Water Quality and Monitoring Track
Journal of South Carolina Water Resources
Volume 2, Issue 1, Pages 80-85, 2015

Model Results and Software Comparisons in Myrtle Beach, SC Using Virtual Beach and R Regression Toolboxes

Matthew J. Neet[1], R. Heath Kelsey[2], Dwayne E. Porter[3], Dan W. Ramage[4], and Adrian B. Jones[5]

AUTHORS: [1]Research Associate, Arnold School of Public Health, University of South Carolina, Columbia, SC 29208, USA. [2]Program Director, Integration and Application Network, University of Maryland Center for Environmental Science, Cambridge, MD 21613, USA. [3]Associate Chair, Arnold School of Public Health, University of South Carolina, Columbia, SC 29208, USA. [4]Systems Programmer II, Arnold School of Public Health, University of South Carolina, Columbia, SC 29208, USA. [5]Web Developer, Integration and Application Network, University of Maryland Center for Environmental Science, Cambridge, MD 21613, USA.

Abstract. Utilizing R software and a variety of data sources, daily forecasts of bacteria levels were developed and automated for beach waters in Myrtle Beach, SC. Modeled results are then shown for beach locations via a website and mobile device app. While R provides a robust set of tools for use in forecast modeling, the software has an extensive learning curve and requires skilled statistical interpretation of results. The Environmental Protection Agency (EPA) created the "Virtual Beach" software package to address these concerns. To evaluate the utility of the more user-friendly Virtual Beach modeling toolbox, predictive models were developed and model results were analyzed using the two software suites. Recommendations were made based on ease of use and several performance measures. Model results indicate the two software toolboxes yield comparable outputs. However, Virtual Beach tends to create more robust model forecasts, while R provides more options for model setup and outputs.

INTRODUCTION

As more people live, work, and play in coastal areas, an increasing need exists to provide robust and timely measures of potential illness risk from fecal water pollution, while ensuring that local economies are not harmed by unnecessary beach closures and advisories. To help accomplish this goal, new forecast tools were developed through the collaborative efforts of the University of South Carolina (USC) Arnold School of Public Health, University of Maryland Center for Environmental Science (UMCES), and National Oceanic and Atmospheric Administration (NOAA). Eight beaches (Figure 1) in the Myrtle Beach Grand Strand area of South Carolina now have daily forecasts for bacteria concentration in swimming waters. Radar-based rainfall estimates and coastal ocean observing system platforms provide real-time environmental data used in these new tools. Enterococci concentration estimates are provided in near real-time. These estimates (forecasts) are then uploaded to a database linked to a website and mobile device application. From here, bacteria concentrations and swim advisories can be seen and compared to EPA water quality criteria for swimming safety.

Previous research and bacterial estimates relied on weekly monitoring program results and a network of rain gauges (Johnson 2007; McDonald 2006). The near real-time models analyzed here offer many advantages and advances over existing monitoring and assessment approaches. First, remote sensing allows rainfall data to be collected and averaged over watersheds. According to Kelsey et al. (2010), areally averaged rainfall values provide more predictive capability for bacteria concentrations than point estimates obtained from rain gauges. Second, remotely sensed data products can be collected, collated, and processed in automated fashion. Computed bacteria concentration estimates can be provided daily and without the need for costly and maintenance intensive rain gauges.

Figure 1. Locations of sampling sites and model areas.

Alternative technologies and software tools have been utilized to model bacteria in coastal waters. EPA's Virtual Beach (VB) software suite was developed for beach recreation areas. This software package provides many statistical tools needed for beach modeling including several of the tools used in previous Myrtle Beach forecasting efforts. In conjunction with the EPA, a need was identified to compare the performance of the existing Myrtle Beach models with those derived from VB.

The purpose of this project was to compare and contrast R and VB modeling software packages in terms of model development procedures and performance results. The Virtual Beach software package is designed to be relatively simple to use by those without statistical background. If the models developed using VB had similar predictive power to those developed using a more manual process in R, it would suggest that VB is a useful tool for developing predictive models for beach bacteria. Bacteria prediction results and the processes used to derive them were analyzed quantitatively and qualitatively when developing new predictive models in the Grand Strand.

METHODS

Data for this analysis were previously collected and summarized as part of a beach water quality prediction project. Data were collected in 2006, 2007, and 2009. These data represented many input and survival factors (Figure 2) necessary for the propagation of bacteria in marine waters. They were collected weekly and were representative of a wide variety of climate and environmental conditions. A common set of data (bacteria concentration, remotely sensed, modeled, and observing system data from varied sources [Table 1]) were included in the models. Enterococci bacteria concentration (culture forming units [CFU]) data were collected approximately weekly from the mid-May to mid-October beach swimming season. These data were compiled into a single .csv file for use in the following modeling processes. In both modeling efforts, multiple linear regression (MLR) was used to analyze multiple explanatory variables.

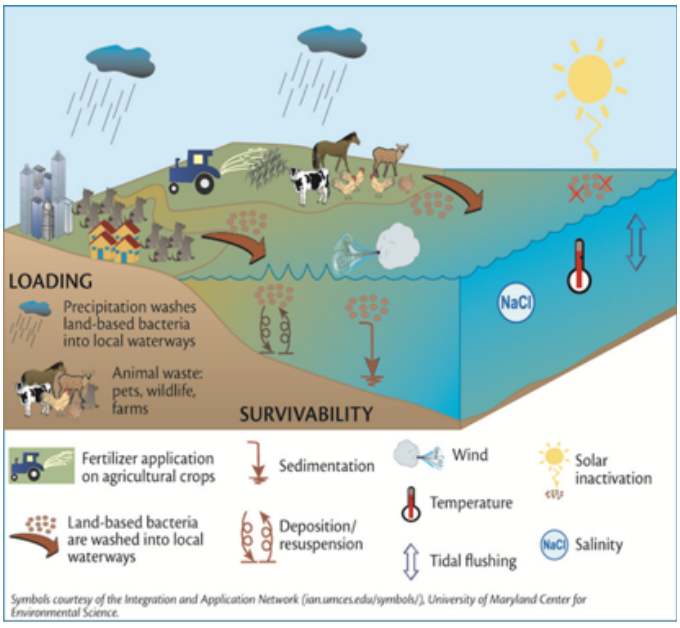

Figure 2. Input and survival factors for bacteria (Kelsey et al. 2010).

Table 1. Remotely sensed, modeled, and observing system independent variables used in the comparisons.

Remotely Sensed/Modeled/Observing System Data
NEXRAD rainfall data
Radar rainfall summaries (24, 48, 72, 96, 120, 144, 168 hours)[1]
24-hour rainfall totals (1, 2 and 3 days)[1]
Number of dry days[1]
Maximum intensity of rainfall 24 hours[1]
Salinity[2]
Tide stage[2]
Water level[2]
Wind speed[2]
Wind direction[2]
Water temperature[2]
[1]prior to sample date
[2]nearest recording station and/or Sun 2 ocean buoy

R Model Development

R, a free statistical software suite, is command-line oriented and must utilize the R language, similar to the S coding of S-Plus. R is open-source and supported and documented by a large user-base (Revolution Analytics 2015; R Core Team 2013).

In R, all potential parameters/predictors for the dependent variable were utilized. The dependent variable, Enterococci concentration, was log transformed to approximate a normal distribution and facilitate further standard statistical analysis. Data were imported via the common .csv file. Sample stations were reassigned as categorical variables so they could be analyzed as potential predictors. To compare results, the "relevel()" command in R was used in the categorical analysis of station location. This allowed the same sample stations to be used for model development in R and VB. No other data pre-processing was performed.

Models were then developed for each of the eight beach regions using linear regression. These locations were delineated based on South Carolina Department of Health and Environmental Control (SCDHEC) sampling station groupings. A backwards, manual selection process was used. The lm, or linear model, function in R was employed. Model "lm is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance…" (R Core Team 2013). Variance inflation, parameter p-value, and model Bayesian Information Criterion (BIC) were used in selecting the models with the highest

predictive power. Because many of the predictors were related (e.g., rainfall averages of different length), variance inflation was evaluated. By deleting parameters with high Variance Inflation Factor (VIF) values (> approximately 10) in the model, unpredictable variance was kept to a minimum. Model selections proceeded by systematically removing parameters from the model until parameter p-values were approximately less than 0.10. BIC was used to evaluate remaining model parameters by removing parameters individually and exploring their effects on BIC. A lower BIC value was more desirable than a higher one. Final models retained parameters with variance inflation values less than 10, p-values generally less than 0.05, and lowest possible BIC values.

Virtual Beach Model Development

The EPA developed Virtual Beach 3 as a decision support tool incorporating suite of statistical software (Cyterski et al. 2013). The tool allows decision-makers and beach managers to predict fecal bacteria concentration using linear relationships between independent and dependent parameters. VB provides a list of model outcomes for the user to analyze (Cyterski et al. 2013).

VB 3 and 2.2 Users' Guides (Cyterski et al. 2013; Cyterski et al. 2012) were utilized as outlines for developing models in VB. The same .csv data file used to develop models in R was analyzed. Dummy variables were created to test whether sample location, a categorical variable, was significant in model predictions. Data were imported and "validation" procedures were performed. Blank columns, rows, columns with missing data, or non-numeric records were deleted. Next, study sites were located along their respective beaches. A map feature, using Google Earth, was provided and an orientation box was created. From this box, an angle was generated which allows a wind, wave, and/or current component to be calculated and used in the modeling process. Since wind speed and direction were collected in the initial dataset, a wind component was generated for wind values perpendicular to the shore (O) and along the shore (A).

Multiple linear regression options were run on both standard and transformed (independent variables) datasets. The standard dataset included raw data with only wind components added. The transformed version contained independent variables that were transformed (e.g., Log10, ln, inverse, square, square root, quad root, polynomial, and exponential functions) and included if they met a 25% threshold for the Pearson correlation coefficient with respect to the dependent variable.

Using the MLR tab, independent variables were chosen in the variable selection tool under model settings. Model fitness can be analyzed using any one of ten model evaluation criteria (e.g., R2, adjusted R2, AIC [Akaike's Information Criterion], BIC, Sensitivity, etc.) under the Control Options tab. BIC was chosen because it tends to limit over-fitting, keeping the number of variables in the model small (Cyterski 2013). Then, VIF levels were set to a maximum of 10 (VB can monitor this automatically). By checking the "Run all combinations box" under the manual option for linear regression modeling and clicking the "Run" button, VB evaluates models generated with all possible combinations of predictors. VB then automatically selects the 10 models with the best performance as determined by the evaluation criterion. The best model, having the lowest BIC (and, in general, the highest adjusted R2), was selected for further evaluation and comparison to the models developed in R.

Performance Metrics

AIC, BIC, adjusted R2, cross validation Mean Square Error of Prediction (MSEP), and Receiver Operator Characteristic curve (ROC) area under the curve (AUC) were used to compare performance of the models developed in R and VB. AIC, BIC, and adjusted R2 values help determine if additional parameters add predictive capacity to the model given the uncertainty introduced by adding an additional predictor. Cross validation allows evaluation of a fixed set of parameters in the final model; it uses random subsets of the original data set to develop parameter estimates and uses the remaining data to validate and compare observed values to the values predicted by the model. ROC curves (like those displayed Figure 3) were utilized to compare true positive to false positive values generated by the model. Curves like those seen in Figure 3 with high true positives (high sensitivity), low false positives (high specificity), and a steep transition are desired. Curves are compared by calculating the AUC. A perfect model would have an AUC=1, and a model with no predictive capability would have an AUC=0.5 (Morrison et al. 2003). In Figure 3, 2.02 represents the log10(104), where 104 is the Enterococci concentration guideline for recreation. The color code and the right scale represent the false positive and true positive rates at a particular decision point. Red represents a decision point approaching 2.7, where false positive and true positive rates are both 0. Blue represents the false positive and true

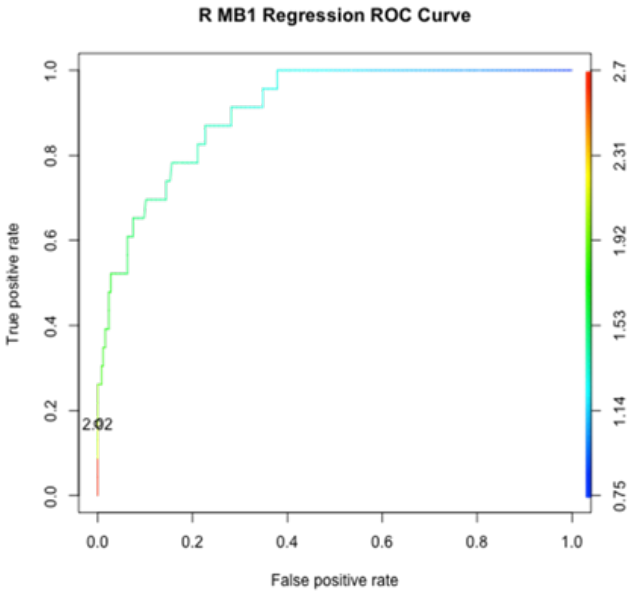

Figure 3. ROC curve for the MB1 site.

positive rates approaching decision point 0.75, where false positive and false negative rates are 1. This can be used to determine the decision rule at an acceptable false positive and false negative rate.

Following evaluation of all model criteria (AIC, BIC, R2, adjusted R2, MSEP, and ROC area) a matrix was generated to compare performance metrics for models at all locations developed in R and VB (Table 2). Each model was given a score of 0, 0.5, or 1 based on a comparison of performance metric values. A score of 1 was given to the most desirable metric value, while the least desirable was scored 0. Where two models tied for the most desirable metric value, a score of 1 was given to both while the remaining model was given a score of 0. Scores for each set of models were tallied. The model with the highest overall point value would represent the model with overall best performance.

A qualitative assessment of the modeling process was also performed. Overall software utility and methodology were evaluated. Ease of use, flexibility, utility of inputs/outputs, etc. were evaluated for R and VB. Each software package was analyzed for simplicity, learning curve required, flexibility of input data and output results, and the overall usefulness of the software.

RESULTS

Results and performance metrics for each model are summarized in Table 2. When first run in R, values for AIC, BIC, and cross validation were very different from VB. This was likely a result of the pre-processing step that VB uses to remove records with missing values for any potential parameters. In R, missing values were removed systematically, only removing records that have missing values for the parameters used in the model. To standardize comparisons, the dataset generated by the pre-processing step in VB was also used in R, resulting in identical data inputs. Model scores were generally highest for the VB model developed with transformed data, next highest for the models generated in VB with non-transformed data, and lowest for the models generated in R. Based on Table 2, VB transformed had a summed score of 37, VB was 21, and R was 16.5. The table also shows the VB transformed column having more green (highest point value) than either of the other two columns, while the R column had more red (no point value) than the other columns.

Table 2. Performance analysis matrix.

Region	Criteria	VB Transformed	VB	R
NMB2	AIC	141	154	158
	BIC	160	171	171
	R^2	0.15	0.08	0.05
	AdjR^2	0.13	0.07	0.04
	MSEP	0.1552	0.1483	0.3657
	ROC area	0.8119	0.8730	0.7959
NMB3	AIC	162	182	178
	BIC	185	202	214
	R^2	0.28	0.20	0.25
	AdjR^2	0.26	0.18	0.22
	MSEP	0.2108	0.1999	0.3759
	ROC area	0.8458	0.7604	0.8087
MB1	AIC	254	265	253
	BIC	283	287	296
	R^2	0.32	0.28	0.34
	AdjR^2	0.30	0.27	0.31
	MSEP	0.2609	0.2561	0.3914
	ROC area	0.8691	0.8704	0.9064
MB2	AIC	237	271	272
	BIC	268	288	299
	R^2	0.31	0.17	0.19
	AdjR^2	0.28	0.16	0.17
	MSEP	0.2805	0.2494	0.4412
	ROC area	0.9240	0.8263	0.8880
MB3	AIC	270	277	269
	BIC	287	292	289
	R^2	0.27	0.21	0.28
	AdjR^2	0.24	0.19	0.25
	MSEP	0.2805	0.2494	0.6713
	ROC area	0.7626	0.7296	0.7632
MB4	AIC	228	235	234
	BIC	250	257	269
	R^2	0.35	0.33	0.36
	AdjR^2	0.33	0.31	0.32
	MSEP	0.3987	0.3870	0.4726
	ROC area	0.8791	0.8791	0.8806
Garden City	AIC	210	218	235
	BIC	236	240	254
	R^2	0.40	0.37	0.29
	AdjR^2	0.38	0.35	0.28
	MSEP	0.4031	0.3887	0.4712
	ROC area	0.8867	0.8747	0.8294
Surfside	AIC	524	524	548
	BIC	557	557	570
	R^2	0.40	0.40	0.34
	AdjR^2	0.39	0.39	0.33
	MSEP	0.8215	0.8215	0.6417
	ROC area	0.8385	0.8385	0.7653
	Green (1pt)	29	13	11
	Yellow(.5pt)	16	16	11
	Red (0pt)	3	19	26
	Total Points	**37**	**21**	**16.5**

DISCUSSION

For investigations of Enterococci bacteria in beach applications, VB and R software can be useful for regression analysis and bacteria predictions for differing reasons; each has its strengths and weaknesses.

Quantitative Comparisons

Performance comparisons suggest that VB can generate more robust models than the simple linear regression manual selection techniques used in R for this assessment. The features of transforming variables and model comparisons using all potential prediction combinations used in VB can somewhat be reproduced in R, but is probably unnecessary,

as these features are built in to the current version of VB. Most importantly, the quantitative comparisons suggest that model development can be improved by using input data sets with predictors that are transformed to create linear relationships with the dependent variable, and using a model selection technique that evaluates all potential combinations of the model parameters.

Qualitative Comparisons

VB and R offer many benefits to potential users. While model results were somewhat comparable, the manner in which model predictions were derived is different. VB enables users to create robust models by running all possible variable permutations. It provides options for transforming independent variables and/or calculating wind A/O values. The VB tool also has an easy to learn graphical user interface (GUI) that utilizes self-explanatory tabs for major functions. VB requires no programming skill and is fairly easy to learn. VB provides users with a no-cost option to expensive commercial-off- the-shelf software tools.

In comparison, R requires use of a command-line programming language and scripting ability. To become proficient in R, time and resources are necessary and would be required to replicate some of the VB options employed here (e.g., calculating potential predictor permutations, transformation of independent variables, etc.). However, R provides some flexibility and options that are currently not available in VB, including automating data input/output, direct linkage to databases, and flexibility in generating descriptive visuals and graphical output. Additionally, predictive models can be developed using a variety of advanced methods in R, and many others are developed every year. Currently, MLR, partial least squares (PLS), and gradient boosting machine (GBM) options are the only options available in VB.

Contributions to the Field

Over the last fifteen years, predictive models for Escherichia coli and Enterococci concentrations have been developed for fresh and marine waters (respectively). Francy et al. (2013) showed that relationships between bacteria concentrations and environmental variables could produce models for use in making near real-time forecasts at inland beaches. Work conducted by Paule et al. (2014) and Francy et al. (2006) utilized MLR analysis to model bacteria from environmental, water, and hydrological data. MLR was utilized by Paule et al. (2014) to determine which hydrogeological factors impacted indicator bacteria concentrations most. Francy et al. (2006) indicated MLR allowed for the determination of beach-specific explanatory variables. Employing similar MLR procedures to evaluate the best variables for bacteria concentration predictions, we also found explanatory variables are unique to beach location. Bacterial models were even developed by Frick et al. (2008) utilizing the VB toolset. Here, weather and environmental data were processed by VB's MLR tool (similar to our efforts in Myrtle Beach) to yield now-casts and forecasts of bacterial concentrations for Huntington Beach, Lake Erie (Frick et al. 2008). Additional modeling efforts incorporated PLS techniques to predict bacteria concentrations and produced similar results to regression efforts (Brooks et al. 2012). The Brooks et al. (2012) study even led to the incorporation of its PLS techniques in VB. The bacterial modeling field continues to expand its statistical modeling tools in an effort to increase accuracy, functionality, and usefulness of predictions for forecasts.

The results of this study are not shocking or ground-breaking. They do, however, reaffirm the importance of making accurate and timely estimates of bacteria in beach waters where permanent swimming advisories may not be in place (e.g., Florida beaches, where sampling is utilized to monitor bacteria levels) to ensure public safety. In SC, these results suggest that SCDHEC could remove permanent advisories and use the model results to determine when advisories should be issued for a particular site. The methodologies and comparisons highlighted in this study can certainly be applied in other beach areas. By utilizing VB, R, MLR, etc., accurate and precise forecasts can be employed by beach managers to ensure public health is impacted minimally. These tools and methodologies can be added to and extend the capabilities of any beach manger's toolbox.

CONCLUSION

Overall, VB is recommended for model development in situations where programming skill is limited. If descriptive graphics and multiple input/output functions are needed, R software should be utilized. To match R's automated data integration, additional programming, support, and funding of VB are recommended to increase tool functionality. The geographic footprint and ensemble modeling approach used here continues to expand; most notably with freshwater bacterial modeling recently completed in the Lower Saluda River of South Carolina and Enterococci concentrations currently being modeled in southwest Florida.

ACKNOWLEDGMENTS

We would like to acknowledge the support we received from NOAA, the EPA and SECOORA, without which this initiative would not have been possible. We would like to thank Dr. Cyterski and the Athens, GA EPA office for meeting with us and answering our VB questions. We would also like to thank the reviewers of this article for their time and thoughtful comments.

LITERATURE CITED

Brooks, W., M. Fienen, and S. Corsi, 2012. Partial Least Squares for Efficient Models of Fecal Indicator Bacteria on Great Lakes Beaches. *Journal of Environmental Management*. 114:470-475. http://dx.doi.org/10.1016/j.jenvman.2012.09.033.

Cyterski, M., 2013. Personal communication: meeting. August.

Cyterski, M., W. Brooks, M. Galvin, K. Wolfe, R. Carvin, T. Roddick, M. Fienen, and S. Corsi, 2013. *Virtual Beach 3: User's Guide*. EPA/600/R-13/311. USEPA, Athens, GA.

Cyterski, M., M. Galvin, R. Parmar, and K. Wolfe, 2012. *Virtual Beach version 2.2 User's Manual*. EPA/600/R-12/024. USEPA, Athens, GA.

Fancy, D., E. Stelzer, J. Duris, A. Brady, J. Harrison, H. Johnson, and M. Ware, 2013. *Applied Environmental Microbiology*. March. 79(5):1676-1688.

Fancy, D., R. Darner, and E. Bertke, 2006. Models for Predicting Recreational Water Quality at Lake Erie Beaches. Scientific Investigations Report 2006-5192. U.S. Geological Survey. 13pp.

Frick, W., Z. Ge, and R. Zepp, 2008. Nowcasting and Forecasting Concentrations of Biological Contaminants at Beaches: A Feasibility and Case Study. *Environmental Science and Technology*. 42(13): 4818-4824.

Johnson, E., 2007. Predictive modeling of enterococcus concentrations at South Carolina tier I beaches. Thesis. University of South Carolina. Columbia, SC. 237pp.

Kelsey, R. H., G. I. Scott, D. E. Porter, D. Edwards, and T. C. Siewicki, 2010. Improvements to shellfish harvest area closure decision-making using GIS, remote sensing, and predictive models. On-line. *Estuaries and Coasts,* 33:712-722. DOI: 10.1007/s12237-010-9264-7

McDonald, E., 2006. Application of ocean observing systems in aiding predictive water quality modeling in Long Bay, South Carolina. Thesis. University of South Carolina. Columbia, SC. 267pp.

Morrison, A. M., K. Cloughlin, J. P. Shine, B. A. Coull, and A. C. Rex, 2003. Receiver Operator Characteristic Curve Analysis of Beach Water quality indicator variables. *Applied and Environmental Microbiology*, 69(11):6405. DOI: 10.1128/AEM.69.11-6411.2003.

Paule, M., S. Memon, B. Lee, U. Raja, C. Sukhbaatar, J. Ventura, D. Jahng, J. Kang, and C. Lee, 2014. Statistical Evaluation of Intra-event Variability of Fecal Indicator in Stormwater Runoff from Different Land uses. *International Environmental Modelling and Software Society (iEMSs)*. 7th Intl. Congress on Env. Modelling and Software. San Diego, CA, USA. Daniel P. Ames, Nigel W.T. Quinn and Andrea E. Rizzoli (Eds.). http://www.iemss.org/society/index.php/iemss-2014-proceedings.

R Core Team, 2013. R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org/.

Revolution Analytics, 2015. What is R? Revolution Analytics - a wholly owned subsidiary of Microsoft. Redmond, WA. http://www.revolutionanalytics.com/what-r. Accessed November 10, 2015.