too weak	13.40				
Definitely too strong	23.40	0.01	0.66	0.52	0.84
Meaty Taste					
Definitely too weak	9.80				
Definitely too strong	17.03				
Pineapple Taste					
Definitely too weak	10.30				
Definitely too strong	36.20	-0.03	-1.12	0.47	0.08
Overall Flavor Blend					
Definitely too weak	13.97				
Definitely too strong	60.50	0.05	-0.58	0.48	0.79

*significant at $\alpha = 0.05$

For Hawaiian Pizza B (Table 8), the respondents significantly penalized the product when found to be having definitely too little amount of cheese and definitely too little amount of meat. The drop in the overall liking when not judged to be at JAR level is 2.21 for the amount of cheese and 2.44 for the amount of meat. Increasing both the amount of cheese and the amount of meat may cause an increase in the overall liking score of Hawaiian Pizza B. Penalty test is not significant when the respondents found the product to be having definitely too big toppings, definitely too much cheese, definitely too much meat, definitely too much pineapple, definitely too strong cheesiness, definitely too strong pineapple taste and definitely too strong overall flavor blend. The mean drops were not obtained for all the remaining attributes because the percentages of respondents were lower than 20%.

4.4 Comparison of ordinary and bootstrapping penalty analysis

Tables 9 and 10 show the comparison of the results of ordinary penalty analysis and bootstrapping penalty analysis for Hawaiian Pizza A and Hawaiian Pizza B respectively. The results include the percentage of respondents for the non-JAR level, mean drop and adjusted bootstrap mean drop.

Based on the results of ordinary penalty analysis (Table 9), the respondents penalized Hawaiian Pizza A when they rated the product as having definitely too little amount of meat, definitely too little amount of pineapple, definitely too weak meaty taste, definitely too strong pineapple taste and definitely too much amount of meat. When penalty analysis is combined with bootstrap method, the respondents penalized the same attributes. However, only the “definitely too little” amount of meat and “definitely too little” amount of pineapple were found to have significant effect on the overall liking score of Hawaiian Pizza A. For Hawaiian B (Table 10), the respondents penalized the product for having definitely too little amount of cheese, definitely too little amount of meat, definitely too strong cheesiness, definitely too much amount of cheese, and definitely too big size of the toppings. However, results of bootstrapping penalty analysis showed that the only significant attributes are “definitely too little” amount of cheese and “definitely too little” amount of meat.

Based on Tables 9 and 10, it can also be observed that the percentages of respondents for the non-JAR levels are almost the same for ordinary penalty analysis and bootstrapping penalty analysis. Also, it can be noted that results obtained for the adjusted bootstrap mean drops are almost twice as the value of the mean drops. This means that when penalty analysis is used together with bootstrap method, the amount of mean drops from the overall liking score is higher compared to an ordinary penalty analysis.

**Table 9. Comparison of Ordinary and Bootstrapping Penalty Analysis
for Hawaiian Pizza A**

Attributes	% Respondents (Ordinary PA)	% Respondents (Bootstrapping PA)	Mean Drop	Adjusted Mean Drop (Bootstrap)
Size of the Toppings				
Definitely too small	3.57	5.77		
Definitely too big	17.86	16.13		
Amount of Cheese				
Definitely too little	17.86	17.00		
Definitely too much	28.57	25.90	-0.60	-1.25
Amount of Meat				
Definitely too little	32.14	32.67	1.23	2.25*
Definitely too much	21.43	20.90	0.10	0.20
Amount of Pineapple				
Definitely too little	28.57	29.10	1.18	2.38*
Definitely too much	7.14	6.67		
Saltiness				
Not salty at all	0	0		
Definitely too salty	17.86	17.50		
Cheesiness				
Definitely too weak	21.43	18.67		
Definitely too strong	25.00	27.27	-0.44	-0.86
Meaty Taste				
Definitely too weak	21.43	20.37	0.83	1.66
Definitely too strong	21.43	20.50	-0.83	-1.65
Pineapple Taste				
Definitely too weak	10.71	9.87		
Definitely too strong	28.57	30.00	0.11	0.15
Overall Flavor Blend				
Definitely too weak	14.29	11.70		
Definitely too strong	42.86	44.27	-0.39	-0.78

*significant at $\alpha = 0.05$

**Table 10. Comparison of Ordinary and Bootstrapping Penalty Analysis
for Hawaiian Pizza B**

Attributes	% Respondents (Ordinary PA)	% Respondents (Bootstrapping PA)	Mean Drop	Adjusted Mean Drop (Bootstrap)
Size of the Toppings				
Definitely too small	10.00	10.73		
Definitely too big	36.67	36.90	0.07	0.22
Amount of Cheese				
Definitely too little	30.00	31.33	1.56	3.14*
Definitely too much	23.33	22.67	0.21	0.49
Amount of Meat				
Definitely too little	23.33	22.60	0.90	1.87*
Definitely too much	26.67	27.00	-0.03	-0.03
Amount of Pineapple				
Definitely too little	10.00	10.23		
Definitely too much	26.67	26.87	-0.22	-0.42

Saltiness				
Not salty at all	13.33	13.47		
Definitely too salty	13.33	12.57		
Cheesiness				
Definitely too weak	13.33	13.40		
Definitely too strong	23.33	23.40	0.33	0.66
Meaty Taste				
Definitely too weak	10.00	9.80		
Definitely too strong	16.67	17.03		
Pineapple Taste				
Definitely too weak	10.00	10.30		
Definitely too strong	36.67	36.20	-0.57	-1.12
Overall Flavor Blend				
Definitely too weak	13.33	13.97		
Definitely too strong	60.00	60.50	-0.26	-0.58

*significant at $\alpha = 0.05$

4.5 Implications on product development

Penalty analysis combined with bootstrap method is a great way to allow the product developers decide as to what sensory attributes should be improved first in order to increase overall liking score.

Based on the results of bootstrapping penalty test, Hawaiian Pizza A was significantly penalized by the respondents for having too little amount of meat and too little amount of pineapple. Since the mean drop is larger for the amount of pineapple, the product developer may increase this first and check if the overall liking score will improve. The developer may also increase the amount of both the meat and pineapple simultaneously. Hawaiian Pizza B, on the other hand, was significantly penalized for having too little amount of cheese and too little amount of meat. Increasing the amount of these ingredients may improve the overall liking score of the product. The developer may also decide to prioritize increasing the amount of cheese first since it has a larger mean drop compared to the amount of meat. However, it should be noted that increasing the amount of any of these ingredients would entail additional cost on the part of the owner.

Adjusting attributes that significantly penalized Hawaiian Pizza A and Hawaiian B may contribute to an increase in overall liking score. However, this is not a guarantee that the hedonic overall liking score will increase by the same amount of mean drops when the attributes not in the JAR levels are corrected. They simply suggest what attributes should be prioritized when improving the product (Rothman, 2007). Caution must be likewise taken as the change in one attribute can cause a significant change in the other attributes.

To determine which of the two pizza products should be improved, commercialized and launched, the results of the overall acceptability percentage, preference test and bootstrapping penalty analysis were compared. Since Hawaiian Pizza B obtained a higher overall acceptability rating and is not significantly preferred over Hawaiian Pizza A, the product developer may focus first on improving attributes that significantly penalized Hawaiian Pizza B.

5. Conclusions and Recommendations

Ordinary penalty analysis and bootstrapping penalty analysis were done on the sensory evaluation data of Hawaiian Pizza A and Hawaiian Pizza B. The respondents penalized Hawaiian Pizza A for having too much amount of cheese, too much of meat, too much pineapple, too strong cheesiness, too strong pineapple taste and too strong overall flavor blend. However, results of bootstrapping penalty analysis showed that only the “too little amount of meat” and “too little amount of pineapple” have significant mean drops on the overall liking. On the other hand, Hawaiian Pizza B was penalized for having too big size of the toppings, too little amount of cheese, too much amount of cheese, too little amount of meat, too much amount of meat, too much amount of pineapple, too strong cheesiness, too strong pineapple taste and too strong overall flavor blend. The only significantly penalized attributes are the “too little amount of cheese” and “too little amount of meat.”

Based on the results of this study, it can be concluded that bootstrap test could be used together with penalty analysis to allow statistical testing of the mean drops. It is then recommended to adjust the level of the attributes that were significantly penalized by the respondents and to conduct another in-house sensory evaluation. Bootstrapping penalty analysis can be applied again to determine if there would be an improvement in the overall liking score and if the mean drops would be corrected. It is also recommended to focus on improving Hawaiian Pizza B over Hawaiian Pizza A since it obtained a higher overall acceptability percentage.

References

- AMERINE, M., PANGBORN, R. and ROESSLER, E., 1965, *Principles of Sensory Evaluation of Food*, New York: Academic Press.
- ANON, 2003, Triangle plots: Graphic display of “just right” scale data, *Research on Research* 56: 1-6.
- BAOSHENG, G., 2005, *Product Testing: Handle with Care*, China: IPSOS Insight.
- CHAMBERS, E., IV and WOLF, M., 1996, Sensory testing methods, *ASTM Manual Series: MNL 26*, West Conshohocken, PA: ASTM International.
- CHERNICK, M., 1999, *Bootstrap Methods: A Practitioner Guide*, New York: John Wiley & Sons, Inc.
- DAVIDSON, A., HINKLEY, D. and SCHECHTMAN, E., 1986, Efficient bootstrap simulation, *Biometrika* 73(3): 555-566.
- DIACONIS, P. and EFRON, B., 1983, Computer-intensive methods in statistics, *Scientific American* 248 (5): 116-130.
- EFRON, B. and GONG, G., 1983, A leisurely look at the bootstrap: The jackknife and cross validation, *The American Statistician* 37: 36-48.
- EFRON, B. and TIBSHIRANI, R., 1993, *An Introduction to the Bootstrap*, New York: Chapman and Hall, Inc.

- FAN, X., 1994, *Does Bootstrap Procedure provide Biased Estimates? An Empirical Examination for a Case of Multiple Regression*, East Lansing, MI: National Center for Research on Teacher Learning.
- GATCHALIAN, M., 1989, *Sensory Evaluation Methods for Quality Assessment and Development*, University of the Philippines, Quezon City.
- GATCHALIAN, M. and BRANNAN, G., 2009, *Sensory Quality Measurement: Statistical Analysis of Human Responses*, Quezon City, Philippines: Quality Partners Company, Ltd.
- JONES, L., PERYAM, D. and THURSTONE, L., 1955, Development of a scale for measuring soldiers' food preferences, *Food Research* 20: 512-520.
- KEMP, S., HOLLOWOOD, T. and HORT, J., 2009, *Sensory Evaluation: A Practical Handbook*, West Sussex, UK: John Wiley & Sons, Ltd., Publication.
- LAWLESS, H. and HEYMANN, H., 1999, *Sensory Evaluation of Food: Principles and Practices*, Gaithersburg, Maryland: Aspen Publishers, Inc.
- LEVINE, D. and STEPHAN, D., 2010, *Even You Can Learn Statistics: A Guide for Everyone who has ever been Afraid of Statistics*, Upper Saddle. River, NJ: Pearson Education, Inc.
- MEILGAARD, M., CIVILLE, G. and CARR, B., 1999, *Sensory Evaluation Techniques*, 2nd ed., Boca Raton, FL: CRC Press.
- MEULLENET, J., XIONG, R. and FINDLAY, C., 2007, *Multivariate and Probabilistic Analyses of Sensory Science Problems*, 1st ed., Iowa, USA: Blackwell Publishing.
- MASON, R. and NOTTINGHAM, S., 2002, *Sensory Evaluation Manual*, Workshop at Naresuan University, Phitsanulok, Thailand, Available at: <http://www.scribd.com/doc/8940001/Sensory-Evaluation-Manual#scribd>
- PACZKOWSKI, W., 2009, *Technical Memorandum on Penalty Analysis*. Data Analytics Corp.
- PERYAM, D.R. and PILGRIM, P.J., 1957, Hedonic scale method for measuring food preferences, *Food Technology* 11:9–14.
- PLAEHN, D., 2009, *Understanding Penalty Analysis*, Insights Now, Inc.
- PLAEHN, D. and HORNE, J., 2008, A regression-based approach for testing significance of “just-about-right” variable penalties, *Food Quality and Preference* 19: 21-32.
- PRELL, P., 1976, Preparation of reports and manuscripts which include sensory evaluation data, *Food Technology*.
- ROTHMAN L., 2007, *The Use of Just-About-Right (JAR) Scales in Food Product Development and Reformulation in Consumer-Led Food Product Development*, Boston: CRC Press.
- SCHMIDHEINY, K., 2010, *The Bootstrap: Short Guides to Microeconometrics*, Universitat Pompeu Fabra.
- STINE, R., 1989, An introduction to bootstrap methods, *Sociological Methods and Research* 18:243-291.
- STONE, H. and SIDEL, J., 2004, *Sensory Evaluation Practices*, 3rd ed., London/New York: Academic Press/Elsevier.
- TIBSHIRANI, R., 1985, *How Many Bootstraps?* Department of Statistics, Stanford University Technical Support. Stanford, California.

- WHISTON S., 2009, *Principles and Applications of Assessment in Counseling*. 3rd ed., Belmont, CA: Brooks/Cole, Cengage Learning.
- XIONG, R., MEULLENET, J. and FINDLAY, C., 2007, *Multivariate and Probabilistic Analyses of Sensory Science Problems*, Blackwell Publishing, 207-212.
- XIONG, R. and MEULLENET, J., 2009, Bootstrapping penalty analysis, *ASTM International* 63-66.

Incidence of Crimes and Effectiveness of Interventions in the National Capital Region: Evidence from Panel Data

Michael Daniel C. Lucagbo

Lianne S. De la Cruz

Jecca V. Narvasa

Micah Jane A. Paglicawan

Jerizza P. Pepito

School of Statistics

University of the Philippines Diliman

Efforts to bring down the incidence of crimes have been intensified by the Philippine National Police (PNP). The index crimes are prioritized among these crimes. The index crimes include theft, robbery, carnapping, and motornapping. Interventions to bring down the incidence of crimes have recently been enacted by the National Capital Region Police Office (NCRPO) of the PNP. These interventions include increases in number of police personnel, mobile patrols, beat patrols, and checkpoints. In this study, the effect of each of these interventions is examined in a panel data analysis using weekly data gathered from all of the police stations in NCR. This paper performs a district-level analysis of the crimes and interventions. The negative binomial regression model for panel data is used to quantify the effects of the interventions on the incidence of index crimes. Results show that some (but not all) of these interventions are effective in reducing crime. The results also show differences in the effects of the interventions across the different districts. Resources should thus be redirected towards these effective strategies. The differences in the effects of the interventions among the different crimes are also studied.

Keywords: index crimes, intervention, panel data, negative binomial regression

1. Introduction

The safety of one's person and security of one's property are widely viewed as basic human rights and are essential to the community's overall quality of life

(Plant and Scott, 2009). People envision a society in which they play an active role in peace and development, engaging in the promotion of social integration as they seek unity in diversity with social justice. However, it has been a global phenomenon that crime is on the rise. Crime has had great impact on individuals and communities and has high and diverse costs. There is terrible physical, material, mental, and emotional injury that is suffered by its victims.

The Philippines has not been spared from the daily occurrence of crimes, especially in metropolis areas like the National Capital Region (NCR). Kiunisala et al. (2004) observe that the rising tide of criminality has disrupted the normal functions of government entities in the region.

The Philippine National Police (PNP) has enacted a set of interventions to address this problem. These interventions, described in detail in this paper, were enacted starting June 2014. This paper aims to determine which interventions are effective and the police districts where they should be deployed. The researchers also look at the trends in crimes in the different districts in NCR, and fit a statistical model that establishes the relationship between the crimes and interventions.

1.1 Significance of the study

Understanding the effects of the interventions implemented by the PNP to reduce crime will provide useful information on improving the system and policy strategies of the PNP. Recognizing the real effects of the interventions results in better risk-management practices and proper allocation and accounting of budget to optimize the reduction of crimes or any unwanted instances, as emphasized by Kelling (2009) in an article assessing the case of the Fire Department of New York.

Argona and Parcon (2004) adopt the hypothesis of Pyle that within general groups of crimes, deterrent effects differ quite substantially from one group to another. The researchers have tested the appropriateness of the theory in the Philippine setting. Identifying the right interventions to reduce a specific crime is premised on the relevance of Pyle's Theorem in the Philippine setting.

One of the benefits of a safe and secure environment is that it encourages investment and economic growth. The Philippine Development Plan (PDP) 2011-2016 recognizes this and therefore argues that crime reduction efforts pave way to promote and sustain public order and internal stability (SEPO, 2013). There is accumulating evidence that many prevention programmes are not only effective in reducing offending and victimization but are also cost-beneficial, resulting in considerable long-term savings, and bringing social and economic benefits well beyond reductions in crime (Shaw and Travers, 2007).

Identifying which interventions significantly contribute to the reduction of crime in the NCR police districts will help policymakers decide on which policies to focus on. These attempts have been made to improve community security throughout the NCR, and eventually throughout the country.

1.2 Scope and limitations

The main objective of this paper is to analyze the effects of eight interventions implemented by the National Capital Region Police Office (NCRPO) of the PNP to reduce robbery, theft, carjacking and “motorjacking” (for two-wheel vehicles), the four index crimes (crimes which are serious in nature and occur with high frequency) considered in the study. Using a panel data fixed effects analysis, the researchers investigate this relationship in each of the five police districts in NCR. The study is based on the data collected by the PNP-NCRPO, from the first week of January 2014 to the last week of September 2014.

The eight interventions included in the study are the following: (1) number of checkpoints per station conducted in the vicinity, (2) number of *Oplan bakal* operations per station conducted in the vicinity, (3) number of mobile patrol operations per station conducted in the vicinity, (4) number of beat patrol operations per station conducted in the vicinity, (5) number of warrants of arrest served, (6) number of *Oplan katok* operations per station conducted in the vicinity, (7) number of *Oplan bulabog* operations per station conducted in the vicinity, and (8) number of additional police personnel per station.

Weisburd et al. (2011) have noted that the clustering of crime at places is very important in criminological inquiry and practical crime prevention. Sherman and Weisburd (1995) support place-specific “micro-deterrence” effects of police policies in reducing crime. These researches provide the motivation to perform district-level analyses to capture the micro-deterrence effects of the interventions.

2. Review of Related Literature

In what follows, previous works related to crimes and interventions are reviewed. Included in this review are studies about the most prevalent crimes in the Philippines, trends in these crimes, and the interventions used to combat them. It also contains researches investigating the relationship between interventions and index crime incidence.

The Overseas Security Advisory Council (OSAC) identifies crime as a significant concern in urban areas of the Philippines (OSAC, 2014). According to the PNP Directorate for Investigation and Detective Management, crimes involving theft, physical assault, and robbery were the top three most common crimes reported to local authorities in 2013. Due to lack of response vehicles, radios, and other essential equipment, the PNP is limited in its capacity to respond and assist victims of crime.

Understanding the human sociological and behavioral condition facilitates the explanation of the reasons behind criminality. Pepa (2013) argues that, in the context of Philippine society, poverty and abuse of power are some of the causes of crimes. Very often, crimes are committed in areas where the urban poor live. Poverty may not strictly be identified as a direct cause of crimes, but certain circumstances that result from a situation of poverty can cause crimes. Another

major cause of crime in the country is the abuse of power by the people who hold position in the society (Pepa, 2013).

The National Crime Prevention Framework produced by Australian Institute of Criminology (AIC) in 2012 suggests that crime prevention strategies be directed towards:

- Addressing the environmental conditions that promote and sustain crime;
- Eliminating risk factors and enhancing protective factors to reduce the likelihood that individuals will engage in offending behavior;
- Strengthening communities by addressing social exclusion and promoting community cohesiveness; and
- Enhancing the capacity of criminal justice agencies to prevent crime and reoffending.

Urban crime prevention strategies include both social crime prevention and situational crime prevention policies. Social crime prevention measures are those which deal with the fundamental causes of crime or the criminogenic factors of crime. On the other hand, situational crime prevention measures, sometimes called defensive prevention measures, are those intended to reduce the opportunities for committing crime, attempting to increase the risk and difficulties of crime through environmental design and management (Sanidad-Leones, n.d.). Efforts done by the country's law enforcers can be divided into three main parts: (1) holistic system approach of the criminal justice system, (2) community-based policing system, and supplemented by (3) international technical cooperation projects (Sanidad-Leones, n.d.).

In this study, the crime prevention interventions examined are described. Two of these, the number of warrants served and number of additional police personnel, are, in a strict sense, not interventions. The former is a measure of the effectiveness of police service in ensuring the arrest of criminals, while the latter simply counts the number of additional police personnel deployed in the police districts. Nonetheless, these two, although not new operations conducted by the PNP, will still be referred to as "interventions." The definitions of the interventions given in this paper are from the PNP Standard Operating Procedures (SOP) in Managing Police Operations.

Police checkpoints

Checkpoints are important in maintaining public safety and reducing criminality in the streets. PNP (2015) describes checkpoints as "the conduct of at least 2 hours police check of vehicular and/or pedestrian traffic by at least 8-man complement in a place normally considered chokepoint or crime hotspot in order to enforce control measures, laws, ordinances and other regulations; police stations shall be required to organize a minimum of 3 teams and conduct simultaneous checkpoints within 8 or 12-hour shift." Moreover, checkpoints must

not cause inconvenience nor intimidate citizens. Instead, they should provide a real sense of safety and security (PNP, 2013).

Fell et al. (2008) and Lacey et al. (1999) assert that sobriety checkpoints not only result in arrests for the driving-while-intoxicated (DWI), but also yield apprehensions for stolen vehicles, illegal firearms, outstanding warrants, and drug violations (Fell, 2013). The strategy of using high-visibility traffic enforcement, where many drivers experience or see the enforcement activity (such as the use of sobriety checkpoints), apparently not only raises perceived probability of apprehension for the DWI, but may also raise the perceived risk of being arrested for other criminal activities (Fell, 2013).

Oplan Bakal operations

Oplan bakal operations are defined by PNP (2015) as the “conduct of spot checks on persons at bar joints and/or areas considered crime prone or hotspots in order to confiscate loose firearms and deadly weapons and prevent them from being used in crimes. It further aims to arrest or neutralize would-be motorcycle riding and walking criminals.”

Mobile patrols

A mobile patrol is described by PNP (2015) as “one or more police officers (usually 2 or buddy system) conducting patrol within designated sector using land based conveyance such as bicycles, motorcycles and motor vehicles in order to maintain order, prevent crimes and enforce laws and local ordinances. The number of operations shall be based on the number of mobile patrols per 8 or 12-hour shift.” Because of their increased efficiency of coverage, car patrols are sometimes preferred over foot patrols. However, some studies (e.g. Houghton Mifflin Harcourt, 2014) prefer foot patrols since car patrols reduce police contacts with citizens. There are three purposes of a patrol: to answer calls, to maintain police presence to deter crime, and to probe suspicious circumstances. Among these three, it is the second one, maintaining police presence, which is the most contentious. The argument for the importance of patrols is that police cars cruising randomly through city streets supposedly create the feeling that the police are everywhere.

Beat patrols

Police beats ensure that the policing needs of local areas are met with a service customized to each individual locality. PNP (2015) defines a beat patrol as “one or two police officers (usually 2 or buddy system) moving about on foot within the designated beat in order to maintain order, prevent crimes and enforce laws and local ordinances. The number of beat patrol operations shall be based on the given number of beats per 8 or 12-hour shift.” Beat policing is a strategy designed to make individual police officers responsible for the community’s policing needs

in a defined geographical area (Queensland Police, 2014). Beat patrols are also commonly known as “foot patrols.” The theoretical expectation of an increased police presence would be crime reductions as a result of an increase in perceived risk by offenders (Andresen and Malleon, 2014).

Number of warrants served

A warrant is a legal process issued by a competent authority, directing the arrest of a person or persons upon grounds stated therein (Philippine Law Library, 2015). In the first half of 2014, 562 warrants of arrest were served, resulting to the arrest of 410 suspects (Aquino Jr., 2014). Many studies have supported the “deterrence hypothesis,” or the idea that crime can be deterred if changes in the costs and benefits of crime are made. For example, Tittle and Rowe (1974) analyze certainty of arrest and crime data. Their study supports the deterrence argument as long as certainty of punishment reaches a critical level. Bailey (1976) also lends general support to the deterrence argument but shows that “the effect of arrest is not uniform for each offense, with different levels of arrest (certainty of apprehension) required to reduce rates for different crimes.”

Oplan Katok operations

Oplan Katok operations are defined by PNP (2015) as “physical verification and accounting of registered firearms through house visitation of known address indicated by the licensed firearm owners/holders. It aims to prevent or address the proliferation of loose firearms and ensures that the licensed firearm holders are residents of the addresses indicated in the data of Firearms and Explosives Office (FEO). Police stations shall be required to conduct a minimum of 40 “katok” operations or house visitations or verification of registered firearms per day.”

A number of house visitations have been conducted by the PNP as part of their effort to eliminate, as much as possible, the use of unlicensed firearms in criminal activities (Purisima, 2013). Many researchers have evinced the significant relationship between safe gun strategies and reduction of crimes (e.g. Sherman, et al., 1996; Kates, 1990). The security risk produced by irresponsible gun ownership and use is one of the major concerns of safe gun strategies.

Oplan Bulabog operations

The PNP (2015) defines Oplan Bulabog operations as the “conduct of saturation drive, spot checks and inspection at the barangays especially in crime prone places targeting unregistered, undocumented and stolen motorcycles.”

Number of additional police personnel

The analysis of Marvell and Moody (1996) affirms the hypothesis that the prevention benefits of hiring more police officers are greater in higher-crime cities than in places where crimes are less prevalent. One major theory about the

crime prevention benefits of hiring more officers is that it reduces police response time. The initial studies of the response time hypothesis produce strong evidence suggesting that a decrease in police response time could lead to the arrest of many more offenders.

Another major theory about the benefits of a larger police force is that it enables police officers to conduct more random patrols. The analysis of Koper (1995) of the Minneapolis Hot Spots Patrol data found a very strong relationship between the length of each police patrol presence (which averaged 14 minutes) and the amount of time the hot spot was free of crime after the police left the scene.

3. Coverage and Variables Used

The data on crimes and interventions come from the PNP National Capital Region Police Office (PNP-NCRPO). The data are given for each police station in all of the five police districts in NCR: Northern Police District (NPD), Eastern Police District (EPD), Manila Police District (MPD), Southern Police District (SPD), and the Quezon City Police District (QCPD).

Table 3.1 shows the police stations covered by the jurisdiction of each district. There are a total of 38 police stations involved in the study. Weekly data on the incidence of the index crimes and the interventions are collected from these police stations.

Table 3.1. PNP NCRPO Districts and Stations

District 1 NPD	District 2 EPD	District 3 MPD	District 4 SPD	District 5 QCPD
Caloocan Malabon Navotas Valenzuela	Marikina Pasig Mandaluyong San Juan	RaxaBago Tondo Sampaloc Sta. Cruz Ermita Sta. Ana J.A. Santos Sta. Mesa Malate Pandacan Meisic	Pasay Makati Paranaque Las Pinas Muntinlupa Taguig Pateros	La Loma Baler Talipapa Novaliches Lagro Batasan Cubao Project 4 Anonas Kamuning Galas Eastwood

Table 3.2 shows the index crimes involved in this study. The table also gives their definition and the way they are measured. These index crimes are the dependent variables in the negative binomial regression runs. The definitions are taken from Bragado (2007).

Table 3.2. Definitions of the Index Crimes Included in the Study

Crime	Definition	Measurement
Robbery	The intent to gain, take any personal property belonging to another, by means of violence or intimidation of any person, or using force upon anything.	Number of reported robbery cases per station
Theft	The intent to gain, take any personal property belonging to another without the consent or knowledge of the owner and without the use of violence against or intimidation of persons or force upon things.	Number of reported theft cases per station
Carnapping	The unlawful taking, with intent to gain, a four-wheel vehicle belonging to another without the latter's consent or by means of violence against or intimidation of person or using force upon things.	Number of reported carnapping cases per station
Motornapping	The unlawful taking, with intent to gain, a two-wheel vehicle belonging to another without the latter's consent or by means of violence against or intimidation of person or using force upon things.	Number of reported motornapping cases per station

Table 3.3 shows the labels and descriptions of the predictors, the eight interventions enacted by the PNP. These are the interventions being enforced by the PNP in its efforts to reduce the incidence of the index crimes.

Table 3.3. Description of the Interventions Enacted by the PNP

Interventions	Description
Police Checkpoints	Number of police checkpoints in station's vicinity
Oplan Bakal	Number of Oplan Bakal operations conducted in station's vicinity
Mobile patrols	Number of mobile patrol operations conducted in station's vicinity
Beat patrols	Number of beat patrol operations conducted in station's vicinity
Warrants served	Number of warrants served per station
Oplan Katok	Number of door-to-door visits in station's vicinity
Oplan Bulabog	Number of Oplan Bulabog operations conducted in station's vicinity
Additional Police	Number of additional police personnel per district/station

The dataset used in this research consists of weekly measurements collected by the PNP-NCRPO for all the 38 police stations in NCR. Since each police station is followed every week for a total of 39 weeks, the entire dataset is a panel dataset.

4. Methodology

The researchers use the fixed-effects negative binomial regression model for panel data to describe the effects of the interventions on the incidence of crimes. The model is introduced by Hausman et al. (1984), who first applied it to study the relationship between patents and R&D expenditures. Following the discussion of

Hausman et al. (1984), the parameter λ_{it} is assumed to follow a Gamma distribution with parameters γ_{it} and δ . Moreover, $\gamma_{it} = e^{X_{it}\beta}$ where X_{it} contains the measurements on the independent variables, and δ is a term common to all cross-sectional units and fixed across time periods. The mean and variance of λ_{it} are then $E(\lambda_{it}) = e^{X_{it}\beta} / \delta$ and $Var(\lambda_{it}) = e^{X_{it}\beta} / \delta^2$. Taking the Gamma distribution for λ_{it} and integrating by parts,

$$\begin{aligned} \Pr(n_{it}) &= \int_0^\infty \frac{1}{n_{it}!} e^{-\lambda_{it}} \lambda_{it}^{n_{it}} f(\lambda_{it}) d\lambda_{it} \\ &= \frac{\Gamma(\gamma_{it} + n_{it})}{\Gamma(\gamma_{it})\Gamma(n_{it} + 1)} \left(\frac{\delta}{1 + \delta}\right)^{\gamma_{it}} (1 + \delta)^{-n_{it}} \end{aligned}$$

which is the negative binomial distribution with parameters γ_{it} and δ . Hausman et al. (1984) mention that computation of maximum likelihood estimates with the use of partial fraction expansions of the gamma and digamma functions permits rapid evaluation. The moments of n_{it} are $E(n_{it}) = e^{X_{it}\beta} / \delta$ and $Var(n_{it}) = e^{X_{it}\beta} (1 + \delta) / \delta^2$. Thus, as Hausman et al. (1984) argue, the negative binomial specification allows for overdispersion (i.e., the variance can be greater than the mean), with the Poisson a limiting case as $\delta \rightarrow \infty$.

5. Descriptive Analysis

Table 5.1 shows the number of reported crime cases per district for the time period covered by the study: December 30, 2013 to September 28, 2014. Among the five police districts, it is the QCPD which has the highest incidence of robbery, theft, and carnapping, and the highest incidence of all crimes combined. The district with the highest incidence of motornapping is SPD. Moreover, it is noteworthy to look at the relatively low levels of crime in EPD and NPD, both for total crimes and for the index crimes considered individually.

**Table 5.1. Reported Incidence of Index Crimes by Police District:
December 30, 2013 to September 28, 2014**

	Robbery	Theft	Carnapping	Motornapping	Total
NPD	728	1,053	68	177	2,026
EPD	623	1,019	93	164	1,899
MPD	2,443	3,451	96	320	6,310
SPD	2,029	4,589	213	426	7,257
QCPD	3,066	5,543	308	256	9,173

Table 5.2 shows the number of operations for some of the interventions hypothesized to reduce the level of crime in the Metro Manila districts. It is SPD

which leads all other districts in number of checkpoints, Oplan Bakal operations, mobile patrols, beat patrols, number of warrants served, and Oplan Bulabog operations. Worth mentioning is the observation that SPD has more than twice as many Oplan Bakal operations, mobile patrols, and Oplan Bulabog operations, as that of any of the other districts. QCPD, on the other hand, has the highest number of Oplan Katok operations, more than twice that of the other districts.

**Table 5.2. Number of Operations for the Interventions:
June 16, 2014 to September 28, 2014**

	Checkpoint	Bakal	Mobile	Beat	Warrant	Katok	Bulabog
NPD	3,750	4,777	12,888	19,146	421	9,456	348
EPD	3,139	4,648	12,400	14,317	409	11,452	70
MPD	6,188	1,527	4,276	17,921	313	20,522	611
SPD	9,640	12,425	31,032	36,005	634	17,006	1,942
QCPD	5,883	5,455	13,734	17,782	89	50,550	392

Figures 5.1 to 5.5 show the weekly incidence of the index crimes for all of the five districts. The figures display the trends in these index crimes, enabling comparisons across the five police districts. The districts generally show dissimilar patterns of the movements of the index crimes across time. Some of the highlights of these patterns are now described. The dates for the weeks covered by the study are given in the Appendix.

Figure 5.1 shows the weekly incidence of crime in NPD. The graph shows that the incidence of theft has been steadily increasing, and so is the incidence of robbery, although at a slower rate. The incidence of carnapping and motornapping appear to show no increasing trend. In the recent weeks, however, the incidence of motornapping has been slowly on the rise. Figure 5.2 shows the weekly incidence of crime in EPD. Once again, the incidence of theft has been increasing, this time at an even sharper rate than what has been observed in NPD. The incidence of robbery is also generally rising. Lastly, the incidences of carnapping and motornapping both fluctuate at a low level. The movement of motornapping, however, is less stable.

Figure 5.3 gives the weekly incidence of crime in MPD. None of the index crimes show any obvious trend, although one might argue that the incidence of theft has been slowly increasing in the latter weeks. Moreover, the sizes of the fluctuations of all four series are roughly constant. Figure 5.4 gives the weekly incidence of crime in SPD. Except for motornapping, which can be seen as having an increasing trend, all other crimes display decreases in number in the latter weeks (something which can be seen through more careful inspection). Visible from the plots is the substantial drop in the levels of theft and robbery in recent

weeks. Lastly, Figure 5.5 shows the weekly incidence of crime in QCPD. Robbery, and motornapping do not display any trends. The incidence of theft, however, has been slowly increasing, while carnapping has steadily dropped.

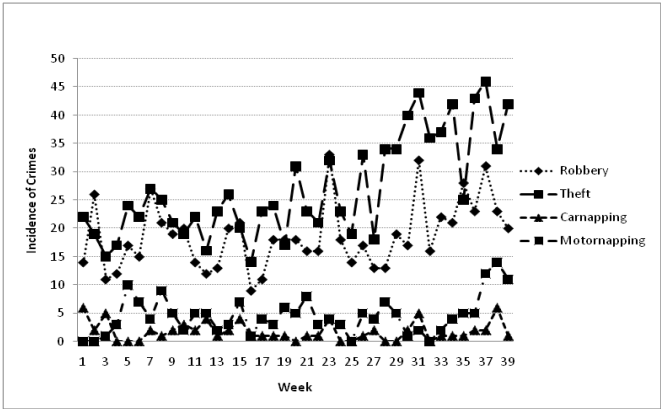


Figure 5.1. Weekly Incidence of Crime in Northern Police District: December 30, 2013 to September 28, 2014

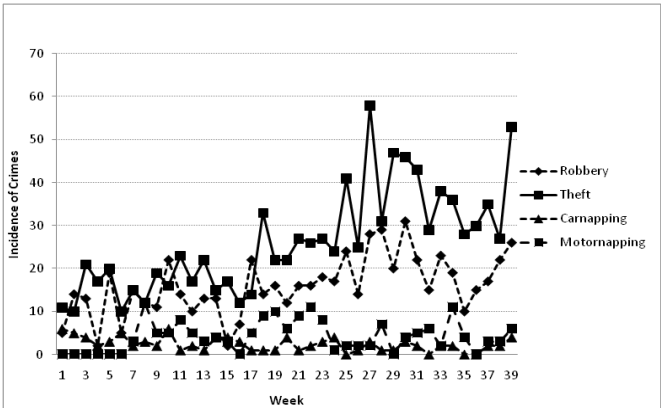


Figure 5.2. Weekly Incidence of Crime in Eastern Police District: December 30, 2013 to September 28, 2014

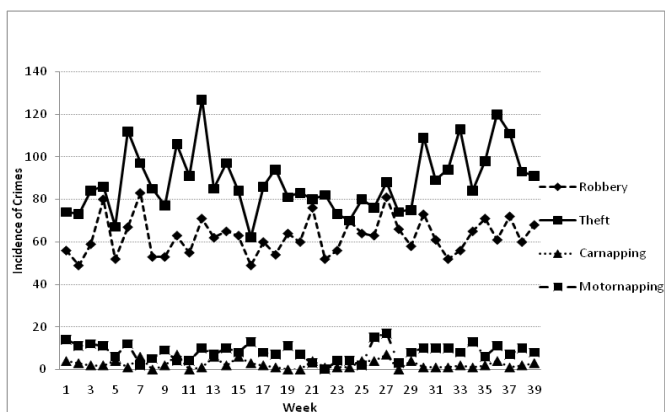


**Figure 5.3. Weekly Incidence of Crime in Manila Police District:
December 30, 2013 to September 28, 2014**

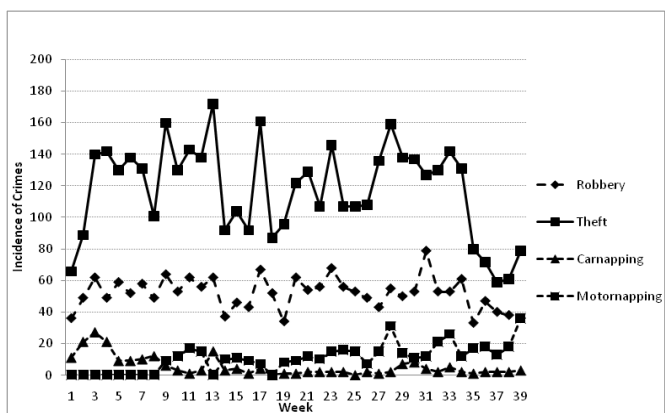


**Figure 5.4. Weekly Incidence of Crime in Southern Police District:
December 30, 2013 to September 28, 2014**

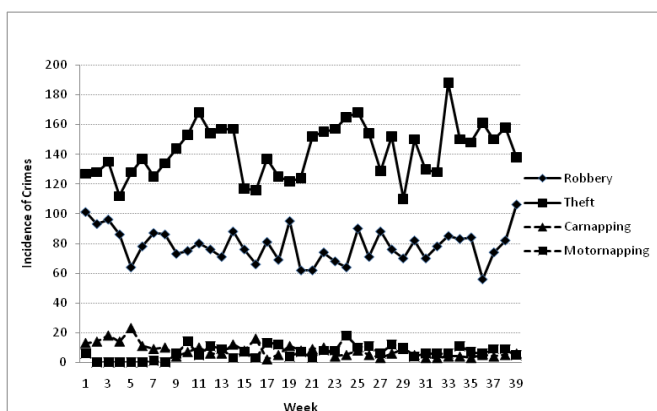


Figure 5.5. Weekly Incidence of Crime in Quezon City Police District: December 30, 2013 to September 28, 2014

6. Negative Binomial Regression Results

Tables 6.1 to 6.5 show the results of the fixed effects negative binomial regression runs for each of the five districts, with asterisks on those interventions which significantly decrease the level of crime. The standard errors of the estimates are enclosed in parentheses. The results of the regression run for NPD are shown in Table 6.1. The effect of mobile patrols is excluded because of multicollinearity issues. The table shows that the number of beat patrols is significant in bringing down the incidence of robbery and motornapping in NPD.

The results for EPD are given in Table 6.2. The results show that the number of mobile patrols is strongly significant in bringing down the level of motornapping in EPD. Meanwhile, the number of Oplan Bulabog operations is significant in reducing the incidence of robbery in EPD.

The results for MPD are shown in Table 6.3. Only the number of beat patrols appears to be significant in bringing down the incidence of crimes in MPD. In particular, it is significant in reducing the incidence of robbery. The SPD results, shown in Table 6.4, show that the number of Oplan Bakal operations, beat patrols, Oplan Katok operations, and additional policemen are significant in reducing some index crimes. In particular, the number of Oplan Bakal operations and beat patrols are significant depressants of robbery. On the other hand, the number of Oplan Katok operations and additional policemen are significant in reducing the incidence of theft.

Lastly, the results for QCPD are given in Table 6.5. The results show that the number of police checkpoints, Oplan bakal operations and mobile patrols are the interventions which are significant in bringing down the incidence of crimes. Increasing the number of police checkpoints brings down the incidence of motornapping, while Oplan Bakal operations and mobile patrols decrease the incidence of carnapping.

Table 6.1. Fixed Effects Negative Binomial Regression Results for Northern Police District

	Robbery	Theft	Carnapping	Motornapping
Police Checkpoints	0.00426 (0.00293)	0.00119 (0.00228)	0.01078 (0.01044)	0.00161 (0.00780)
Oplan Bakal	-0.00041 (0.00223)	0.00277 (0.00199)	0.00522 (0.00921)	0.01741 (0.00655)
Beat Patrols	-0.00159*** (0.00057)	-0.00026 (0.00048)	-0.00293 (0.00239)	-0.00467*** (0.00134)
Warrants Served	0.05573 (0.02251)	0.01946 (0.01898)	-0.07836 (0.10668)	0.00899 (0.04423)
Oplan Katok	-0.00026 (0.00048)	0.00021 (0.00040)	-0.00023 (0.00140)	-0.00004 (0.00117)
Oplan Bulabog	0.00515 (0.00500)	0.00266 (0.00410)	0.02454 (0.01426)	0.00093 (0.01166)
Additional Police	0.00409 (0.00219)	0.00275 (0.00204)	0.00666 (0.01105)	0.02437 (0.00533)
Constant	3.49297*** (0.89513)	2.76870*** (0.53095)	0.74874 (0.72349)	1.62805** (0.79262)

*Sig. at 10%; **Sig. at 5%; ***Sig. at 1%

Table 6.2. Fixed Effects Negative Binomial Regression Results for Eastern Police District

	Robbery	Theft	Carnapping	Motornapping
Police Checkpoints	0.00812 (0.00228)	0.00281 (0.00158)	0.00984 (0.00801)	0.01106 (0.00374)
Oplan Bakal	0.00121 (0.00173)	0.00279 (0.00129)	0.00161 (0.00537)	0.00171 (0.00343)
Mobile Patrols	0.00048 (0.00100)	0.00025 (0.00072)	-0.00534 (0.00441)	-0.01381*** (0.00336)
Beat Patrols	-0.00002 (0.00082)	0.00033 (0.00067)	0.00133 (0.00287)	0.00670 (0.00162)
Warrants Served	-0.00704 (0.01402)	0.00733 (0.00983)	-0.04329 (0.04804)	-0.00833 (0.03147)
Oplan Katok	-0.00015 (0.00053)	-0.00022 (0.00040)	-0.00135 (0.00185)	0.00367 (0.00110)
Oplan Bulabog	-0.12512** (0.05996)	-0.02061 (0.03528)	-0.07821 (0.19714)	0.03360 (0.07197)
Additional Police	0.00108 (0.00291)	0.00126 (0.00228)	0.01027 (0.00935)	0.01060 (0.00533)
Constant	1.47904*** (0.34145)	2.12561*** (0.33456)	0.37235 (0.49537)	-0.61740* (0.32189)

*Sig. at 10%; **Sig. at 5%; ***Sig. at 1%

**Table 6.3. Fixed Effects Negative Binomial Regression Results for
Manila Police District**

	Robbery	Theft	Carnapping	Motornapping
Police Checkpoints	0.00357 (0.00158)	0.00280 (0.00150)	0.00667 (0.00759)	0.00477 (0.00421)
Oplan Bakal	0.00007 (0.00329)	0.00161 (0.00303)	-0.00240 (0.01647)	0.00571 (0.00807)
Mobile Patrols	0.00059 (0.00132)	0.00031 (0.00115)	0.00608 (0.00570)	-0.00437 (0.00411)
Beat Patrols	-0.00118** (0.00052)	0.00004 (0.00047)	-0.00146 (0.00291)	-0.00153 (0.00154)
Warrants Served	-0.00701 (0.00989)	-0.01434 (0.00968)	0.00608 (0.08197)	0.01851 (0.02466)
Oplan Katok	-0.00007 (0.00016)	0.00016 (0.00013)	-0.00189 (0.00181)	0.00027 (0.00030)
Oplan Bulabog	-0.00466 (0.00303)	-0.00352 (0.00245)	-0.03936 (0.05475)	0.00570 (0.00507)
Additional Police	0.00530 (0.00242)	0.00192 (0.00231)	0.00374 (0.01251)	0.00868 (0.00662)
Constant	3.14723*** (0.36233)	2.53232*** (0.19443)	1.05586 (1.02662)	0.60498* (0.31561)

*Sig. at 10%; **Sig. at 5%; ***Sig. at 1%

**Table 6.4. Fixed Effects Negative Binomial Regression Results for
Southern Police District**

	Robbery	Theft	Carnapping	Motornapping
Police Checkpoints	0.00110 (0.00084)	0.00035 (0.00077)	-0.00004 (0.00300)	0.00169 (0.00182)
Oplan Bakal	-0.00106* (0.00062)	-0.00028 (0.00055)	-0.00049 (0.00281)	0.00333 (0.00168)
Mobile Patrols	0.00044 (0.00022)	0.00035 (0.00022)	-0.00161 (0.00112)	0.00131 (0.00055)
Beat Patrols	-0.00065** (0.00029)	-0.00026 (0.00027)	0.00036 (0.00134)	-0.00057 (0.00078)
Warrants Served	0.00791 (0.00798)	0.00864 (0.00739)	-0.01474 (0.04291)	0.00164 (0.02024)
Oplan Katok	0.00025 (0.00018)	-0.00038* (0.00022)	-0.00197 (0.00124)	0.00065 (0.00031)
Oplan Bulabog	0.00024 (0.00066)	0.00105 (0.00059)	0.00223 (0.00198)	0.00026 (0.00150)
Additional Police	-0.00243 (0.00176)	-0.00457** (0.00181)	0.00605 (0.00757)	-0.00440 (0.00400)
Constant	3.18544*** (0.35934)	2.37177*** (0.15410)	0.04901 (0.25021)	-0.10585 (0.22893)

*Sig. at 10%; **Sig. at 5%; ***Sig. at 1%

**Table 6.5. Fixed Effects Negative Binomial Regression Results for
Quezon City Police District**

	Robbery	Theft	Carnapping	Motornapping
Police Checkpoints	0.00070 (0.00268)	0.00069 (0.00232)	0.01151 (0.00885)	-0.02513** (0.01056)
Oplan Bakal	0.00102 (0.00116)	0.00284 (0.00097)	-0.00909* (0.00537)	0.00557 (0.00366)
Mobile Patrols	-0.00062 (0.00113)	0.00002 (0.00100)	-0.00755* (0.00444)	0.01466 (0.00335)
Beat Patrols	-0.00028 (0.00072)	-0.00015 (0.00058)	-0.00291 (0.00286)	-0.00162 (0.00214)
Warrants Served	-0.03557 (0.05189)	-0.02191 (0.04223)	0.18790 (0.18911)	-0.20730 (0.16933)
Oplan Katok	0.00000 (0.00028)	-0.00003 (0.00023)	0.00040 (0.00101)	0.00086 (0.00100)
Oplan Bulabog	0.00906 (0.00708)	0.00503 (0.00615)	-0.00522 (0.03086)	0.01616 (0.01855)
Additional Police	-0.00049 (0.00107)	0.00016 (0.00088)	0.00302 (0.00426)	-0.00356 (0.00306)
Constant	3.11297*** (0.29726)	2.81419*** (0.16800)	0.66418** (0.29497)	-0.19719 (0.27784)

*Sig. at 10%; **Sig. at 5%; ***Sig. at 1%

7. Conclusions and Recommendations

There are index crimes in the NCR which display an increasing trend: slowly increasing in number since the start of 2014. The historical plots in Figures 5.1 to 5.5 show how the incidence of index crimes has changed over time. The incidence of theft has increased quite rapidly in NPD, while in the EPD its increase is even sharper. Robbery incidence has also increased gradually in these two districts. Moreover, the incidence of motornapping has steadily increased in SPD.

One of the efforts exerted by the PNP in improving the crime situation in the country is by enacting different interventions which this paper has described. Some of these interventions are effective in reducing certain crimes, while some are not. Moreover, the relationship between the crimes and interventions differs across the police districts. This paper has identified the interventions which bring down the level of crime in the different police districts in NCR.

The district-level results show that there is always at least one intervention which significantly decreases at least one of the index crimes considered: robbery, theft, carnapping and motornapping. In the NPD, results show that increasing beat patrols is effective in crime reduction, particularly the crimes of robbery and motornapping. Increases in the number of mobile patrols and Oplan Bulabog operations produce significant decreases in the respective levels of motornapping and robbery in the EPD.

In the MPD, increasing the number of beat patrols is the effective in reducing the incidence of robbery. In the SPD, Oplan Bakal, beat patrols, Oplan Katok, and addition of policemen all have significant diminishing effects on crime. Lastly, the number of police checkpoints, Oplan Bakal operations and mobile patrols are the effective interventions in Quezon City.

The results show that increasing the number of Oplan Bakal operations, beat patrols and Oplan Bulabog operations are those which significantly reduce the number of robbery cases. Oplan Katok operations and addition of policemen are the interventions which can lower the incidence of theft in SPD. The interventions which can bring down the level of carnapping are the Oplan Bakal operations and mobile patrols in QCPD. Lastly, the increasing the number of police checkpoints, mobile patrols and beat patrols reduce the incidence of motornapping.

References

- ANDRESEN, M. and MALLESON, N., 2014, Police Foot Patrol and Crime Displacement: A Local Analysis, *Journal of Contemporary of Criminal Justice* 30 (2) (May 2014), Sage Journals. Available at: <http://ccj.sagepub.com/content/30/2/186.full.pdf+html>.
- AUSTRALIAN INSTITUTE OF CRIMINOLOGY, 2012, National crime prevention framework, Australian Government. Available at: http://www.aic.gov.au/crime_community/crimeprevention/ncpf.html.
- AUSTRALIAN INSTITUTE OF CRIMINOLOGY, 2014, Crime prevention approaches, theory and mechanisms, Australian Government, Available at: http://www.aic.gov.au/publications/current%20series/rpp/100-120/rpp120/07_approaches.html.
- BAILEY, W.C., 1976, Certainty of arrest and crime rates for major felonies: A research note, *Journal of Research in Crime and Delinquency* 13(2): 145-154.
- BRAGADO, F., 2007, Fundamentals of criminal investigation, Philippines: Philippine College of Criminology Manila Law College Textbook Committee.
- FELL, J.C., 2013, The effects of increased traffic enforcement on other crime, *Proceedings of the 2013 Australasian Road Safety Research, Policing & Education Conference, Pacific Institute for Research and Evaluation*, Available at: <http://acrs.org.au/files/arsrpe/Paper%2029%20-%20Fell%20-%20Traffic%20Leg%20&%20Enforcement.pdf>
- FELL, JC., TIPPETTS, AS. and LEVY, M., 2008, Evaluation of seven publicized enforcement demonstration programs to reduce impaired driving: Georgia, Louisiana, Pennsylvania, Tennessee, Texas, Indiana, and Michigan, *Annals of Advances in Automotive Medicine*, National Center for Biotechnology Information, Available at: <http://www.ncbi.nlm.nih.gov/pubmed/19026220>.
- HAUSMAN, J., HALL, B. and GRILICHES, Z., 1984, Econometric models for count data with an application to the patents-R&D relationship, *Econometrica* 52 (4): 909-938.
- HOUGHTON MIFFLIN HARCOURT, 2014, Police strategies, *Cliffs Notes*, Available at: <http://www.cliffsnotes.com/more-subjects/criminal-justice/police-function/police-strategies>

- KATES, D.B., 1990, Guns, murders, and the constitution, Pacific Research Institute. Available at: <http://www.constitution.org/211/2ndschol/115kat.pdf>
- KELLING, G.L., 2009, How New York became safe: The full story, Special issue 2009, *City Journal*, Available at: http://www.city-journal.org/2009/nytom_ny-crime-decline.html
- KIUNISALA, R., GUIAM, R. and CABANLIT, E. JR., 2004, Crime prevention campaign and leadership styles of General Santos City police precinct chiefs: Their effects to crime rate, crime solution efficiency and community support and involvement, 9th National Convention on Statistics (NCS), National Statistical Coordination Board. Available at: http://www.nscb.gov.ph/ncs/9thncs/papers/publicOrder_CrimePrevention.pdf
- KOPER, C.S., 1995, Just enough police presence: Reducing crime and disorderly behavior by optimizing patrol time in crime hot spots, *Justice Quarterly*, Taylor and Francis Group. Available at: <http://www.tandfonline.com/doi/abs/10.1080/07418829500096231>
- LACEY, J. H., JONES, R. K. and SMITH, R. G., 1999, Evaluation of Checkpoint Tennessee: Tennessee's Statewide Sobriety Checkpoint Program, Technical Report, National Highway Traffic Safety Administration. Available at: <http://www.nhtsa.dot.gov/people/injury/research/ChkTenn/ChkptTN.html>
- LATEST NEWS PHILIPPINES, 2013, 'Oplan Katok' yields thousands of loose guns. Available at: <http://latestnewsphilippines.com/national-news/oplan-katok-yields-thousands-of-loose-guns>
- MARVELL, T.B. and MOODY, C. E., 1996, Specification problems, police levels, and crime rates, *Criminology* 34 (4): 609–646, Wiley Online Library. Available at: <http://onlinelibrary.wiley.com/doi/10.1111/j.1745-9125.1996.tb01221.x/abstract>
- Office of the President of the Philippines, 2014, PNoy cites declining crime incidence in Metro Manila, Gov @ Work, Philippine Law Library. Available at: http://president.gov.ph/gov_at_work/pnoy-cites-declining-crime-incidence-metro-manila
- OVERSEAS SECURITY ADVISORY COUNCIL, 2014, Philippines 2014 crime and safety report. Available at: <https://www.osac.gov/pages/ContentReportDetails.aspx?cid=15294>
- PEPA, R. F., 2013, The true state of the Philippines: Crime in a culture of corruption. Available at: <http://newsjunkiepost.com/2013/07/21/the-true-state-of-the-philippines-crime-in-a-culture-of-corruption>
- PHILIPPINE NATIONAL POLICE, 2015, PNP Standard Operating Procedures (SOP) in Managing Police Operations.
- PHILIPPINE NATIONAL POLICE, 2013, PNP Guidebook on Human Rights Based-Policing. Available at: http://pnp.gov.ph/portal/images/stories/publicrelations/PNP_GUIDEBOOK_opt.pdf
- PLANT, J.B. and SCOTT, M. S., 2009, *Effective Policing and Crime Prevention* 68: 133-150. Available at: <http://www.popcenter.org/library/reading/pdfs/mayorsguide.pdf>
- QUEENSLAND POLICE, 2014, Police Beats, State of Queensland (Queensland Police Service). Available at: https://www.police.qld.gov.au/programs/cscsp/p_beat
- SANIDAD-LEONES, C. V., n.d, The current situation of crime associated with urbanization: Problems experienced and countermeasures initiated in the Philippines. Available at: http://www.unafei.or.jp/english/pdf/RS_No68/No68_13VE_Leones1.pdf
- SENATE ECONOMIC PLANNING OFFICE, 2013, Crime statistics at a Glance, Senate of the Philippines. Available at: <https://www.senate.gov.ph/publications/AAG%202013-05%20-%20Crime%20Statistics.pdf>

- SHAW, M., and TRAVERS, K., 2007, Strategies and best practices in crime prevention in particular in relation to urban areas and youth at risk, *Proceedings of the Workshop held at the 11th UN Congress on Crime Prevention and Criminal Justice, International Centre for the Prevention of Crime*, Available at: http://www.crime-prevention-intl.org/fileadmin/user_upload/Publications/Strategies_and_Best_Practices_in_Crime_Prevention_urban_Areas_and_Youth_at_Risk_ANG.pdf.
- SHERMAN, L. W., SHAW, J. W. And ROGAN, D. P., 1995, The Kansas City gun experiment, *Research in Brief*, National Institute of Justice, US Department of Justice. Available at: <https://www.ncjrs.gov/pdffiles/kang.pdf>
- SHERMAN, L.W. and WEISBURD, D., 1995, General deterrent effects of police patrol in crime hot spots: A randomized, controlled trial, *Justice Quarterly* 12: 4, Center for Evidence-Based Crime Policy. Available at: <http://cebcp.org/wp-content/onepaggers/ShermanWeisburd1995.pdf>.
- TITTLE, C.R. and ROWE, A.R., 1974, Certainty of arrest and crime rates: A further test of the deterrence hypothesis, *Social Forces* 52(4): 455-462.
- WEISBURD, D., GROFF, E. R. and SUE-MING, 2011, Understanding developmental crime trajectories at places: Social disorganization and opportunity perspectives at micro units of geography, Executive summary, United States Department of Justice. Available at: <https://www.ncjrs.gov/pdffiles1/nij/grants/236058.pdf>

**Appendix: Dates for the Weeks Covered by the Study:
December 30, 2013 to September 28, 2014**

Week	Date	Week	Date
1	December 30-January 5	21	May 19-25
2	January 6-12	22	May 26-June 1
3	January 13-19	23	June 2-8
4	January 20-26	24	June 9-15
5	January 27-February 2	25	June 16-22
6	February 3-9	26	June 23-29
7	February 10-16	27	June 30-July 6
8	February 17-23	28	July 7-13
9	February 24-March 2	29	July 14-20
10	March 3-9	30	July 21-27
11	March 10-16	31	July 28-August 3
12	March 17-23	32	August 4 -10
13	March 24-30	33	August 11-17
14	March 31-April 6	34	August 18-24
15	April 7-13	35	August 25-31
16	April 14-20	36	September 1-7
17	April 21-27	37	September 8-14
18	April 28-May 4	38	September 15-21
19	May 5 -11	39	September 22-28
20	May 12-18		

The Recursive Alpha (RALph) Coefficients: Quantifying Inter-item Cohesion under Indirect Range Restriction

Michael Van B. Supranes

John Francis J. Guntan

Joy Pauline Adrienne C. Padua

Joseph Ryan G. Lansangan

School of Statistics

University of the Philippines Diliman

Range restriction is a known cause of underestimation in the Cronbach's Alpha reliability coefficient. The estimate of the Cronbach's Alpha is usually adjusted to minimize bias, but existing methods require information about the population. In the case of indirect range restriction however, such information may not be readily or intuitively available. A data-driven bootstrap-based estimator that requires minimal assumptions about the unrestricted population, called the Recursive Alpha (RALph) coefficient, is therefore proposed. Based on the simulation studies, the two versions of the Ralph coefficient perform best when the information associated to the range restriction is strongly correlated with the characteristic being measured, and when the true reliability coefficient Alpha is high. Also, the RALph coefficients are found to be effective in minimizing the error in estimating Alpha under strong presence of range restriction. Moreover, considerations on the length of the instrument, scale of the responses, and sample size aid in minimizing the error of the proposed coefficients. In support of the simulation results, an empirical study using behavioral data on social media users is carried out, and evidently, the RALph coefficients are far better than the ordinary Cronbach's Alpha estimate.

*Keywords: range restriction, adjusted Cronbach's Alpha,
bootstrap sampling*

1. Introduction

Reliability is one of the important concepts in building quality and useful tests or assessment instruments. Together with and as anchor of validity, reliability is used to evaluate an instrument's dependability on measuring a certain characteristic. Measuring an instrument's reliability is a vital step in many psychological and educational researches. Currently, there are many coefficients that gauge instrument reliability, the most common of which, and relatively among the oldest and regarded as one of the most recorded measure of reliability, is the Cronbach's Alpha.

The Cronbach's Alpha (or simply, Alpha) was developed by Lee Cronbach (1951) to provide a measure of internal consistency of a test or scale. It describes the extent to which the items in an instrument measure a single construct (Tavakol, 2011). Alpha considers the average inter-item correlation in its computation and must be used under a strong belief of unidimensionality among the items being analyzed. It is a function of the number of items in an instrument, the average variance of responses and the average inter-item covariance. Hence, Alpha can be viewed as a function of inter-item correlations (inter-relatedness) and the number of items in an instrument. Being a function of variances and covariances, its accuracy as a measure highly depends on the range of scores observed from the sample. Many have explored the properties of Alpha and/or its estimators. Some have proposed better alternatives or coefficients, yet Alpha remains to be the most popular reliability coefficient. At the same time, Alpha is one of the most frequently misused reliability coefficient in published researches (Sijtsma, 2009; Knapp, 1991). Sijtsma (2009) provides a thorough discussion of these common errors in using Alpha.

Range restriction is a common problem in estimating reliability measures. This commonly happens when there is a preselection or censoring method applied to the respondents in the experimental stage of the instrument. Similarly, range restriction may be viewed as the case when the data collected is from a sample of the restricted population that is used for inference for the unrestricted population (Le and Schimdt, 2006). For instance, a new instrument that is built to test or detect a behavioral disorder among children was only tested for reliability among children ages 7 to 9 years old, yet the instrument reliability is inferred over the whole (unrestricted) population of children. The range criterion (age for the given example) would more likely restrict the possible behavioral test scores, if age is an important factor in the prevalence of the behavioral disorder. In the case of using the Alpha as a reliability measure, a low estimate may be computed not due to the lack of inter-relatedness of the items, but rather due to the nature of the sample itself.

It has been shown in many related literature that range restriction causes underestimation in the estimates of the reliability coefficient and other measures indexed by correlations (Sackett et al., 2007). There are two types of range

restriction. Direct range restriction is the condition wherein the interests involve two or more correlated variables in the population and the sample is taken from a population that was restricted based on at least one of the variables of interest. Indirect range restriction is the condition wherein a correlated variable outside of the main interests was used as a basis for restricting the population hence, the range of values in the population is indirectly restricted. It has been shown that direct and indirect range restrictions have different effects on estimation, such as in meta-analysis, correlations, and reliability measures (Hunter et al., 2006).

In general, range restriction may be introduced to the process due to selection bias or simply by chance. There are formulated adjustments in or methods for estimation in case of range restriction. Some of the methods include imputation of the unobserved values, adjustment of the correlation estimates before estimation, or adjustment of the initially computed estimate. As an adjustment of the (usual) Cronbach's Alpha estimator, the Range-Restricted Cronbach's Alpha, α_{adj} , is formulated as (Gulliksen, 1950):

$$\alpha_{adj} = 1 - \left(\frac{S_{res}}{S_{pop}} \right)^2 (1 - \alpha) \quad (1)$$

The adjustment is made to address the underestimation issues of the Alpha estimator under indirect range restriction. The adjustment uses $\frac{S_{res}}{S_{pop}}$, the ratio of the standard deviations of scores in the range-restricted sample and in the unrestricted population. Hence, the adjustment can only be computed upon prior knowledge of the true standard deviation of scores in the entire population. It is commonly possible to access the true population standard deviation for cases under direct range restriction, since the restriction is actually based on one of the variables of interest. The case is different however, for indirect range restriction, in which an external factor was used as a basis of restriction and thus it is possible that there is no knowledge on the true value of the population standard deviation.

In cases in which the population standard deviation is unknown, truncation of the correlation estimates under a truncated normal distribution assumption may be used (Alexander et al., 1984). However, it may be hard to view the individual item scores as normally distributed, since instruments are usually composed of discretized scales (i.e., a dichotomous scale or a Likert scale). Moreover, large sample approximations may not apply in usual applications wherein the test sample is usually small. Under such conditions, the potential of bootstrapping may be exploited. Population conditions can be approximated using bootstrap method, and thus bootstrapping may be useful when the population standard deviation is unknown. Moreover, biased-corrected and accelerated bootstrap

confidence interval may be used for assessing precision of the reliability measure (Kelley and Cheng, 2012).

In this study, an adjustment on the inter-item cohesion (or reliability coefficient) estimation as well as the possible effects of the degree of range restriction vis-à-vis scale levels, number of items included, and sample size are of primary interest. As a potential solution to underestimation under range restriction of the reliability coefficient and as an alternative when population quantities cannot be assumed and/or derived, the bootstrap-based estimators are proposed. Characterization of these estimators is done through simulation under different scenarios which take into account the identified factors.

2. Methodology

Factors considered

Indirect range restriction uses an external variable in restricting the population of interest, call this screening variable. The screening variable is usually correlated to the items or components of the main instrument. The screening variable filters certain observations depending on a criterion or acceptance range. As the correlation between the item scores and the screening variable increases, the dispersion of the instrument (total) scores reduces and the range of the screening variable narrows down. Hence, it is hypothesized that when the screening variable is weakly correlated, the restriction on the instrument scores will appear to be done at random points within the screening variable's range.

The second factor considered is the proportion of the population that is retained after applying a restriction criterion, this is termed as coverage. It is hypothesized that if the coverage is low, i.e., a shorter range of values is observed, underestimation effects of indirect range restriction may be more severe.

Note that these two factors, the correlation of the screening variable and the extent of coverage, may not be observed by the researcher. The strength of correlation of the screening variable and the amount of coverage are unknown in most cases. To further characterize the qualities of the proposed procedures, the number of items in the instrument, the scale of the response, and the sample size are also considered.

Proposed Estimators

Three estimators of the reliability coefficient are compared and analyzed in the study – the usual Cronbach's Alpha estimator, and the two versions of the proposed recursive reliability coefficient, which are modifications of Gulliksen's (1950) formulation.

$$\alpha = \frac{K\bar{c}}{1 + (K - 1)\bar{c}} \quad (2)$$

Assuming a standardized data set, the Cronbach's Alpha α is computed as a function of the number of items K and the average inter-item correlation $\bar{c}$, as shown in equation (2). In case of direct range restriction, a Range Restricted Cronbach's Alpha α_{adj} was formulated by Gulliksen (1950), as shown in equation (1).

The problem of range restriction can be viewed as having an unrestricted population that supersedes a restricted population from which the sample is actually taken. With such view of range restriction, it is proposed to take several bootstrap samples, similar to the process of taking a sample from a restricted population (see Figure 1). The key here is the assumption of an "unrestricted" sample that supersedes the observed sample, if the population was actually unrestricted.

Two forms of adjustments are then proposed. Bootstrap sampling is mainly used, with the final adjusted estimate as the average of all the adjusted estimates computed from each bootstrap sample. The method makes adjustment possible even without the knowledge of the true population variance. Performance of the estimators are investigated under certain conditions discussed in the next section.

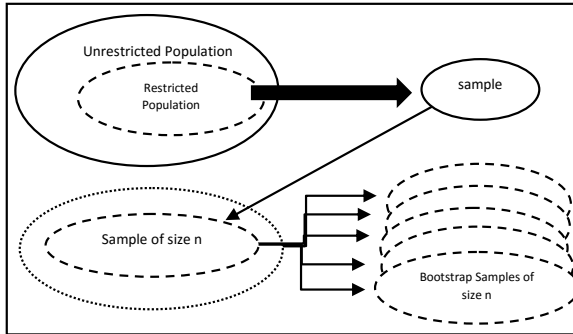


Figure 1. (Top) Range Restriction and (Bottom) Resampling from Restricted Sample

The proposed adjusted Alpha coefficients are referred to as Recursive Alpha coefficients or simply RAlph coefficients. First of the two RAlph coefficients is the bootstrapped range-restricted-equivalent Alpha using Gulliksen's (1950) formulation:

$$RAlph1 = \alpha_{adj}^* = \frac{\sum_{i=1}^r \left(1 - \left(\frac{s_i}{s_a} \right)^2 (1 - \alpha_i) \right)}{r} \quad (3)$$

where r is the number of bootstrap replicates. RAlph1 is a bootstrapped modified Gulliksen's formulation, where the s_i^2 is the variance (of scores) of a bootstrap sample and s_a^2 is the adjusted variance (of scores) of the sample. The adjusted

variance of the sample is computed as the sample variance s^2 (of total points) multiplied by a ratio of ranges, as shown in equation (4).

$$s_a^2 = \left[\left(\frac{mK - K}{X_{(n)} - X_{(1)}} \right) s \right]^2 \tag{4}$$

The range in the numerator of the adjustment fraction is the full potential range. The full potential range is simply the difference between the highest and lowest possible score from the instrument. In the formulation, K is the number of items in the instrument and m is the highest possible point in an item with an integer scale that starts at 1 (e.g. in a 5-point Likert scale, $m = 5$). On the other hand, the observed range, $X_{(n)} - X_{(1)}$, is the difference between the observed minimum and maximum scores. The rationale for the adjustment is that when a random variable is bounded, the changes in the range become a good indicator of changes in variability and/or in location.

The second RAlph coefficient is constructed such that other factors are considered in the adjustment, i.e., the sample size and the number of items in the instrument. The estimator is a bootstrapped item-size adjusted Cronbach’s Alpha $\alpha_{K,n}^*$, which, with r bootstrap replicates, is computed as:

$$RAlph2 = \alpha_{K,n}^* = \frac{\sum_{i=1}^r \left(1 - \left(\frac{s_i}{s_a} \right)^{\frac{n-K+0.5}{n+K-0.5}} (1 - \alpha_i) \right)}{r} \tag{5}$$

RAlph2 coefficient is a modification of equation (3), wherein the power of the ratio of standard deviations is adjusted depending on the sample size n and the number of items K . Note that the power $\frac{n-K+0.5}{n+K-0.5}$ is numerically derived to yield an acceptable estimator of Alpha. Also, the adjustment is restricted to $n \geq K$, otherwise, RAlph2 will suffer from underestimation.

Simulation procedure

There were six factors considered in data simulation (as listed in Table 1). Three of the six factors were population characteristics – the correlation of the screening variable, the coverage of the restricted population, and the true value of Alpha. The other three were factors involved in the instrumentation and/or conduct of reliability analysis – the scale of the instrument, the number of items in the instrument, and the sample size. For each scenario, there were 500 sampling

iterations, and 500 bootstrap replicates were used in computing the bootstrap-based estimators for each sampling iteration.

Table 1. Factors Considered in Data Simulation

Correlation of the Screening Variable	around 0.30 – weakly correlated around 0.80 – strongly correlated
Range Restriction Coverage	70% covered – high coverage 30% covered – low coverage
Population Cronbach’s Alpha	0.40 – weak internal consistency 0.90 – strong internal consistency
Scale of Items	2 – dichotomous scale 5 – five-point Likert scale 10 – ten-point rating scale
Number of items/components	5 – few items 15 – many items
Sample Size	30 – small sample 100 – large sample

The simulation was done in stages. The first stage involved simulating population values from a multivariate normal distribution. The mean vector was set to a zero vector, while the correlation matrix was constructed using the number of items and the population Alpha as constraints. Then, inter-item correlation values were sampled from a uniform distribution, whose average was equal to the average correlation and such that the inter-item correlations exhibited variation (to mimic presence of best subset of items). The screening variable correlated to the item scores was then constructed.

The next stage was simulating range restriction. A subset of the multivariate normal data was taken, depending on the coverage of the scenario being simulated. Then, 500 simple random samples without replacement was selected from each of the restricted population data.

In the next stage, the normally distributed item scores in each sample were discretized depending on the scale set for the simulated scenario. The normal values were categorized using percentiles as boundaries. After which, a simulated total score per observation was computed. Finally, across the different scenarios, the point estimators were then computed. Also, the bias, mean absolute deviation (MAD), and percent of the time underestimated/overestimated of the point estimators were computed.

3. Results and Discussion

Factors affecting the errors of alpha under indirect range restriction

As expected, the ordinary Cronbach’s Alpha estimator (referred to as alpha, different from the population value Alpha) underestimates the true value of Alpha

under the presence of indirect range restriction, but there are certain cases in which the error is not as worse as others. All scenarios with strongly correlated screener and low coverage have underestimation 98% of the time to 100% of the time (across sampling iterations). Relatively low percentages of the time underestimation are observed among scenarios having nondichotomous scales (5 or 10 point scales), low population Alpha values, or a weakly correlated screening variable. Looking at the bias and the mean absolute deviation, Alpha tends to be more underestimated and inaccurate using the alpha as estimator, when either the screening variable is strongly correlated, the coverage is lower, the scale is shorter, there are fewer items being analyzed, or the sample size is smaller.

The mean absolute deviation (MAD) is especially sensitive to the significance of the screening variable in restricting the item scores. For certain cases, the ordinary alpha is better than the RAlpha adjustment (especially RAlpha1), particularly when the screening variable is weakly correlated (see Table 2). That is, if the screening variable does not have a significant relationship with the characteristic being measured, then there's no reason to worry about indirect range restriction affecting alpha. The ordinary alpha also performs comparably as RAlpha2 when the screening variable is weakly correlated.

Table 2. Average Mean Absolute Deviation of Ordinary alpha versus RAlpha coefficients by Correlation of the Screening Variable

Correlation of the Screening Variable	Estimators		
	alpha	RA1	RA2
0.2	0.09429	0.17280	0.09222
0.8	0.38674	0.18706	0.17274

With respect to coverage, the average error when using alpha is generally higher when the coverage is low. However, the degree of error due to coverage depends on the correlation of the screening variable to the characteristic being measured (see Table 3). The average bias of alpha tends to worsen as the coverage decreases under strongly correlated screening variable.

Table 3. Average Bias of alpha by Coverage and Screening Variable

Coverage of Restricted Population	Correlation of Screening Variable	
	0.2	0.8
0.3	-0.07565	-0.51496
0.7	-0.05761	-0.25654

The average MAD using alpha under indirect range restriction also varies depending on the instrument and the sample size (see Table 4). As expected,

instruments with fewer items or with shorter scale tend to be more sensitive to range restriction. In addition, the average MAD is smaller for large sample sizes.

Table 4. Average Mean Absolute Deviation (MAD) of Alpha by Number of Items, Scale and Sample Size

Factors		Average MAD
Number of Items	5	0.28891
	15	0.19213
Number of Responses in the Scale	2	0.28852
	5	0.22180
	10	0.21124
Sample Size	30	0.26066
	100	0.22038

Mean Absolute Deviation (MAD) and bias of alpha and the proposed adjustments

Table 5 shows the MAD and the percent of the time underestimation is observed for the different scenarios involving high population Alpha, strongly correlated screener and low coverage. RAlph1 generally has a smaller MAD and mostly overestimates. On the other hand, RAlph2 always underestimates Alpha, but it has lower MAD than alpha. In general, under this scenario, the RAlph1 is relatively superior to the RAlph2 in terms of MAD.

Table 5. Mean Absolute Deviations and Percentages of Underestimation using Alpha and Proposed Adjustments under High Population Alpha, Strongly Correlated Screener and Low Coverage

Cases			Mean Absolute Deviation			Percentage of Underestimation		
Sample Size	Scale	No. of items	alpha	RAlph1	RAlph2	alpha	RAlph1	RAlph2
100	10	15	0.2153	0.0280	0.0828	100%	1%	100%
100	2	15	0.3907	0.0709	0.2325	100%	15%	100%
100	5	15	0.2414	0.0145	0.1069	100%	100%	100%
30	10	15	0.2289	0.0444	0.1506	100%	0%	100%
30	2	15	0.4127	0.0362	0.3208	100%	3%	100%
30	5	15	0.2541	0.0327	0.1754	100%	87%	100%
100	10	5	0.3112	0.0134	0.1074	100%	26%	100%
100	2	5	0.6931	0.1463	0.3601	100%	65%	100%
100	5	5	0.3700	0.0148	0.1404	100%	100%	100%
30	10	5	0.3286	0.0250	0.1332	100%	14%	100%
30	2	5	0.7255	0.0733	0.3597	100%	33%	100%
30	5	5	0.3924	0.0206	0.1698	100%	94%	100%

Table 6 shows the MAD of scenarios in which the conditions cause severe underestimation with a low population Alpha. Consistently, the RAlph coefficients show improvement on the MAD. In most cases, RAlph2 is better than RAlph1. In addition, RAlph2 has lesser tendency to overestimate the low true value Alpha.

Table 6. Mean Absolute Deviations and Percentages of Underestimation using Alpha and Proposed Adjustments under Low Population Alpha, Strongly Correlated Screener and Low Coverage

Cases			Mean Absolute Deviation			Percentage of Underestimation		
Sample Size	Scale	No. of items	<i>alpha</i>	<i>RAlph1</i>	<i>RAlph2</i>	<i>alpha</i>	<i>RAlph1</i>	<i>RAlph2</i>
100	10	15	0.4576	0.4575	0.1013	100%	0%	6%
100	2	15	0.4431	0.2362	0.1137	100%	0%	19%
100	5	15	0.4568	0.4225	0.0703	100%	0%	89%
30	10	15	0.5240	0.5034	0.1854	100%	0%	94%
30	2	15	0.5192	0.3567	0.3001	100%	1%	99%
30	5	15	0.5259	0.4806	0.2082	100%	50%	100%
100	10	5	0.8431	0.2615	0.1501	100%	0%	83%
100	2	5	0.6817	0.1516	0.3217	99%	0%	86%
100	5	5	0.8214	0.1894	0.2100	98%	0%	94%
30	10	5	0.9001	0.3552	0.2008	100%	0%	85%
30	2	5	0.7371	0.1499	0.3539	100%	0%	90%
30	5	5	0.8896	0.3030	0.2489	100%	20%	96%

Table 7 contains the MAD of the estimators when range restriction is expected to have minimal effect on a high population Alpha. In general, minimal MAD is observed for alpha when the screener is weakly correlated and the coverage is high. The MAD of the RAlph coefficients is comparably on the same level for each scenario. These suggest that the formulation of the RAlph coefficients are able to adapt to the severity of restriction for high values of Alpha.

Table 8 contains the MAD of the estimators when range restriction has minimal effect on a low population Alpha. The MAD of RAlph1 is highest in almost all cases. Hence, RAlph1 may not be a good adjustment method under low population Alpha with weak-correlated screener and high coverage. On the other hand, RAlph2 performs slightly better than the other estimators on certain cases. RAlph2 may not be good as the ordinary Alpha but is more accurate than RAlph1.

Table 7. Mean Absolute Deviations and Percentages of Underestimation using alpha and Proposed Adjustments under High Population Alpha, Weakly Correlated Screener and High Coverage

Cases			Mean Absolute Deviation			Percentage of Underestimation		
Sample Size	Scale	No. of items	alpha	RA1ph1	RA1ph2	alpha	RA1ph1	RA1ph2
100	10	15	0.0146	0.0207	0.0106	72%	6%	41%
100	2	15	0.0729	0.0683	0.0725	85%	32%	72%
100	5	15	0.0194	0.0114	0.0134	100%	100%	100%
30	10	15	0.0249	0.0402	0.0210	74%	66%	72%
30	2	15	0.0753	0.0460	0.0764	91%	90%	92%
30	5	15	0.0290	0.0276	0.0250	100%	100%	100%
100	10	5	0.0171	0.0139	0.0158	62%	0%	55%
100	2	5	0.1044	0.1024	0.1048	72%	6%	68%
100	5	5	0.0254	0.0238	0.0254	97%	93%	98%
30	10	5	0.0316	0.0201	0.0262	70%	38%	64%
30	2	5	0.1157	0.1083	0.1189	76%	68%	75%
30	5	5	0.0385	0.0261	0.0358	98%	98%	98%

Table 8. Mean Absolute Deviations and Percentages of Underestimation using alpha and Proposed Adjustments under Low Population Alpha, Weakly Correlated Screener and High Coverage

Cases			Mean Absolute Deviation			Percentage of Underestimation		
Sample Size	Scale	No. of items	alpha	RA1ph1	RA1ph2	alpha	RA1ph1	RA1ph2
100	10	15	0.0706	0.4439	0.2264	56%	0%	0%
100	2	15	0.1229	0.2328	0.0631	64%	0%	0%
100	5	15	0.0758	0.4096	0.1917	88%	0%	27%
30	10	15	0.1519	0.4936	0.1352	58%	0%	22%
30	2	15	0.2008	0.3447	0.1306	62%	0%	28%
30	5	15	0.1563	0.4694	0.1247	78%	0%	58%
100	10	5	0.0769	0.2215	0.1094	54%	0%	7%
100	2	5	0.1276	0.1197	0.1277	61%	5%	23%
100	5	5	0.0818	0.1482	0.0786	87%	84%	87%
30	10	5	0.1450	0.3280	0.1475	50%	0%	13%
30	2	5	0.1905	0.1206	0.1570	56%	0%	25%
30	5	5	0.1505	0.2673	0.1262	69%	47%	69%

Table 9 summarizes the accuracy of the estimators depending on the situation of the population and range restriction. The average MAD is computed as the average of the MAD values for the scenarios under the specific set of screener

correlation, coverage, and population Alpha. Clearly, the RAlph coefficients work better than the ordinary alpha under a strongly correlated screener. Under a high screener correlation, RAlph1 has a lower average MAD than RAlph2 when the true Alpha value is high, while RAlph2 has a better average MAD than RAlph1 when the true Alpha value is low. Also, RAlph2 has a comparable average MAD as the ordinary alpha under a weakly correlated screener.

Table 9. Average MAD of Alpha and Proposed Adjustments under Different Levels of Screener Correlation, Coverage and Population Alpha

Cases			Average MAD		
Screener Correlation	Coverage	Population Alpha	alpha	RAlph1	RAlph2
0.20	0.70	0.40	0.1292	0.2999	0.1348
0.20	0.70	0.90	0.0474	0.0424	0.0455
0.20	0.30	0.40	0.1427	0.3017	0.1349
0.20	0.30	0.90	0.0579	0.0471	0.0537
0.80	0.70	0.40	0.3574	0.3050	0.1630
0.80	0.70	0.90	0.1593	0.0776	0.1276
0.80	0.30	0.40	0.6500	0.3223	0.2054
0.80	0.30	0.90	0.3803	0.0433	0.1950

Table 10 shows the average MAD of the different estimators by number of items, type of scale and sample size. In terms of these researcher-controllable factors, RAlph1 has a lower MAD under lower number of items, shorter scales, or larger sample size. While RAlph2 has a lower MAD under higher number of items, nondichotomous scales, or larger sample size. For RAlph2, note that the average MADs of a 5-item instrument and a 15-item instrument are comparable. Moreover, the differences in the average MAD for the different sample sizes are also small for the RAlph coefficients.

Table 10. Average MAD of alpha and Proposed Adjustments by Number of Items, Type of Scale and Sample Size

Factors		Alpha	RAlph1	RAlph2
Number of Items	5	0.2889	0.1393	0.1471
	15	0.1921	0.2206	0.1179
Number of Responses in the Scale	2	0.2885	0.1627	0.1870
	5	0.2218	0.1765	0.1089
	10	0.2112	0.2005	0.1016
Sample Size	30	0.2607	0.1969	0.1451
	100	0.2204	0.1630	0.1199

4. Real Data Application

Using an experimental social media behavior instrument, the Alpha is computed for a cloud nine construct. The cloud nine construct measures the tendency of an internet user to use social media for enjoyable/entertainment purposes. The data is based on an exploratory study conducted among the students of University of the Philippines Diliman. The data consisted of varying types of students and social media users. The instrument used has a 10-point scale. For the purpose of the study, the cloud nine construct is tested using 5 items only. These items are derived from an initial exploratory factor analysis of the original and larger data.

In the original data, the Alpha for the cloud nine construct is 0.9160. An artificial indirect range restriction is then applied to the sample data such that higher total scores for cloud nine is more likely to be retained. Restriction was based on other factors that are found to be correlated to the cloud nine construct, and after the restriction, the original sample containing 276 responses were narrowed down to 92 responses (i.e., the range restricted sample).

The computed alpha for the range restricted sample is at 0.5404. Applying the proposed adjustments, the RAlph1 coefficient overestimated the original Alpha with an estimate of 0.9818. On the other hand, RAlph2 coefficient gave an estimate of 0.8934, which is closest to and slightly underestimates the original Alpha.

5. Conclusion and Recommendations

There are two potential uses of the RAlph coefficients. First, a high difference between alpha and RAlph2 can serve as a good indicator of severe range restriction. Second, both forms of the RAlph coefficients are able to lessen the bias due to range restriction. Hence, RAlph1 and RAlph2 are relatively more robust to range restriction than the usual Cronbach's Alpha.

Both forms of the RAlph coefficients may help in regaining the accuracy loss due to range restriction without requiring strong assumptions such as a truncated normal distribution or a known population standard deviation of scores. RAlph1 and RAlph2 are accurate in the presence of severe range restriction, i.e., strongly correlated screener and low population coverage. Moreover, both are best in estimating high population values of Alpha.

In general, RAlph2 may be preferred over RAlph1. Although RAlph1 is more accurate than RAlph2 in severe cases with high population Alpha, RAlph1 tends to highly overestimate, when restriction is not severe or the population Alpha is low. As overestimating Alpha imposes worse misinformation and misinterpretation than underestimating Alpha, the RAlph2 is then generally preferred over RAlph1.

Both alpha and RAlph2, generally, have the same effects due to changes in sample size, length of scale, and number of items. For Alpha and RAlph2, error is minimized when the sample size is larger, the scale is longer (e.g. 10-point

rating scale) and the instrument is relatively long. On the other hand, RAlph1 is generally performs better when there are fewer items and shorter scales (e.g. dichotomous scales).

Since RAlph1 and RAlph2 are transformations of alpha, its dependability as a reliability measure is also anchored to the properties of alpha. The instrument being tested must be unidimensional, and the instrument must not be too long, as it is sensitive to increasing number of items. The RAlph coefficients are based on data-driven approximations of unknown population conditions, while existing parametric methods are exactly derived from assumed distributions. Hence, Gulliksen's adjustment may still be a better option when the population standard deviation is known, and other normal-based correction methods on correlations may still be best when there's no reason to question normality. Further studies are needed in comparing the accuracy of RAlph coefficients and existing parametric adjustments under unconventional conditions, such as nonnormal, skewed or multimodal distributions.

References

- ALEXANDER, R, ALLIGER, G. and HANGES, P., 1984, Correcting for range restriction when the Ppopulation variance is unknown, *Applied Psychological Measurement* 8(4): 431-437.
- CHRISTMAN, A. and VAN AELST, S, 2006, Robust estimation of Cronbach's alpha, *Journal of Multivariate Analysis* 97(7): 1660-1674
- CRONBACH, L., 1951, Coefficient alpha and the internal structure of tests, *Psychometrika* 16: 297-334.
- CRONBACH, L., 1990, *Essentials of Psychological Testing*, 5th ed., New York: Harper and Row
- GULLIKSEN, H., 1950, *Theory of Mental Tests*, New York, Wiley.
- GUTTMAN, L., 1945, A basis for analyzing test-retest reliability, *Psychometrika* 10(4): 255-282
- HELD, J. and FOLEY, P., 1994, Explanations for accuracy of the general multivariate formulas in correcting for range restriction, *Applied Psychological Measurement* 18(4): 335-367.
- HUNTER, J., SCHMIDT, F., and LE, H., 2006, Implications of direct and indirect range restriction for meta-analysis methods and findings, *Journal of Applied Psychology* 91: 594-612.
- IACOBUECC, D. and DUHACHEK, A., 2003, Advancing alpha: Measuring reliability with confidence, *Journal of Consumer Psychology* 13(4): 478-487.
- KELLY, K. and CHENG, Y., 2013, Estimation of and confidence interval formation for reliability coefficients of homogeneous measurement instruments, *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 8(2): 39-50.
- KNAPP, T.R., 1991, Coefficient alpha: Conceptualizations and anomalies, *Research in Nursing and Health*, 14, 457-460.

- LE, H. and SCHIMDT, F., 2006, Correcting for indirect range restriction in meta-analysis: Testing a new meta-analytic procedure, *Journal of Applied Psychology* 11(4): 416-438.
- LI, J., CHAN, W., and CUI, Y., 2011, Bootstrap standard error and confidence intervals for the correlations corrected for indirect range restriction, *British Journal of Mathematical and Statistical Psychology* 64: 367-387.
- LI, J., CUI, Y., GIERL, M. and CHAN, W., 2012, Bootstrap confidence intervals for range-restricted coefficient alpha, Poster presented at the 2012 Annual Meeting of American Educational Research Association (AERA), Vancouver, Canada. Abstract retrieved from [http://www2.education.ualberta.ca/educ/psych/crame/docs/April%202012/AERA%20conference%202012%20\[Johnson\].pdf](http://www2.education.ualberta.ca/educ/psych/crame/docs/April%202012/AERA%20conference%202012%20[Johnson].pdf).
- SACKETT, P., LIEVENS, F., BERRY, C., and LANDERS, R., 2007, A cautionary note on range restriction and predictor intercorrelations, *Journal of Applied Psychology* 92: 538-544.
- SIJTMA, K., 2009, On the use, the misuse, and the very limited usefulness of Cronbach's alpha, *Psychometrika* 74(1): 107-120.
- TAVAKOL, M. and DENNICK R., 2011, Making sense of Cronbach's alpha, *International Journal of Medical Education* 2:53-55.
- WIBERG, M. and SUNDSTORM, A., 2009, A comparison of two approaches to correction of restriction of range in correlation analysis, *Practical Assessment, Research and Evaluation* 14(5).
- YURDUGUL, H., 2008, Minimum sample size for Cronbach's coefficient alpha: A Monte-Carlo study, *H. U. Journal of Education* 35: 397-405.

Guidelines for Authors

The Philippine Statistician (TPS) is the official scientific journal of the Philippine Statistical Association, Inc. (PSAI). It considers papers resulting from original research in statistics and its applications. Papers will be sent for review on the assumption that this has not been published elsewhere nor is submitted in another journal.

Aims and Scope

The Journal aims to provide a media for the dissemination of research by statisticians and researchers using statistical method in resolving their research problems. While a broad spectrum of topics will be entertained, those with original contribution to the statistical science or those that illustrates novel applications of statistics in solving real-life problems will be prioritized. The scope includes, but is not limited to the following topics:

- Official Statistics
- Simulation Studies
- Survey Sampling
- Time Series Analysis
- Nonparametric Methods
- Econometric Theory and Applications
- Other Applications
- Computational Statistics
- Mathematical Statistics
- Statistics Education
- Biostatistics
- Experimental Designs and Analysis

In addition to research articles, the Journal will have the following sections that may appear in some of its issues (but not necessarily in all):

- *Letters to the Editor*. This section will provide a forum for the airing of opinions on issues pertinent to the statistical community or offers commentaries on articles that have appeared in the journal.
- *Notes* section will include notices and announcements of upcoming events, conferences, calls for papers.
- *Review* section will present reviews on statistics books and software.
- *Teacher's Corner* shares experience of some teachers related to statistical education.

Articles submitted for the four special sections above will be reviewed only by the Editor and/or Associate Editors.

Submission of Manuscript

Only unpublished manuscripts will be considered. They will be refereed and evaluated on content, language and presentation. The article in MS word format, without author's identification should be sent as email attachment to Erniel B. Barrios, Editor, *The Philippine Statistician*: ernielb@yahoo.com / barrios.

erniel@gmail.com. A separate file containing the title of the paper, authors(s) (corresponding author identified) and their affiliations and complete address should be included in a separate file to be emailed along with the main article. To avoid delays and difficulties in submission, authors should follow instructions on style prescribed below.

A printed page in the Journal will have a maximum amount of space of 4.5" by 8.5."

Organization of Manuscript

The manuscript should be written in 8.5" x 11" page using Times New Roman font size 12 with 1" margin in all sides.

Manuscripts must be organized in the following manner:

- *Title*: This should be brief and concise.
- *Abstract and Key Words*: An abstract of at most 250 words must be submitted with the manuscript. It precedes the article text. The abstract should summarize objectives, results, and main conclusions, but it should not contain any graph or complex mathematical notation and no references. Three to six keywords should be identified.
- *Article Text*: Sections should be concise and numbered in a decimal system. Tables, Figures, and Artwork may be used within the body of the article, should be numbered consecutively. Tables, Figure Titles and Legends, Figure Artwork must be strategically placed in the article text. The original files for the tables, figures, and artwork should also be submitted to facilitate typesetting. Authors must obtain written permission to reproduce or adapt all or part of a figure from a copyrighted source. Mathematical equations cited in the text should be numbered consecutively. Numbers should be placed at the rightmost margin of the equation line in parenthesis. Matrices should appear in bold and vectors in italics. All other symbols should appear in italics. The preferred software for equations are Mathtype and MS Equation Editor, which are add-ins of MS Word.
- *Acknowledgments*: An acknowledgment section may be included at the end of the article. This section should acknowledge financial assistance in the form of grants or university funding, assistance by individual colleagues, and any other pertinent information. This section will be inserted by the author only upon acceptance of the paper.
- *References*: All references included in the list at the end of an article must be cited in the text. References are cited in the text in the following format: (Author, Year). Up to two authors can be cited in the text. If there are three or more authors, only the first author will be cited in the text, e.g. (Author et al., Year). The following format will be followed in the listing references:

Journal Article

LANDAGAN, O. and BARRIOS, E., 2007, An estimation procedure for a spatial-temporal model, *Statistics and Probability Letter* 77(4): 401-406.

Book

KOTTAK, C., 1991, *Anthropology: The Exploration of Human Diversity*, 5th ed., New York: McGraw-Hill, Inc.

Book Chapter

FINK, E. and PRATT, K., 2008, Indexing compressed time series, in Last, M., Kandel, A. and Bunke, H., eds., *Data Mining in Time Series Databases*, Singapore: World Scientific, pp. 43-66.

Internet Document

MUNDLAK, Y., LARSON, D. and BUTZER, R., 2002, Determinants of agricultural growth in Indonesia, the Philippines, and Thailand, V. 1, World Bank Working Paper 2803, The World Bank. Available at: http://econ.worldbank.org/external/default/main?pagePK=64165259&piPK=64165421&theSitePK=469372&menuPK=64216926&entityID=000094946_02032604542948

For particulars about the style, please download the Philippine Statistician Style Guide at www.philstat.org.ph.

- *Appendices*: A single appendix is headed, “APPENDIX: FOLLOWED BY A DESCRIPTIVE TITLE”. If there are two or more appendices, they should be labeled, “APPENDIX A,” “APPENDIX B,” and so on.
- *Editorial Style*: In addition to content, manuscripts are evaluated on their conciseness and clarity. Thus, the Journal gives premium to well-written and well-structured papers that will be of interest to a wide segment of the readership. Manuscripts and reviews that have been accepted for publication will be copy edited in accordance to accepted rules of correct grammar, usage, spelling, and punctuation. To avoid common problems of style, for guidelines on style, usage, and the preparation of technical manuscripts for publication, the following reference may be consulted:

The Chicago Manual of Style (14th ed.) (1993),
Chicago: University of Chicago Press.

Editorial Notes

Use quotation marks only when a standard term is used in a nonstandard way and to indicate the beginning and ending of a direct quotation.

1. Hyphens are used when two or more adjectives or an adjective and a noun together modify another noun; for example, *goodness-of-fit test* is the equivalent of *test for goodness of fit*. Most words with prefixes such as sub and non are not hyphenated, for example, *subtable*, *nonnormal*.
2. Italics are used to introduce important terms, when appropriate; they are to be used sparingly to indicate emphasis.

3. Abbreviations and acronyms should be minimized; those that are used are spelled out on their first appearances in the manuscript with the shortened form given in parentheses, for example, best linear unbiased estimate (BLUE).
4. Numbers under 10 are spelled out when they are not part of an equation or an expression containing symbols.
5. The sign % is always used when giving a specific percentage, for example, 23%, not 23 percent. Otherwise use the word percent.

Copyright Transfer Form

Authors of accepted papers will be required to submit an author copyright transfer form before the final release of the journal.



PHILIPPINE POSTAL CORPORATION
MEGA MANILA AREA
Liwasang Bonifacio 1000 Manila

Permit No. _____

Expiration Date : _____

SWORN STATEMENT

The undersigned ERNIEL B. BARRIOS editor of THE PHILIPPINE STATISTICIAN, frequency of Issue BI-ANNUAL language ENGLISH in which printed at PSAI OFFICE office of publication entered as SECOND CLASS (first, second and / or third class) office of entry UP POST OFFICE after having duly sworn to in accordance with law, hereby submit the following statement of ownership, management, circulation, etc., which is required by act 2580 as Amended by Commonwealth Act 201.

ADDRESS

Editor SCHOOL OF STATISTICS Bldg, Magsaysay Ave.
UNIVERSITY OF THE PHILIPPINES, DILIMAN, QUEZON CITY
Owner PHILIPPINE STATISTICAL ASSOCIATION, INC
Publisher PHILIPPINE STATISTICAL ASSOCIATION, INC
Office of Publication RM 214 PSSC BLDG., COMMONWEALTH AVE., DILIMAN, QUEZON CITY

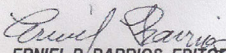
If publication is owned by a corporation, stockholders owning one percent of more of the amount of stocks NOT APPLICABLE Bond Holders, mortgage or other security holders owning one percent or more of the total amount of Security NOT APPLICABLE

In case of daily publication average number of copies printed and circulated of each issue during the preceding month of NOT APPLICABLE

1. Sent to paid subscribers _____
 2. Sent to other than paid subscribers _____
- Total _____

In case of publication other than daily total number of copies printed and circulated at the last issue date 2015

1. Sent to paid subscribers 200
 2. Sent to other than paid subscribers 100
- Total 300

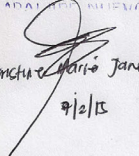

ERNIEL B. BARRIOS, EDITOR

Designation and Signature

SUBSCRIBED AND SWORN TO, before me this JUL 0 2 20 day of 20 in the City of MAKATI CITY, Philippines Affiant exhibiting to me his/her UP ID No. 021032410 issued at Quezon City, on 2014.

Doc. No. 536
Page No. 108
Series No. 1
Series of 2015

JOSE L. NOTARY PUBLIC
NOTARY PUBLIC - MAKATI CITY
COMMISSION NO. 14168
UNTIL DECEMBER 31, 2016
BP NO. 0481345 PROL NO. 3283
TR NO. 4750464 Makati J.S.B.
459 CAMINO DE LA FE ST
ADALDO BUENO, MAKATI

Rec. by  Notary Public

9/2/15

Please accomplish in three (3) copies

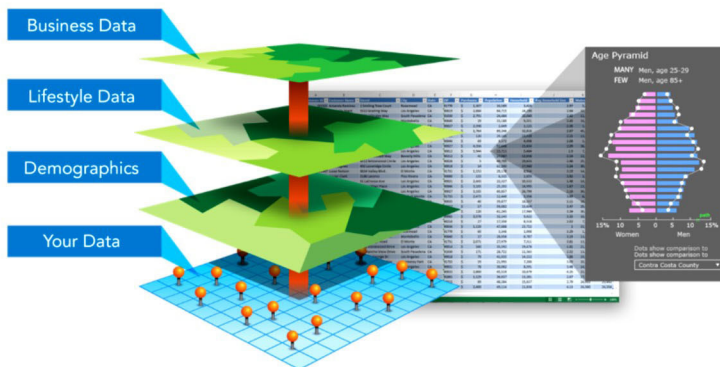
rvmyap314

The Leading GIS and Mapping Solutions Provider in the Philippines

Geodata Systems is an information technology company focused on providing geographic information systems (GIS) and mapping solutions in the Philippines since 1991. Geodata is the Sole Authorized Philippine Distributor of Environmental Systems Research Institute (Esri) — the global company that has been at the forefront of GIS for more than 40 years. Geodata Systems is an ISO 9001:2008 Quality Management Systems Certified company.

PRODUCTS:

- Esri's Complete ArcGIS Suite of Products
- Esri Solutions with Imagery
- Trimble GPS Devices for GIS & Mapping
- Harris' ENVI software for Image Processing and Analysis
- Schneider Electric's ArcFM for Facilities Management Solution for Utilities
- Innovyze's Water Distribution Modeling and Management Solution for ArcGIS
- GeoDRRM Appliances
- Digital Philippine Map, Philippine Digital Atlas, Address Maps, Metro Manila Street Maps



Advancing Statistics with the use of a Location Platform is the way to go to support the accomplishment of our country's shared Sustainable Development Goals in 2030.

This location platform will greatly enhance the planning, production, and dissemination of statistical information products visualized and accessed through map representation of statistics.

For more information, contact us:



Geodata Systems Technologies, Inc.
19/F Strata 100 Building, F. Ortigas Jr. Road, Ortigas Center, 1605 Pasig City
Phone: +632 637 4447 to 49; 633 7787 to 88 Fax: +632 633 6873
E-mail: marketing@geodata.com.ph Website: www.geodata.com.ph



Visual Analytics

Opportunity
at your fingertips.

The answers you need, the possibilities you seek – they're all in your data. SAS® helps you quickly see through the complexity and find hidden patterns, trends, key relationships and potential outcomes. Then easily share your insights in dynamic, interactive reports.



Try Visual Analytics and see for yourself
sas.com/VAdemo



SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. © 2016 SAS Institute Inc. All rights reserved. 511779NLS 0015



Purchasing Managers' Index (PMI) Philippines

A Leading Economic Indicator

**A product of
Foundation of the Society of Fellows
in Supply Management (SOFSM)
and
Philippine Institute for Supply Management (PISM)**

Distributed By:

I-Metrics Asia-Pacific Corporation
info@imetricsasia.com

3F Unit B&F Raha Sulayman Bldg.
#108 BEnavidez St. LGaspi Village, Makati City
Contact Nos. : +632-750-1468 | +632-887-1166

Editorial Board

Editor-in-Chief

Erniel B. Barrios, University of the Philippines Diliman

Associate Editors

Lisa Grace S. Bersales, University of the Philippines Diliman and
Philippine Statistics Authority

Zita V. Albacea, University of the Philippines Los Baños and
Philippine Statistical Research and Training Institute

Alex C. Gonzaga, University of the Philippines Manila

Joseph Ryan G. Lansangan, University of the Philippines Diliman

Advisory Board

Roberto S. Mariano, University of Pennsylvania

Bimal K. Sinha, University of Maryland Baltimore County

Alexander R. De Leon, University of Calgary

Managing Editor

Angela D. Nalica, University of the Philippines Diliman

The Philippine Statistician is the official scientific journal of the Philippine Statistical Association. Inc. (PSAI). It considers papers resulting from original research in statistics and its applications. Papers will be sent for review on the assumption that this has not been published elsewhere nor is submitted in another journal.

Copyright 2015 by the Philippine Statistical Association, Inc., PSSCenter, Commonwealth Ave., Diliman, Quezon City.

ISSN 2094-0343

PHILIPPINE STATISTICAL ASSOCIATION, INC.
THE BOARD OF DIRECTORS and OFFICERS for 2015

MEMBERS OF THE BOARD

Individual Members

Zita VJ. Albacea
Carmelita N. ERICTA
Ana Maria L. Tabunda

Jose Ramon G. Albert
Diwa C. Guinigundo
Winecito L. Tan

Imelda E. De Mesa
Iluminada T. Sicat

Institutional Members

ABS-CBN Broadcasting Corporation
Ateneo Social Science Research Center
Bangko Sentral ng Pilipinas
Philippine Statistics Authority
Philippine Statistical Research and
Training Institute
SAS Institute (Philippines), Inc.

Represented by
Ruthie A. Floresta
Michael A. Cuesta
Rosabel B. Guerrero
Candido J. Astrologo Jr.

Virginia N. Gañac
Maxima V. Garin

PSAI Regional Chapter

Represented by
Milagros R. Baldemor
President, PSAI Region 1

EXECUTIVE COMMITTEE AND OFFICERS

President
Vice President
Secretary
Treasurer
Ex officio Member
(Immediate Past President)

Jose Ramon G. Albert
Ana Maria L. Tabunda
Winecito L. Tan
Diwa C. Guinigundo
Nieves L. Osorio

WORKING COMMITTEES IN 2015

Advocacy
Chair: Jose Ramon G. Albert
Annual Conference
Chair: Romeo S. Recide
Co-chair: Carmelita N. ERICTA
Co-chair: Milagros R. Baldemor

Finance
Chair: Diwa C. Guinigundo
Membership
Chair: Cynthia C. Vallesteros
Co-chair: Chona P. Guatlo

Annual Meeting and Christmas Party
Chair: Candido J. Astrologo Jr.
Co-chair: Ruben V. Litan

Nomination and Election
Chair: Ludivinia D. Gador
Co-chair: Ester P. Abitona

Institutional Development
Chair: Ana Maria L. Tabunda
Co-chair: Maxima V. Garin

Statistics Education
Chair: Zita V.J. Albacea
Co-chair: Isidoro P. David

Institutional Training
Chair: Imelda E. de Mesa
Co-chair: Iluminada T. Sicat

Publications and Information
Chair: Erniel B. Barrios
Editor, The Philippine Statistician: Erniel B. Barrios

Search and Awards
Chair: Winecito L. Tan

Website Administrator: Erniel B. Barrios
Book Review: Arturo Y. Pacificador Jr.

PSAI SECRETARIAT

Managing Director: Teresita G. Abad
Secretariat Assistants: Stephanie Rae C. Andres and Reynaldo B. Belluga Jr.