*exception reporting, outbreak, surveillance, prediction*

Gavin MCCABE[*], David GREENHALGH[*], George GETTINBY[*],
Eileen HOLMES[**], John COWDEN[***]

# PREDICTION OF INFECTIOUS DISEASES:AN EXCEPTION REPORTING SYSTEM

In this paper prediction methods are discussed in the context of developing an exception reporting system for laboratory reports. The detection of outbreaks and longer term trends is briefly addressed, before a consideration of data types and availability to be used in evaluating the prediction methods. Four general prediction methods are outlined and the selection of data to which they are applied is examined. Both real and simulated data are used to evaluate the prediction methods and a strategy for an exception reporting system is proposed.

## 1. INTRODUCTION

There has recently been a growing expectation that institutional bodies should be adequately prepared for, and respond rapidly to, events which impinge on public life. In few areas is this more true than in public health and it is clear that the early detection of possible outbreaks of infection is vital to this objective.

In the UK, diagnostic microbiology laboratories participate in a voluntary system of reporting organisms identified from samples taken from patients for clinical reasons when they present to their family doctors or hospitals. In Scotland, these reports are collated by the Scottish Centre for Infection and Environmental Health (SCIEH), and in England and Wales by the Communicable Disease Surveillance Centre (CDSC). Although these reports are valuable there is no formal threshold of what constitutes an outbreak or a potential outbreak. While an outbreak may be defined as "an epidemic limited to localised increase in the incidence of a disease" [1], this begs the question of what is an epidemic. In fact, in practice, an outbreak is simply "more reports than would be expected". At national level, therefore, outbreak detection depends upon the expert, but subjective, judgement of staff in the national surveillance centre. On identifying a potential outbreak staff access routinely available information from each report, to assess whether or not further investigation is warranted. The combination of limited resources and increasingly vast numbers of different organisms being reported each week means that the development of more

[*] Department of Statistics and Modelling Science, University of Strathclyde, Livingstone Tower, 26 Richmond Street, Glasgow, G1 1XH, U.K.

[**] R&D Support Unit, Clinical Sciences Building, Salford Royal Hospital NHS Trust, Stott Lane, Salford, M6 8HD

[***] Scottish Centre for Infection and Environmental Health, Clifton House, Clifton Place, Glasgow, G3 7LN, U.K.

automated detection systems is vital in order to support consultants and epidemiological practitioners in their surveillance activities.

However, in attempting to detect potential outbreaks how can a decision routinely be made whether or not it is appropriate to raise an alarm? A system that repeatedly yields false alarms will eventually be ignored by practitioners, but the failure to detect even the clearest outbreak renders the system redundant. With thousands of different organisms monitored nationally on a weekly basis, it is impractical to design systems specifically tailored to volumes of individual organisms and thus the development of generically applicable detection procedures is essential.

While there have been published methods looking at the automated detection of possible outbreaks for individual organisms or health events [2-4], relatively few exist that deal with the diversity of organism characteristics presented to national centres [5-7]. Regardless of the breadth of their focus, most published detection systems can be broken down into two main components: the first element predicts the expected count for a specific organism for a given week in the future; the second element makes a comparison of the observed and predicted counts, and on the basis of this comparison makes a decision whether or not to trigger an alarm. Such systems should not be called ‘outbreak detection systems’ as outbreak detection is not the role of an automated system, but rather the responsibility of experts in the field related to each organism. The term ‘exception reporting system’ is more accurate as its function is to report to practitioners exceptional data points, which can then be further investigated if warranted.

In this paper approaches to the development of the prediction element of an exception reporting system are discussed and, using data from the Scottish Centre for Infection and Environmental Health, an evaluation of how such procedures can be used to make useful predictions as part of an overall, generic and automated system is given.

## 2. MATERIALS

Taking into account all the various typings and sub-typings of organisms, SCIEH’s databases currently contain information on 2364 different organisms. These databases are updated on a weekly basis with reports from the diagnostic microbiology laboratories throughout Scotland. Most reports arise from isolates from specimens sent to these laboratories from general practitioners, hospitals, environmental health officers and others for clinical reasons. For each report, information is provided on the case’s sex, age, geographical location, as well as the timing of the report, the reporting laboratory and health board, and additional clinical information. These resultant reports are grouped into six broad areas: *Escherichia coli* infections, non-viral infections, viral infections, *Salmonella* infections, tuberculosis infections and samples taken from a veterinary source. Examples of the observed reports for two organisms can be seen at the top of Figures 1 and 2 at the end of this paper.

In order to facilitate the testing of various prediction methods it was necessary to use both real and simulated data. The need for simulated data arose due to the absolute control over the patterns and characteristics of the data that simulation yields, while the need for validation on real data is immediately clear. However, due to the sheer number of different organisms available, it was necessary to reduce the number of real data series used and make a selection of organisms that would characterise in some broad sense the diversity of features exhibited in the range of data. Of the 2364 organisms available within SCIEH's databases, only 70 have an average weekly report rate that is greater than or equal to one, for the period from 1990 to 2001 inclusive. Each of these 70 organisms were selected as it was felt by practitioners that exception reporting was most valuable when applied to the more common organisms, as opposed to the rarer organisms within which points of exception are generally much more evident and identified much more rapidly. Alongside those 70 organisms, a further 35 were selected pseudo-randomly from the range of organisms whose average weekly report rate was below one, selected in such a manner so as to reflect the distribution of organisms across the scale from zero to one.

These 105 organisms were used to both provide real data on which to test the prediction methods, and to simulate data again for testing the prediction methods. Four data characteristics stood out as clearly requiring explicit attention in order to make useful predictions based on historical data for a generic, automated exception reporting system:

1) the average weekly reporting rate;
2) the magnitude of trend seen across the years of data;
3) the magnitude of seasonality seen within the data; and
4) whether or not outbreaks were present within the historical data.

Each of the 105 organisms was classified on the basis of their average weekly reporting rates over the period 1990 to 2001, and their trends and seasonalities were highlighted using correlograms. This use of correlograms was particularly important for the rarer organism for which any seasonal pattern is frequently much more difficult to observe by eye. It became clear that, as well as the cases where no trend or seasonality was evident, there were two distinct magnitudes of both trend and seasonality, now categorised as 'strong' and 'weak'. As might be expected, across the range of reporting rates (zero to above 70), the numerical magnitude of seasonal peak associated with e.g. strong seasonality would vary. In order to create sensible simulated data, relationships were drawn between the reporting rate and numerical magnitude of each trend and seasonality category using the 105 selected organisms. The seasonal amplitude was found to be given by MATERIALE × $rate^{\delta}$, where $\gamma$ was 272.14 and $\delta$ equalled 0.2758 for the weak category and the strong category was specified by a $\gamma$ value of 239.99 and a $\delta$ value of 0.7021. The gradient of the trend was specified by $\alpha + \beta \times rate$, where $\alpha$ was 0.0031 for the weak category and 0.0054 for the strong, and $\beta$ equalled 0.0015 and 0.0035 for the weak and strong categories respectively. Obviously the magnitude of outbreaks included within the simulated data was much more arbitrary, but nevertheless their inclusion was important so as to isolate the effect historical outbreaks had on the various prediction methods tested. Twenty years of data were simulated for each combination of weekly reporting rate (0.1, 0.5, 1, 3, 5, 10, 20 and 50), trend (none, weak and strong), seasonality (none, weak and strong) and outbreaks (included or not included).

## 3. PREDICTION APPROACHES TO ANALYSIS

### 3.1. PREDICTION METHODS

Examining previously published detection systems (both those applied to single organisms or health events, and those few that were applied to a broader range of organisms), a variety of prediction methods were isolated as being of potential use for a generic, automated exception reporting system. While many detection systems used prediction methods that were far too dependent on user input or computationally intensive, four main prediction methods came to the fore with a mixture of complexities, namely:

1) exponentially weighted moving averages (EWMAs);
2) the mean of selected historical data;
3) zero-inflated Poisson modelling (ZIP); and
4) generalised linear modelling (GLM).

EWMA, also known as exponential smoothing, is a method that has been used in a variety of areas including nosocomial infection surveillance data in connection with suspected outbreaks of gentamicin resistance among *Pseudomonas aeruginosa* bacteria [8]. Exponential smoothing derives its name from the fact that it is formed by a weighted moving average that has geometric weights that lie on an exponential curve, i.e. weights which decrease by a constant ratio each step back [9]. Therefore the one-step-ahead forecast made at time N for a time series, $x_1$, $x_2$, …, $x_N$, is given by

$$\alpha x_N + \alpha(1-\alpha)x_{N-1} + \alpha(1-\alpha)^2 x_{N-2} + \cdots + \alpha(1-\alpha)^N x_1$$

where $\alpha$ is the smoothing constant, taking values between zero and one.

The process of taking the mean requires no introduction or explanation. However the data of which the mean was taken is of greater interest. Unlike the EWMA method which was applied to all the data received so far, the mean was taken of a selection of the historical values. The selection was made in order to incorporate particular data characteristics of interest without having to explicitly model them and parallels some of Stroup's work which takes the mean of a selection of historical values [6, 10]. This selection of data will be explored later in Section 3.2.

The GLM method makes the same selection of historical values as chosen for the mean prediction method and then takes the prediction process a step forward. Using the same approach and assumptions as seen in the method suggested by Farrington, Andrews, Beale and Catchpole [5], the prediction is made using a generalised linear model and the quasi-Poisson link function. The selection of historical data (called baseline values) is used as the response variable. Although Farrington et al. proposed the use of one predictor, i.e. the week corresponding to the baseline value, initial investigations suggested that this predictor alone did not always perform satisfactorily. Often when a year upon year increase was observed within the data, the prediction would consistently fall below the observed values and so fail to properly pick up on the pattern of change in overall level. In order to correct for this problem, two further predictors were explored.

First, the introduction of a linear year term in addition to the weekly term as a predictor. The second possible solution to this problem stems from incorporating a yearly term via five indicator variables representing the five baseline years. For the predictions, removing the indicator variable for the most recent of baseline years and setting the other indicator variables to be zero, results in predictions that are biased towards the most recent information but still impacted on by that of previous years. In order to attempt to reduce the effect of historical outbreaks, Farrington et al. re-weight the GLM so as to down-weight those baseline values with large residuals.

ZIP modelling is again a regression based approach but instead of making the assumption that the data comes from a quasi-Poisson distribution, the data is assumed to come from a distribution that is similar to a Poisson distribution but with more zeros than would be expected under normal

Poisson structures [11-13]. This sort of distribution theoretically fits in much more accurately with what is expected of the largest section of SCIEH's collected data, i.e. organisms that have a rate of less than one count per week.

### 3.2. CHOICE OF HISTORICAL DATA

The possible choices of historical data to be used as the basis of any predicted value fall into two different categories. The first choice is to make use of all the historical values observed, however this approach was only utilised by the EWMA prediction method. The alternative strategy that was used for the remaining prediction methods, was to reduce the amount of data that contributes to a prediction by taking only a selection of the historical data, but in that selection aiming to improve the prediction ability by isolating data of specific interest. This approach has been used in a variety of published work, including Farrington et al. [5] and Stroup et al. [6, 10].

The choice made by Farrington et al. [5] of a seven week window of data centred on the current week's position within the year and applied over the past five years has both parsimony and simplicity about the manner in which it addresses its aims. The act of choosing a window of data allows for the incorporation of a seasonal effect through the data choice rather than explicit modelling. The length of window has been chosen in such a way as to enable flexibility over the exact positioning of the seasonal pattern, i.e. if a seasonal peak were to move slightly from year to year the window is sufficiently wide in order to still capture a seasonal pattern that remains relevant. Taking the last five years together allows, in part, for fluctuations across the years and in particular means that an outbreak in the last year will not have as detrimental effect as would be the case if only the previous year's data were used. The choice of using historical data from only the last five years crucially also makes a clear decision regarding the expiry time of the data, i.e. data that are more than five years old are likely to be less relevant to making predictions for the current time period due to such factors as increased/decreased organism occurrences, changes in the reporting system, and changes in the classification of organisms. Practitioners have accepted this to be pragmatic in the construction of an exception reporting system. This choice of 35 historical, or baseline, values was therefore used in conjunction with the mean, GLM and ZIP prediction methods.

## 4. RESULTS AND GENERAL FINDINGS

Initially using the simulated data, all of the prediction methods outlined above were tested on each combination of data characteristics mentioned in Section 2. For both the GLM and ZIP prediction methods, each linear combination of predictors was individually tested. Each method was evaluated on its ability to find and match the patterns within the data presented. Due to the number of combinations of generated data patterns and prediction methods, it was necessary to perform this evaluation numerically as well as visually, utilising correlograms of the resultant differences between the observed and predicted values. Just as before, these correlograms could be used to pick up on any remaining trend or seasonal pattern and the effect of outbreaks lying in the historical data could be visually examined in the differences. Once a decision had been made on the basis of the simulated data, this decision was ratified using the real rather than simulated data.

It was found that ultimately the GLM approach to prediction excelled among the prediction methods. While the ZIP method often worked equally well, its much greater computational demands meant that it could not be usefully scaled up for application on large numbers of organisms. Equally, although the EWMA approach often worked very well when examined solely on a numerical basis, visual examination of the observed and predicted values highlighted the fact that along with accurately matching trend or seasonality in the data, it would also 'match' any outbreaks seen and thus make the outbreaks nearly impossible to detect via the differences between the observed and expected values.

Although it was initially hoped that one prediction method would be found that outperformed the others in all circumstances, it instead transpired that two different GLM prediction methods were necessary. For those organisms with a reporting rate less than or equal to five, the systematic component of the log-linear model was

$$\begin{gathered} \mathit{log(count)} = \alpha + \beta\mathit{*week} + \gamma_1\mathit{*baseline\ year\ 1} + \gamma_2\mathit{*baseline\ year\ 2} \\ + \gamma_3\mathit{*baseline\ year\ 3} + \gamma_4\mathit{*baseline\ year\ 4} \end{gathered}$$

where $\alpha$, $\beta$, $\gamma_1$, $\gamma_2$, $\gamma_3$ and $\gamma_4$ are all constants, week takes an integer value from the interval [-3,3] to represent from which position within the seven-week window cast over the historical data the baseline count comes and, for example, baseline year 1 is an indicator variable whose value is one if the baseline count comes from the earliest of the five baseline years and is of zero otherwise. Complementing this, for those organisms with a reporting rate greater than five, the systematic component of the log-linear model was

$$\mathit{log(count)} = \alpha + \beta\mathit{*week} + \gamma\mathit{*\ year}$$

where $\gamma$ is a constant and year is the year from which the baseline count comes.

As was stated earlier, this choice of prediction scheme was ratified using the 105 'real' data sets. Two examples are shown for illustration of the prediction scheme in practice: Salmonella enteritidis and Haemophilus influnzae, the first having a reporting rate just below 30 and the second with a reporting rate of 2.2. Figures 1 and 2 show the plots for each of these organisms in turn. The first plot in each figure is of the observed reports collated by SCIEH between 1995 and 2001 with the predicted values for each week plotted with a dashed line. The second plot in Figure 1 shows the differences between these observed and expected values using the log-linear prediction model, and the last plot in each figure is the correlogram of these differences. While the organism data presented are very different, in each case the prediction model is seen to pick up the overall patterns within the data and produce predicted values that, when subtracted from the observed counts, leave behind the main data points likely to be of interest to an exception reporting system. The correlograms of these differences then highlight the removal of both trend and seasonality from within the data.

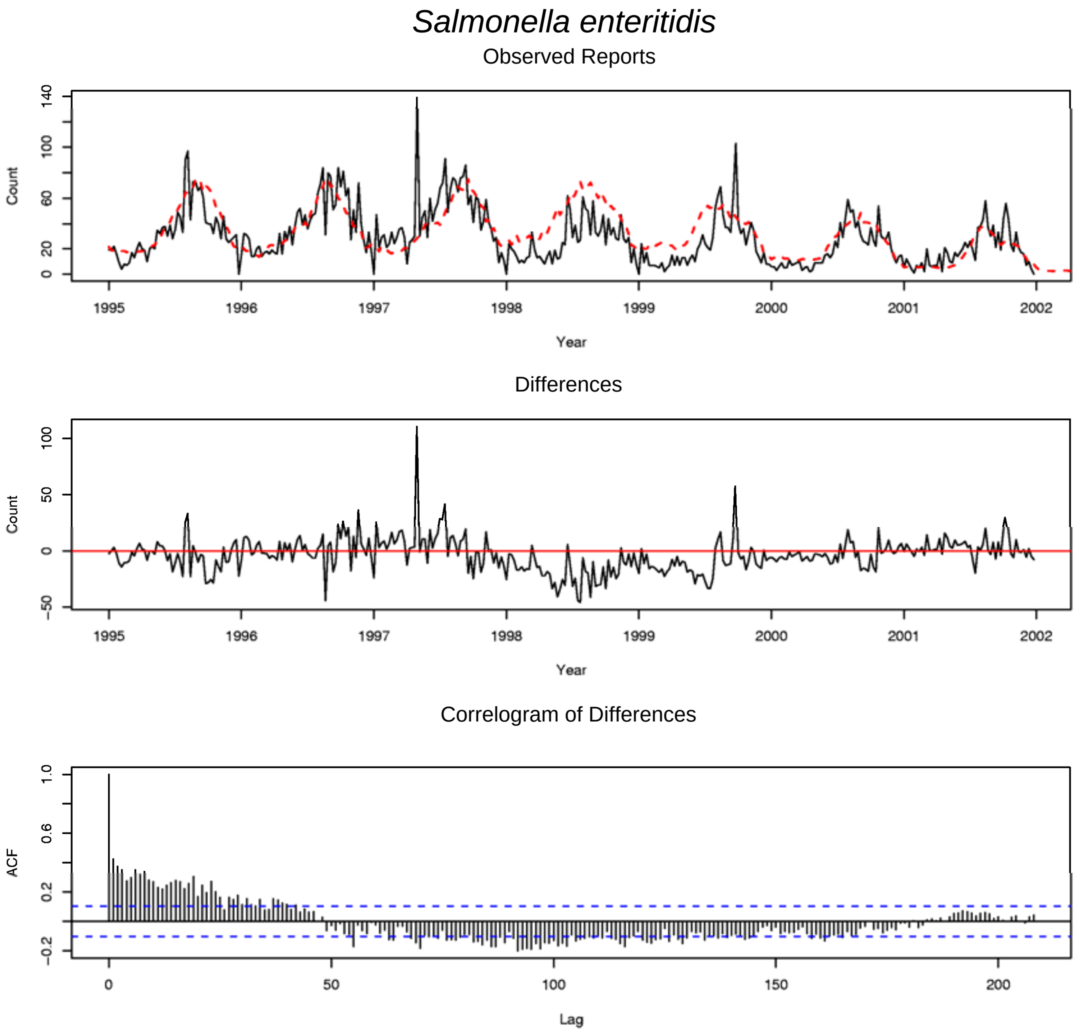
Across the range of 105 organisms tested, it was found that the selected prediction methods performed well in light of the data presented to them and for only 1 out of 105 organisms was there evidence of the prediction model being clearly inadequate, namely Hepatitis C.

## 5. CONCLUSION

As the first component of an exception reporting system, a prediction method was sought that addressed five key competencies:

1) the ability to deal with the necessary range of organism reporting rates;
2) the ability to accommodate a trend within the data;
3) the ability to accommodate a seasonal pattern within the data;
4) the ability to handle outbreaks occurring in the historical data without particular detriment;
5) the ability to act without excessive computing cost and complexity.

Using both simulated and real data from the SCIEH, it transpired that not one, but two prediction methods were required due to their differing behaviour under varying underlying reporting rates. These two chosen prediction methods will now be used to produce expected values for further work developing an exception reporting system, i.e. the comparison between the observed and expected values that will form the basis of alarm decisions.



*Salmonella enteritidis*

Observed Reports

Differences

Correlogram of Differences

Fig.1. Salmonella enteritidis plots

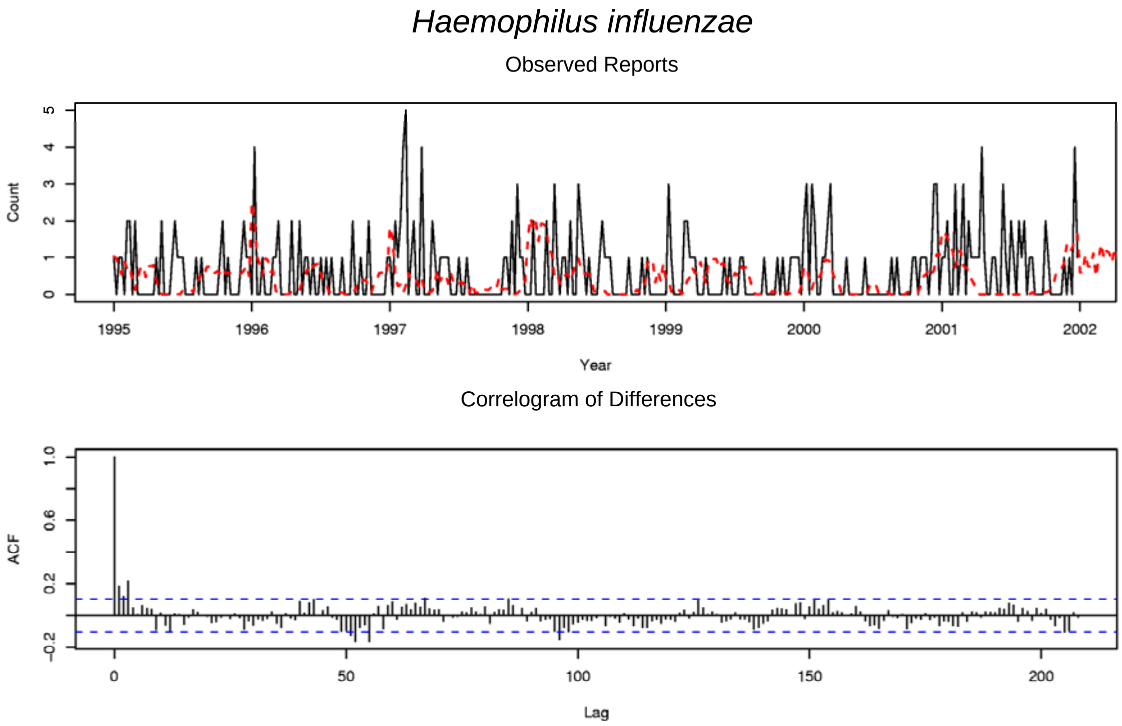


Fig.2. Haemophilus influenzae plots

## BIBLIOGRAPHY

[1] LAST J.M., ed. A dictionary of epidemiology, 3rd ed, Oxford University Press, 1995.

[2] COSTAGLIOLA D., When is the epidemic warning cut-off point exceeded?, European Journal of Epidemiology, Vol. 10, No. 4, pp.475-476, 1994.

[3] WATIER L., RICHARDSON S., HUBERT B., A time-series construction of an alert threshold with application to S-bovismorbificans in France, Statistics in Medicine, Vol. 10, No. 10, pp.1493-1509, 1991.

[4] TOUBIANA L., FLAHAULT A., A space-time criterion for early detection of epidemics of influenza-like-illness, European Journal of Epidemiology, Vol. 14, No. 5, pp.465-470, 1998.

[5] FARRINGTON C.P., ANDREWS N.J., BEALE A.D., CATCHPOLE M.A., A statistical algorithm for the early detection of outbreaks of infectious disease, Journal of the Royal Statistical Society Series A - Statistics in Society, Vol. 159, No. 3, pp.547-563, 1996.

[6] STROUP D.F., WHARTON M., KAFADAR K., DEAN A.G., Evaluation of a method for detecting aberrations in public-health surveillance data, American Journal of Epidemiology, Vol. 137, No. 3, pp.373-380, 1993.

[7] STERN L., LIGHTFOOT D., Automated outbreak detection: A quantitative retrospective analysis, Epidemiology and Infection, Vol. 122, No. 1, pp.103-110, 1999.

[8] NGO L., TAGER I.B., HADLEY D., Application of exponential smoothing for nosocomial infection surveillance, American Journal of Epidemiology, Vol. 143, No. 6, pp.637-647, 1996.

[9] CHATFIELD C., The analysis of time series - An introduction, 5th ed, Chapman & Hall/CRC, 1996.

[10] STROUP D.F., WILLIAMSON G.D., HERNDON J.L., Detection of aberrations in the occurrence of notifiable diseases surveillance data, Statistics in Medicine, Vol. 8, No. 3, pp.323-329, 1989.

[11] XIE M., HE B., GOH T.N., Zero-inflated Poisson model in statistical process control, Computational Statistics & Data Analysis, Vol. 38, No. 2, pp.191-201, 2001.

[12] LAMBERT D., Zero-inflated Poisson regression, with an application to defects in manufacturing, Technometrics, Vol. 34, No. 1, pp.1-14, 1992.

[13] RIDOUT M., DEMETRIO C.G.B., HINDE J., Models for count data with many zeros, Proc. 19th International Biometric Conference, pp.179-192, Cape Town, 1998.