***THE INTERNATIONAL JOURNAL OF BUSINESS & MANAGEMENT***

# Managing and Predicting the Number of Health Insurance Claims in Ghana Based on Big Data and Time Series Analysis: A Case Study of Kumasi Metropolis, Ghana

**Wu Jiying**
Associate Professor, School of Finance and Economics, Jiangsu University, China
**Jean-Jacques Dominique Beraud**
Master Student, School of Finance and economics, Jiangsu University, China
**Isaac Adjei Mensah**
Ph.D. Student, Department of institute of Applied Systems and Analysis,
Jiangsu University, China

***Abstract:***
*This study seeks to improve the process of managing and securing personal information (name, age, place of residence, telephone number, disease and so on) of the National Health Insurance Scheme (NHIS) subscribers of Ghana. Data was collected from the National Health Insurance Authority (NHIA) in Kumasi, Ghana on claims dating from 2010 to 2016. Big data methodology and Seasonal Auto Regressive Integrated Moving Average (SARIMA)*$(2,1,0)(2,0,0)^{12}$ *model was used to study and analyze the trend of health insurance claims and its future prediction claims.Our result showed an overall decrease in the number of claims from 2010 to 2016 and revealed that, the NHIA is faced with challenges in handling claims submitted to it by its clientele. An increasing trend was predicted from 2017 to 2022based on the model. Henceforth, the challenges of NHIA from the results of our study are attributable to inflated claims by clientele. NHIA efficiency can be improved by establishing a system that validates and controls claims.*

***Keywords:*** *Big data, security, information, NHIS, times series analysis*

## 1. Introduction

Insurance is a mechanism which allows interested individuals to make periodic fund contributions (premiums) into a central domain 'pool' which is used to compensate people who suffer the loss for which the contributions were made. For years, a globally encountered problem is the high cost of health care services, attributed to the increasing demand for these services and an apparent slow in the supply (Mc Connell, 1999). The National Health Insurance (NHI) also known in other countries as Statutory Health Insurance (SHI) is a scheme that is enshrined in the constitution of most countries with the goal to insure its citizens against the cost of health care. The administration of the scheme may be done by the private sector, the public sector, or a combination of the two depending on the source of funding. Every country has its own source of funding depending on the nature of the scheme that is being run. In countries like Australia, which uses the 'Medicare System' and the United Kingdom (UK), which also uses the National Health Insurance, have their major source of funding from general taxation. The case is not different from Canada. However, countries like Belgium and Germany have their source of revenue generated by employers and employees into a contribution fund called the "Sickness Fund". As such, the funds do not come directly from the government or private payments (Wikipedia). In 1985, patients were charged with for health care services in Ghana a compulsory fee. This was initiated by the Act of Hospital Fees regulations (LI 1313). The introduction of these user fees at government hospitals and clinics as part of the cost share policy by the government of Ghana in the health sector, placed a lot of financial burden especially on the less endowed. The system was such that, patients were required to make on the spot payment for health care services delivered to them. This was referred to as the "Cash and Carry System". This strategy made accessibility to health care services more difficult since most of the citizens were not able to meet the requirements. Because of the above-mentioned problem, the "Cash and Carry System" was phased out and the National Health Insurance Scheme (NHIS) was introduced into the country in 2003 by the National Health Insurance Acts 650, with the aim of improving the accessibility to health care and decreasing the cost borne in accessing health care services. It was an idea that was initially conceived by the then government in 2003, with the goal to provide equitable access and financial coverage from basic health care services to the citizens. Nowadays with the increased number of subscribers, gathering all this information has become a problem since the number of subscribers is increasing exponentially. The National Health Authority (NHA) decided to upgrade and computerize their system by creating a network system to save all the subscribers personal information and facilitate the payment of claims. The usage of big data management and security is a necessity in terms of gathering and securing the information of their subscribers since that information are vital for the on-going services that they are offering to their customers. The major source of funding to health care services in Ghana is the Government, which is generated from tax revenue. According to Kunfaa (1996) in most developing countries, the Central government is the main financier of health services. He as well indicated that on the average, 2.4% of the GDP is spent on health care needs by the government. Since the advent of the NHIS Act 650, 2003, Ghana was solely on the "Cash and Carry "System. The existence of this law has led to the establishment of 145 Mutual Health Insurance Schemes at the Districts, Municipals and Metropolitan areas in the country. It also established the National Health Insurance Authority to seek to the implementation of the National Health Insurance Policy, which made accessibility to basic health care services very easy to the Ghanaian citizens. It was realized that, some of the claim's providers inflate the number of claims just to receive payments. One major worry is the fact that, after auditing the information provided on hospital folders does not correlate with the number of claims forms from the service providers. Ashanti region was a major victimized region for claims inflation from service providers. In 2016, it was recorded that 19% of total people had the claims but the claim form indicated 30%, which laid to a delay in the payment of the claims. The National Health Authority decided to pause the registration of new subscribers and investigate the inflations of claims. Because of these mentioned problems, the research was carried out to identify and manage the actual number of people who registered for the scheme by recommending a process of handling big data. Which will help the agency to have a unique database where no one-can manipulate. We as well analyze the pattern or trend of the claims in the preceding years and predict the future possible claims.

## 2. Literature Review

In recent years, modeling and predicting health insurance claims has been one of the major interesting research subjects because modeling the number of insurance claims is an essential part of insurance pricing. A considerable amount of literature has been published on modeling number of claims using count regression analysis. These studies allow identification of risk factors and prediction of the expected frequency. Many researchers present a new model for panel data, where the interpretation of the model, the probability distribution, the properties of the model, and its first moments are shown. In some models, it is shown why some intuitive models (past experience as regressors, multivariate distributions, or copula models) involving time dependence cannot be used to model the number of reported claims. Statistical tests to compare the nested models are explained and a Vuong test is used to compare the fitting of non- nested models. However, most of them did not tackle the management and security of the subscriber's information that is a very delicate point because if the information is not well manage and secure it will result in the inflation of claims and leakage of personal information that constitute a danger to the subscribers of the National Health Insurance. Big data is used to refer to the ever-increasing amount of information that organizations are storing, processing and analyzing due to the enlarged number of information sources in use. Within the next decade, the amount of information managed by enterprise data centers will grow considerably including the number of IT professionals who will be needed to work with such data. Data volumes continue to expand as they take in an ever-wider range of sources, much of which is in unstructured form. Most organizations want to extract value from that data to identify and study the opportunities for the business that it contains. However, the centralized nature of big data stores creates new security challenges because they contain personal information that can be used against an individual or state for example. Traditional tools are not, on their own, up to the task of processing the information, the data it contains, let alone ensuring it is secured in the process.

The ever-increasing integration of highly diverse enabled data generating technologies in medical, biomedical and healthcare fields. The growing availability of data at the central location that can be used in need of any organization from pharmaceutical manufacturers to health insurance companies to hospitals have primarily make healthcare organizations and all its sub-sectors in face of a flood of big data as never before experienced. While this data is being hailed as the key to improving health outcomes, gain valuable insights and lowering costs, the security and privacy issues are so overwhelming that healthcare industry is unable to take full advantage of it with its current resources. Managing and harnessing the analytical power of big data, however, is vital to the success of all healthcare organizations. Karim' study presents the state-of-the-art security and privacy issues in big data as applied to healthcare industry and discuss some available data privacy, data security, users' accessing mechanisms and strategies (Karim Abouelmehdi, 2017). It is therefore important to prepare for subsequent data curation and integration at the point of data capture. To date, health care industry has not fully grasped the potential benefits to be gained from big data analytics. While the constantly growing body of academic research on big data analytics is mostly technology oriented, a better understanding of the strategic implications of big data is urgently needed. To address this lack, (Yichuan Wang, 2016) in their study, examined the historical development, architectural design and component functionalities of big data analytics.

From content analysis of 26 big data implementation cases in healthcare, they were able to identify five big data analytics capabilities: analytical capability for patterns of care, unstructured data analytical capability, decision support capability, predictive capability, and traceability. They also mapped the benefits driven by big data analytics in terms of information technology (IT) infrastructure, operational, organizational, managerial and strategic areas. In addition, they recommended five strategies for healthcare organizations that are considering adopting big data analytics technologies. Their findings will help healthcare organizations understand the big data analytics capabilities and potential benefits and support them seeking to formulate more effective data-driven analytics.) Jing et al stated in their paper that Electronic medical records (EMR) and health insurance claims data offer two potential data sources for researchers to examine healthcare utilization patterns and the cost of care. In particular, combining the clinical and epidemiological variables typically available in EMR with cost information available in the claims data is not only intuitively sensible, but also increasingly more feasible with growing standardization of EMR across healthcare delivery systems. In this study, they compare EMR and claims data within a cohort of rheumatoid arthritis patients who received care from Geisinger Health

System (GHS) and had concurrent Geisinger Health Plan (GHP) coverage. They also developed a cost “imputation” method to obtain GHP claims-based cost estimates within EMR, even for those who did not have GHP coverage. Their findings confirm that, there is significant disagreement between EMR and claims data. They suggested that each represent a different set of clinical phenomena.

Their study also illustrates different factors to consider for researchers in choosing one data source over the other in conducting clinical research, (Jing Hao et al, 2017). Tulone and Madden, 2016came out with a method for estimating the values of sensors in a wireless sensor network using time series forecasting. Specifically, their approach was based on Autoregressive models built at each sensor to predict local readings. Nodes transmit these local models to a sink node, which uses them to predict sensor values without directly communicating with sensors. When needed, nodes send information about outlier readings and model updates to the sink. They showed that this method can dramatically reduce the amount of communication required to monitor the readings of all sensors in a network, and proved that their method provides correct, user-controllable error bounds on the predicted values of each sensor. (Hibon and Makridakis, 2015) in their paper tried to study the Box-Jenkins methodology to ARIMA models and determine the reasons why in empirical tests, it is found that the post-sample forecasting accuracy of such models is worse than much simpler time series methods. They concluded that the major problem is the way of making the series stationary in its mean (i.e., the method of differencing) that has been proposed by Box and Jenkins. If alternative approaches are utilized to remove and extrapolate the trend in the data, Auto Regressive Moving Average (ARMA) models outperform the corresponding methods involved in the great majority of cases. In addition, it is shown that using ARMA models to seasonally adjusted data slightly improves post-sample accuracies while simplifying the use of ARMA models. It is also confirmed that transformations slightly improve post-sample forecasting accuracy, particularly for long forecasting horizons. Finally, it is demonstrated that Auto Regressive (AR) (1) and AR (2), or their combination, produce as accurate post sample results as those found through the application of the Box-Jenkins methodology.

## 3. **Methodology**

### *3.1. Big Data Management*

Digital technologies have become indispensable in the healthcare sector. They helped in the achievement of better and cheaper healthcare. Policymakers are nowadays focusing strongly on eHealth (electronic Health) and collaboration between the various players in the healthcare ecosystem. It is crucial for the NHIA not to miss the boat. In for-profit organizations, digitalization revolves around customer centricity. They focus on how they can ensure that patients are given control, access to all their information and can make and reschedule appointments themselves, whilst enabling them to be monitored at every stage in the journey. So, besides information about their passage through a particular hospital, they need to know about all their hospital admissions and/or tests, wherever they may be. The government is encouraging hospitals to form networks, so the exchange of information between hospitals in a network and between the various networks is crucial. It means the exchange of information can help for further follow-up of the patient once they have been discharged from the hospital. That involves all kinds of external parties: GPs, physiotherapists, home care teams, residential care centers etc. Information can only be exchanged efficiently if it is standardized and the relevant processes and systems are integrated. Brecht planned to present a uniform IT roadmap, which could then help to put the further digitalization of the hospital sector on the right track. In order to get a better picture of the IT challenges, he first consulted various external parties, which work with and for hospitals or specialize in IT for the healthcare sector. Claude Hopkins believes that, accurate information points to the shortest, safest and cheapest route to any destination. Big data helps us to choose more effective modes of communication and more valuable business models. Studies from various sectors show that the transition point where scalability begins to bind is likely to arise in one of four general directions, generally referred to as the four Vs (Volume, Velocity, Variety and Veracity) of big data.

As a result, many organizations that collect process and analyze big data turn to NoSQL databases as well as Hadoop and its companion tools, including:

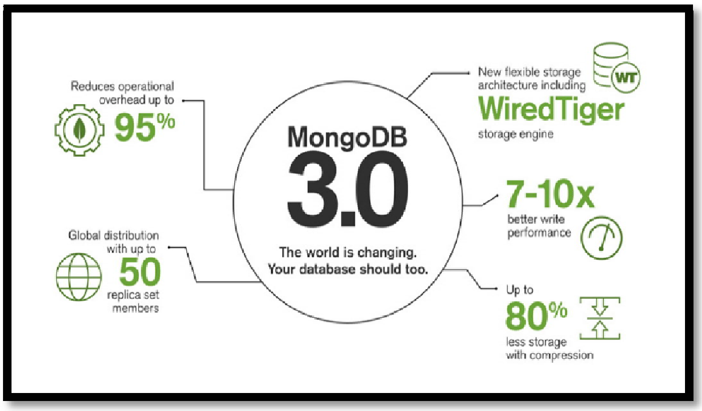

Reduces operational overhead up to
95%
Global distribution with up to
50
replica set members
MongoDB
3.0
The world is changing. Your database should too.
New flexible storage architecture including
WiredTiger
storage engine
WT
7-10x
better write performance
Up to
80%
less storage with compression

*Figure 1: MongoDB processing*

MongoDB is the modern, start-up approach to databases. It is good for managing data that changes frequently or data that is unstructured or semi-structured.

Yarn: a cluster management technology and one of the key features in second-generation Hadoop.

MapReduce: a software framework that allows developers to write programs that process massive amounts of unstructured data in parallel across a distributed cluster of processors or stand-alone computers.

Spark, an open-source parallel processing framework that enables users to run large-scale data analytics applications across clustered systems.

Kafka: a distributed publish-subscribe messaging system designed to replace traditional message brokers.

| Processor | Application in Healthcare System |
|---|---|
| MongoDB | Healthcare companies rely on MongoDB to address a broad variety of use cases while at the same time meeting compliance standards and improving healthcare outcomes. Some examples of healthcare solutions built on MongoDB include:<br>360-Degree Patient View: Patient data is mostly a challenge, as it is often too great to manage and secure to be effectively retrieved and analyzed. It helps healthcare providers to create applications that easily consolidate data from any source to more efficiently serve patients, reduce potential errors in patient care, and improve patient outcomes.<br>Population Management for At-Risk Demographics: The power of sophisticated data analysis is particularly compelling when it can be used to help save lives. One example is when insurers can reduce preventable patient illnesses by using advanced population management applications. A healthier patient population is also good business as it helps insurers reduce costs and increase margins.<br>Lab Data Management and Analytics. MongoDB helps healthcare providers make better use of lab data by enabling real-time analytics and data visualization. The result is new insights to better serve patients and new revenue streams for providers. |
| Hadoop | Real-time data analysis and distribution is an ongoing activity for healthcare organizations. Patient records are one of the many crucial information resources like, claims, finance records, customer relationship management (CRM) systems, business and care partner organization references, research logs and physician correspondence. Much of this data is unstructured and changes constantly. In addition, it is usually spread across multiple sources and departments. Getting access to this valuable data and factoring it into clinical and advanced analytics is critical to improving care and outcomes, incentivizing patient behavior and driving efficiencies. Hadoop software help solve the above mention problems. It also helpsin:<br>• Building sustainable healthcare systems and health information exchanges<br>• Improving clinical treatment effectiveness and reducing readmission rates<br>• Reducing medical errors and supporting collaboration<br>• Detecting claims fraud and other attempts to misuse medical resources. |
| Hafka | Its goal is to improve health care through the meaningful use of health information technology in order to:<br>• Improve healthcare quality and coordination so that outcomes are consistent with current professional knowledge<br>• Reduce healthcare costs, reduce avoidable overuse<br>• Provide support for reformed payment structures<br>Claims are the documents providers submit to insurance companies to get paid. A key component of the Health Insurance Portability and Accountability Act (HIPAA) is the establishment of national standards for electronic healthcare transactions in order to improve efficiency by encouraging the widespread use of Electronic Document Interchange (EDI) between healthcare providers and insurance companies. Claim transactions include International Classification of Diseases (ICD) diagnostic codes, medications, dates, provider IDs, the cost. |
| Virtual Security | In the healthcare industry, virtual machine (VM) technology can provide benefits to energy and hardware costs, efficiency, security, and maintenance. As users express a desire to be more mobile in the workplace, the opportunity to implement virtualization is significant to every employee within an organization. What’s more, patients are coming to expect more personalized care that comes from the quick retrieval of accurate records made possible by end-user usability. VMs are the key to a truly modern healthcare organization, and understanding how they operate is critical to the success of deploying any virtualization solution. |

*Table 1: Usage of Big Data in Healthcare*

### *3.2. Time Series Model*

A time series model for the observation (Yt) is the joint distribution of the sequence of random variable Yt. The basic time series models include:

- Trend Model: This is when the mean of the model is a function of time. It is given by $Y_t = M_t + Z_t$, where $Z_t \sim$IID $(0,\sigma^2)$ and Mt is the trend function.
- Trend and Seasonality: A time series model may consist of both trend and seasonal component i.e. Yt = Mt + Zt + Xt with Zt ~IID (0, σ2) where Mt is the trend

component and Xt is the seasonal component.

Any time series data should exhibit at least one of the following components.

- Trend component.
- Seasonal or periodic variations
- Cyclical variation
- Irregular variation

To forecast any time series data, it is very crucial to assess stationarity of the data. Stationarity is done to remove some short-term fluctuations or remove seasonal fluctuations. A time series data is stationary if the mean and the variance of the series are invariant or independent of time. A time series data is strictly stationary if the joint distribution implies the following basic conditions.

That is for any value of t, $f(X_1, X_2 \cdot\cdot\cdot, X_t) = f(X_{1+k}, X_{2+k}, \cdot\cdot\cdot, X_{t+k})$ where k is an integer. This means that the joint distribution remains fixed if all time periods are moved a constant number of periods. The following is a time series plot of the annual number of earthquakes in the world with seismic magnitude over 7.0, for 99 consecutive years.

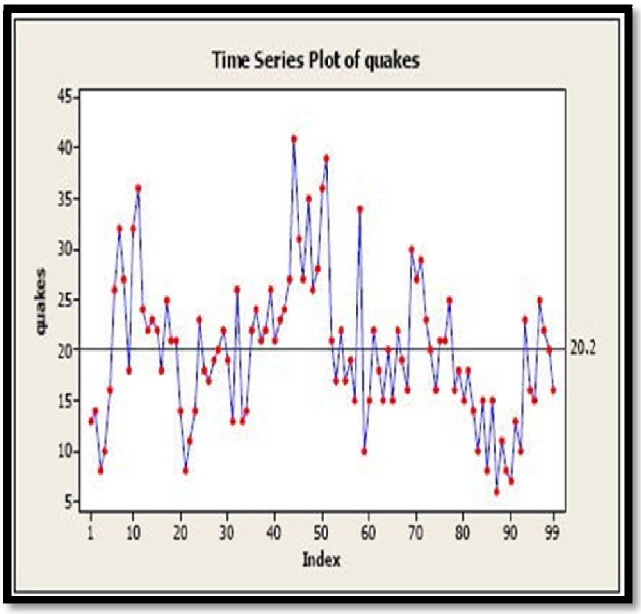


*Figure 2: Time Series Plot of Earthquakes*

Some features of the plot above:

- There is no consistent trend (upward or downward) over the entire time span. The series appears to slowly wander up and down. The horizontal line drawn at quakes = 20.2 indicates the mean of the series. Notice that the series tends to stay on the same side of the mean (above or below) for a while and then wanders to the other side.
- Almost by definition, there is no seasonality as the data are annual data.
- There are no obvious outliers.
- It is difficult to judge whether the variance is constant or not

Forecasting is the art and science of predicting future events based on historical data. The categories of forecasting are grouped into two, which is Qualitative and Quantitative methods. Time series methods of forecasting use historical data as the basis of estimating future outcomes. Some of these forecasting time series techniques (Extrapolation methods) include:

- Trend-Based Regression
- Auto-Regression
- Moving Averages
- Exponential Smoothing

For the purpose of this study, we will elaborate on Exponential Smoothing for the prediction of the claims. This forecasting technique bases forecasts on weighted average of past observations, with more weight on recent observations. The Exponential Smoothing is comprised of three main methods, which are:

Simple/ One-Parameter Exponential Smoothing: this is used when there is no obvious trend in the data. The simple exponential smoothing is written as:

$St = \alpha X_1 + (1-\alpha)S_{t-1}$, $t > 0$ where $\alpha$ is the smoothing factor, and $0 < \alpha < 1$.

In other words, the smoothed statistics St, is a simple weighted average of the current observation and the previous smoothed statistics $S_{t-1}$

Double or Two-Parameter Exponential Smoothing: this is used when there is a trend but no seasonality in the time series data. For any period, t, the smoothed value i.e. $S_t$is given us:
$S_t = \alpha X_t + (1 - \alpha)\ (S_{t-1}+U_{t-1})$. . . Smoothed observation
$U_t = \beta(S_t - S_{t-1}) + (1 - \beta)U_{t-1}$. ..... Trend equation
Moreover, the forecast value is $X_{t+1} = U_t + S_t$ with the assumption that $S_2 = X_1$and $U_2 = X_2 - X_1$ where α is the smoothing data factor and β is the trend smoothing factor with, $0 < \alpha < 1$ and $0 < \beta < 1$.

The Triple or Holts Winters Exponential Smoothing: this is also used when there is both trend and seasonality in the time series data, which was named after Holts and Winters. Due to the complexity by hand calculations, Statistical Software packages are used.

## 4. Data Analysis

### *4.1. Data Summary*

The data collected is a secondary monthly data recorded from January 2010 to December 2016 in Kumasi.

| Min | 1st QU. | Median | Mean | 3rd QU. | Max |
|---|---|---|---|---|---|
| 670.0 | 998.5 | 1165.0 | 1154.0 | 1310.0 | 1590.0 |

*Table 2: Summary Statistics of the Data*

The minimum number of claims in our data is 670 and the maximum value is 1590 claims. The overall mean of our data is 1154 claims and the overall median of our data is 1165 claims. In compiling our data, we realized that the NHIA's claims-vetting system is not well equipped to identify abnormal behavior among service providers. Claims offer a wealth of information on expenditure patterns, but most of the data captured by NHIA are not analyzed. We also realized that the data collected is not in a format of conductive analysis. Addressing these issues is a costly and time-consuming process. These issues remain common in all regions. Second, the data captured by the current system are insufficient to verify the accuracy of the appropriateness of the treatment. Moreover, the claims submitted to the NHIA are crosschecked manually which can lead to so many mistakes. One can mistakenly omit a claim or over count, the number of claims submitted. There is a serious need of putting in place a system that can replace the manual system adopted by the NHIA. We proposed the use of one software listed about for their data management. That system can be linked directly to the NHIA so that they can control and monitor each claim by various hospitals.

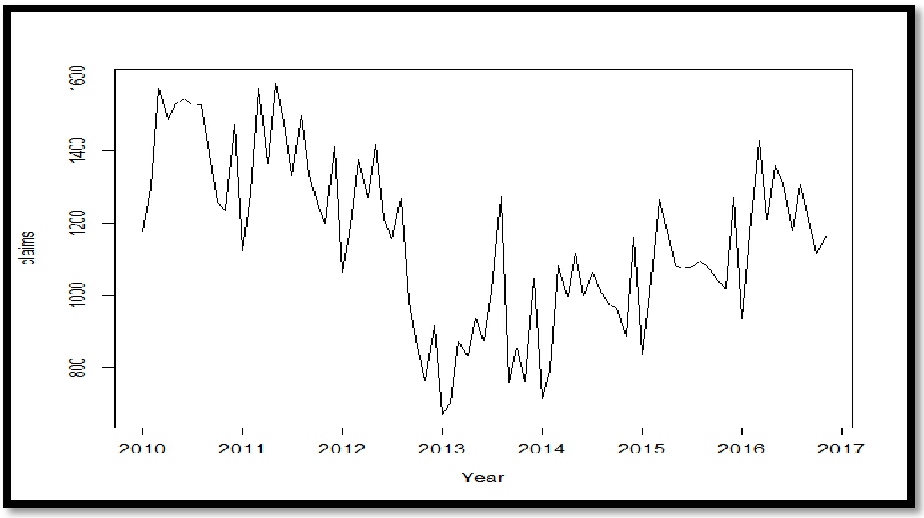


*Figure 3: Time Series Plot of the Number of Claims*

The time series plot in figure 3 above shows how the entire data set is distributed over the time interval space. The overall trend depicted by the plot shows a negative trend. However, a critical look at the plot shows that there was a mild decrease in the number of claims from 2010 to 2011 (negative trend). The number of claims then decreases considerably from the year ending of 2010 to 2012. The least number of claims recorded in February 2013 that was 670 and the highest number of claims (1590) recorded in June 2011. During the next month (March) of the same year 2013, the number of claims rose throughout the years persistently to 2016.
This is clearly shown in figure 4 below:

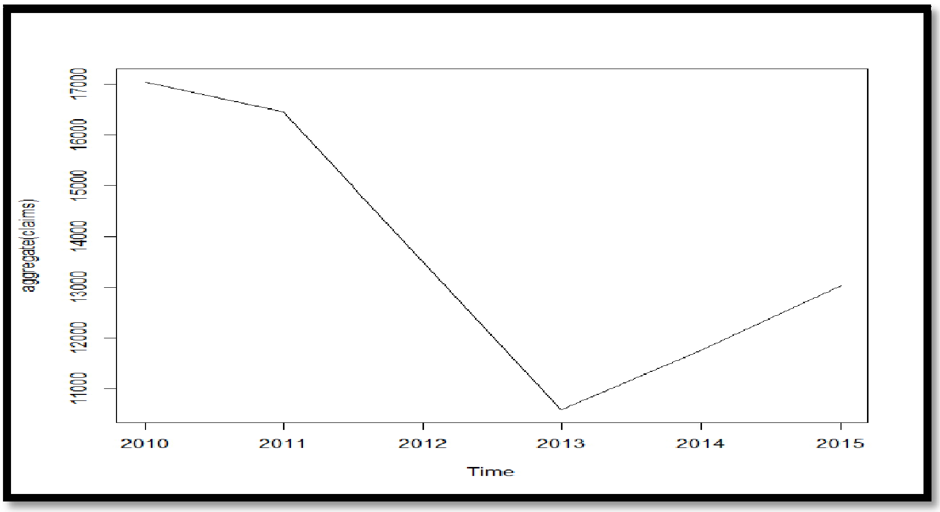


*Figure 4: Claims Aggregate over the Various Years*

The sharp decrease of the number of claims from 2011 to 2013 was because most people were not relying of the NHIS whenever they were sick. A survey conducted revealed that the health facilities under the NHIA were not accessible to most patients due to the long distance they have to cover. Some of them complained about the lack of health personnel in the various facilities, which lead to the delay in attending to them. They have to wait several hours before someone will attend to them irrespective of their condition. Hence, they prefer going to private health facilities where they will be attended to in a short period. Some also complained about the fact that the NHIS does not cover the majority of their expensive whenever they visit the hospital. Most hospitals complain about the delay in the payment of the previous claims, which enable them to attend to NHIS subscribers. Inefficient Claims Processing Claims processing by NHIA is labor-intensive and inefficient. Claims are vetted on an individual basis. Most claims are evaluated manually, even the relatively small share that are electronically submitted. The rise in 2013 was because the NHIA re-strategize it system by expending the coverage of patient's expenditure in the various hospitals. They also increased the number of personnel in the various facilities, which reduced the waiting of patients. It really helped in the increment of the number of NHIS subscribers.

We now look at the sample ACF and the PACF plot of the data

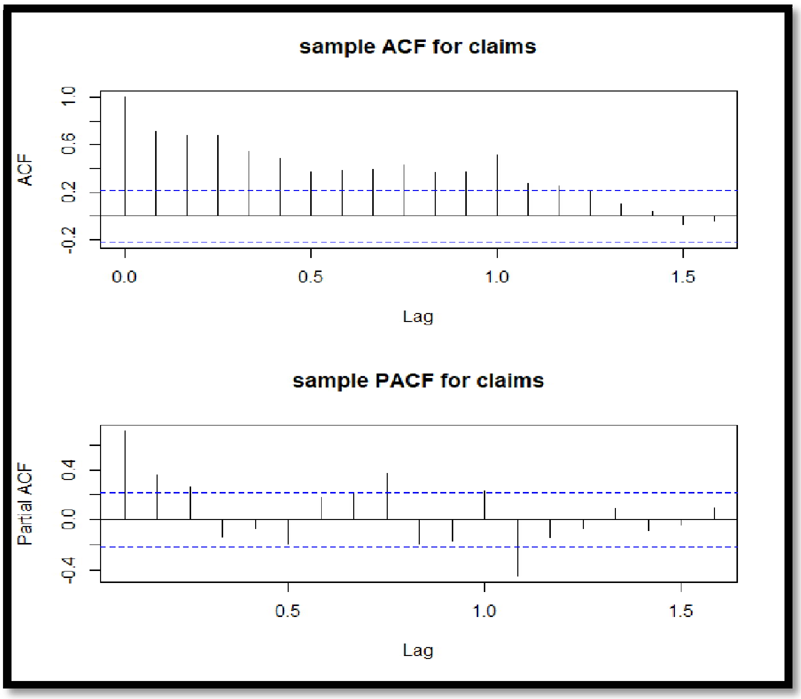


*Figure 5: ACF and PACF Plot of NHIS Claims*

From the plots in figure 5 above, we realize that is an ARIMA model because of the tapering effect of the ACF and the PACF, or simply put, because both the ACF and the PACF tails off. We can as well see from both the PACF and the ACF that there is trend in our data. We thus proceed to check if our data is stationary or not using the KPSS test for stationarity.

### *4.2. Test for Stationarity*

Null Hypothesis: The data is stationary Vs Alternate Hypothesis: The data is not stationary data: claimsKPSS Level = 1.0272, Truncation lag parameter = 2, p-value = 0.01.
Since our p-value is smaller than the significance level i.e. α = 0.05, we reject the null hypothesis and conclude that our data is not stationary. Hence, we need to difference our data to achieve stationarity.

### *4.3. Test for Stationarity*

Null hypothesis: The data is stationary Vs Alternate hypothesis: The data is not stationary KPSS Test for Level Stationarity Data: diff (claims, differences = 1)
KPSS Level = 0.057319, Truncation lag parameter = 2, p-value = 0.1
Since the P-value is greater than the significance level i.e. α = 0.05, we fail to reject the null hypothesis and conclude that our data is stationary. The first difference has made our data stationary.

### *4.4. Test for Normality*

We now consider the Shapiro-Wilk normality test to check if our data is normally distributed or not. Shapiro-Wilk normality test.

Null hypothesis: The data is taken from a normal distributed population. Alternate hypothesis: The data is not taken from a normally distributed population.
data: diff (claims, differences = 1)
data: claims W = 0.98057, p-value = 0.2423

Since the p-value is greater than the level of significant α = 0.05 we fail to reject the null hypothesis and conclude that our data is normally distributed. We use the Quantile-Quantile plot to further justify our point.

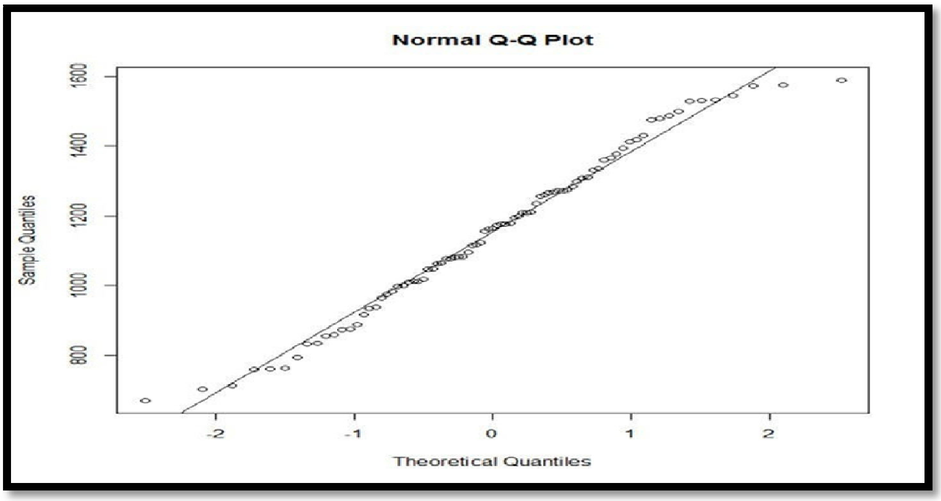


*Figure 6: Q-Q Plot of the Data*

From the Q-Q plot of the data, we realize that our data is normally distributed since most of the points lie on the 45$^{o}$ line.

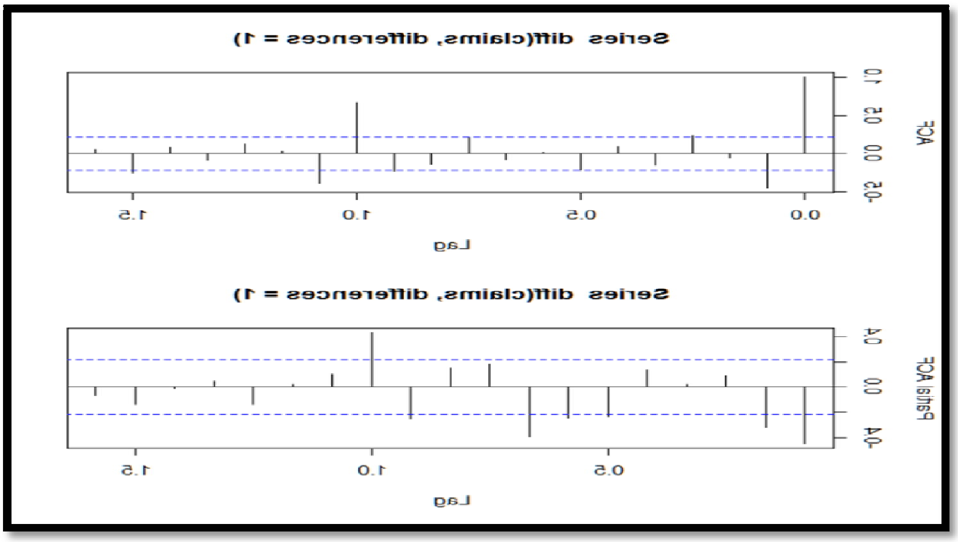
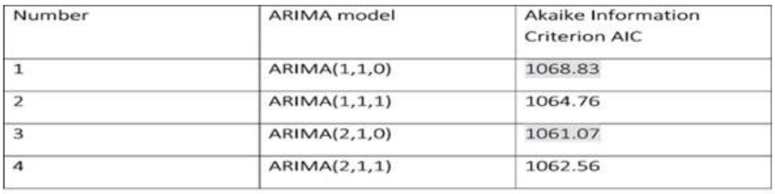


*Figure 7: ACF and PACF Plot of the First Difference of Claims*

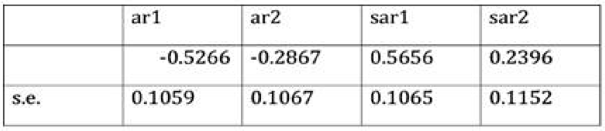
From the correlogram in figure 7 above, we can use both the ACF and PACF to identify the model that fit our data. We realize that the PACF cuts off after lag 2, which shows that the order of the non-seasonal AR part of the model is 2. Using the R software precisely the forecast package, we were able to get the Akaike Information Criterion (AIC) of our models. The table below contains this information.



| Number | ARIMA model | Akaike Information Criterion AIC |
|---|---|---|
| 1 | ARIMA(1,1,0) | 1068.83 |
| 2 | ARIMA(1,1,1) | 1064.76 |
| 3 | ARIMA(2,1,0) | 1061.07 |
| 4 | ARIMA(2,1,1) | 1062.56 |

*Table 3: Akaike Information Criterion*

The best model that fit our data is the one with the smallest AIC. In our case, the ARIMA (2, 1, 0) model is the one with the smallest AIC.Using the auto.arima function in the forecast package from R, the output below was obtained.
Series: claims ARIMA (2, 1, 0) (2, 0, 0) [12]
Coefficients: σ2 estimated as 9302: log likelihood = -496.73



| | ar1 | ar2 | sar1 | sar2 |
|---|---|---|---|---|
| | -0.5266 | -0.2867 | 0.5656 | 0.2396 |
| s.e. | 0.1059 | 0.1067 | 0.1065 | 0.1152 |

*Table 4: Model Parameters*

AIC = 1003.46 AICc = 1004.25 BIC = 1015.5

The proposed model for the data is ARIMA (2,1,0) (2,0,0) [12] which is the best model that fit our data and can be used in the prediction of future monthly claims.

*4.5. Model Diagnostic*

A first step in diagnostic checking of fitted models is to analyze the residuals from the fit for any signs of non–randomness, which is to say, the residuals must be uncorrelated to be a fit model. If the residuals are non-random then the model is not adequate or fit for forecasting. R has the function tsdiag ( ), which produces a diagnostic plot of a fitted time series model:

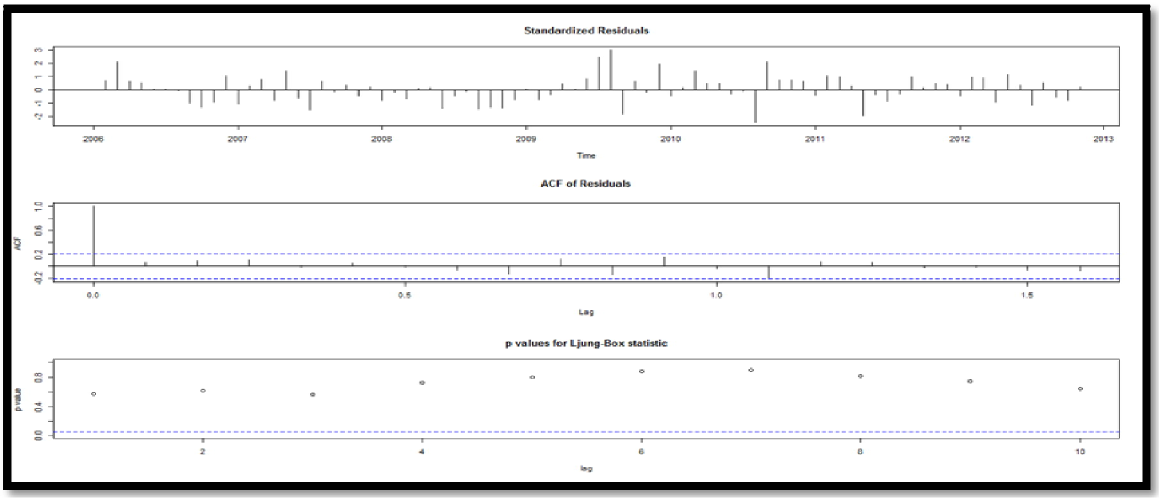


*Figure 8: Output from Tsdiag*

The autocorrelation of the residuals and the p-values of the Ljung–Box statistic for the first 10 lags. Since all the p-values of the residuals are greater than 0.05 we can say that, the residuals are uncorrelated or independent.

We perform a normality test of the residuals by using the Shapiro-Wilk test of residuals to check if the residuals are indeed normally distributed or not.

Null Hypothesis: The residuals are normally distributed Vs Alternate Hypothesis: The residuals are not normally distributed. Shapiro-Wilk normality testdata: resid(fit)W = 0.98856, p-value = 0.6785 since our p-value is greater than the significant level i.e. $\alpha = 0.05$, we fail to reject the null hypothesis and conclude that the residuals are normally distributed. We then use the Ljung–Box test to determine the adequacy of the model. Null hypothesis: The model is adequate Vs Alternate hypothesis: The model is not adequate data: fit-residualsX-squared = 21.501, df = 20, p-value = 0.3682

Since the p-value is greater than the significant value $\alpha = 0.05$, we fail to reject the null hypothesis and conclude that the model is adequate. We can move to the next step, which is the prediction of our data.

*4.6. Model Prediction*

From our analysis, the model achieved is SARIMA (2,1,0) (2,0,0) [12], we now predict the number of claims for the next five years.

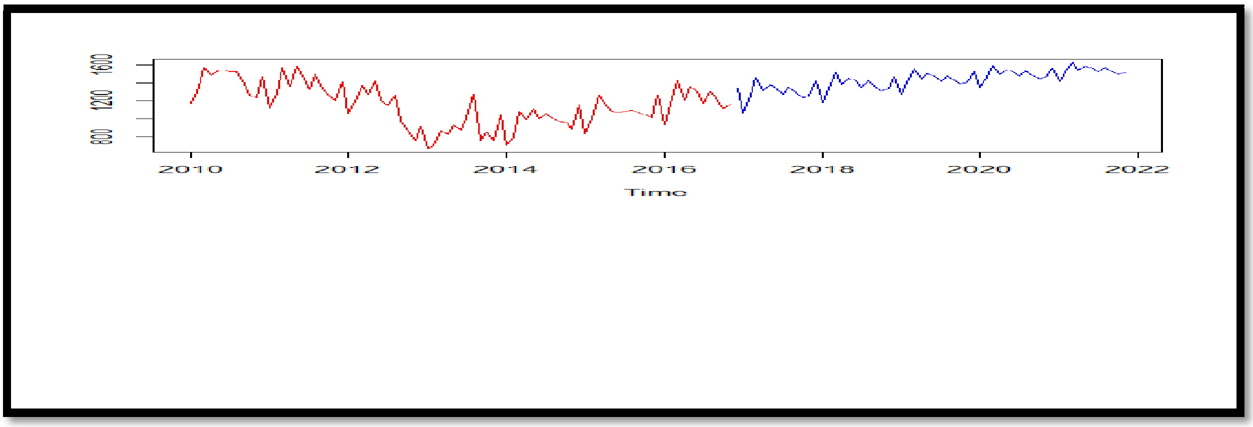


*Figure 9: Five Years Forecast of the Number of Claims*

The blue graph in figure 9 represent the future claims over the 5 years intervals. It reveals that from 2017 to 2022 the number of claims will increase compare to the previous years. It shows that the measure taken the NHIA to improve and facilitate the claim holders are working and people are willing to register more.

## 5. Conclusion

Digital technologies have become indispensable in the healthcare sector. Henceforth, the NHIA must adapt to this new technology since they help to achieve better and cheaper healthcare. We found that the NHIA has a poor system of handling their database, which led to the inflation of their claims. In addition, there was an overall decreasing trend from 2006 to 2013, though there appeared to be an increase in the number of claims from 2009 to 2013. Contrary to this, the forecast appears to have an overall increasing trend. This suggests a cyclical component in the patronage of NHIS.

The overall model that was obtained was SARIMA (2,1,0) (2,0,0) [12] and the corresponding equation is given by:$(1 - \Phi s1B12 - \Phi s2B24 + \Phi 1\Phi s1B24 - \Phi 1B + \Phi 1\Phi s2B36 - \Phi 2B2 + \Phi 2\Phi s1B36 + \Phi 2\Phi s2B48)\ (xt - xt{-}1) = \varepsilon t$,Where B is the backward shift operator and the parameter estimates are:

$\Phi 1 = -0.4880$, $\Phi 2 = -0.2471$, $\Phi s1 = 0.5323$, and $\Phi s2 = 0.2557$

The model was used to forecast the number of claims from 2013 to 2018. The forecasted number of claims had an increasing trend. Three factors determine the size and efficiency of claims expenditures in Ghana: coverage expansion, behaviors of service providers and National Health Insurance Scheme (NHIS) members, and the internal management of the National Health Insurance Authority (NHIA). Insurance intends to remove financial barrier for accessing care by expanding coverage. Policymakers' ability to influence this dimension is inherently limited. However, the authorities can affect the behavior of service providers and patients through measures to address adverse selection during enrollment, the suboptimal composition of the benefits package, low levels of cost-consciousness, and weak performance incentives. The NHIA can also enhance its own internal efficiency by reforming its systems for claims processing, provider oversight, and member engagement. Without a mechanism to collect member feedback, NHIS claims information and beneficiary characteristics cannot be verified. Reimbursements are based on provider claims that are vetted through a manual desk review. There is no second source of information with which to verify claims. Without member engagement, false or erroneous claims are likely to go undiscovered. The current system cannot identify patients who seek care frequently, even though they represent a large share of claims expenditures. About 20 percent of the subscribers in Kumasi visit health care facilities at least five times during a given year, yet they account for more than half of total outpatient expenditures. Some of them suffer from chronic diseases, while others may be visiting multiple facilities in order to find the best care option. However, in some cases these repeat patients may reflect fraudulent claims. Developing systems to identify and track these patients could help improve the efficiency of claims expenditures and reduce errors and abuse.

We also recommend that:

- Future claims submitted to the office should be weighed against the forecasted values to check for inflation of claim.
- Necessary measures should be put in place for the increasing trend in the forecast.
- A system must be put in place to prevent the inflation of claims, which harm the good function of the NHIS.
- The NHIA must patronage the installation of a system across all the hospital under its jurisdiction, which can be linked to their database. It will help them manage and control the flow of their subscribers whenever they visit any hospital.

## 6. References

i. Yichuan Wang, L. K., Terry Anthony Byrd (2016). "Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations." Elsevier 11: 3-13.

ii. Abdel-Aal, R. E. and Mangoud, A. A. (1998). Modelling and forecasting monthly patient volume at a primary health care clinic using univariate time-series analysis. Computer methods and programs in biomedicine, pages 235–247.

iii. Abubakari, F. (2012). Time Series Analysis on Membership Enrollment of National Health Insurance Scheme. KNUST-Kumasi.

iv. Antwi, S. (2012). A logistic Regression Model for Ghana National Health Insurance Claims. P.R. China: Jiangsu University.

v. Bowerman and O'Connell. (1993). Forecasting and time series: An applied approach. in the duxbury advanced series in statistics and decision sciences., the duxbury advanced series in statistics and decision sciences, (third ed.). Belmont: Duxbury.

vi. Box, G. E. and Jenkins, G. M. (1976). Time series analysis: Forecasting and control. San Francisco.

vii. Cochrane, D. and Orcutt, G. H. (1949). Application of least squares regression to relationships containing autocorrelated error terms. Journal of the American Statistical Association, pages 32–61.

viii. Dobre, I., a. A. A. (2008). Modelling unemployment rate using box-jenkins procedure. Journal of Applied Quantitative Method, 3(2).

ix. Durbin, J. and Watson, G. S. (1950,1951). Biometrika. testing for serial correlation in least squares regression. 37-38, 409-428,159-178.

x. Hibon, M. and Makridakis, S. ((accessed 2015)). Arma models and the box jenkins methodology. Fontainebleau, France: INSEAD.

xi. Kirchgassner, G., Wolters, J., and Hassler, U. (2012). Introduction to Modern Time Series Analysis. Berlin: Springer Press.

xii. Lu, C.-J., Lee, T.-S., and Chiu, C.-C. (2009). Decision support systems. Financial Time Series Forecasting Using Independent Component Analysis and Support Vector Regression, pages 115–125.

xiii. Mesike, G. C., Adeleke, I. A., and Ibiwoye, A. (2012). Predictive actuarial modeling of health insurance costs. International journal of mathematics and computation, pages 14(1),1–2.

xiv. Taylor, J. (2008). A comparison of univariate time series methods for forecasting intraday arrivals at a call center. oxford.

xv. Tinbergen, J. (1939). Statistical Analysis of Business Cycle Theories. A Method and Its Application to Business Cycle Theory.

xvi. Tindogo, M. R. (2013). Forecasting utilization by subscribers of the national health insurance. Kumasi: KNUST.

xvii. Tulone, D. and Madden, S. (2016). Time series forecasting for approximate query answering in sensor networks.

xviii. Warren, M. P. (1919). Indices of business conditions. Review of Economic Statistics, pages 5–107.

xix. Benston, G., Bromwich, M., Litan, R. E. and Wagenhofer, A. (2004) Following the Money: The Enron Failure and the State of Corporate Disclosure, Brookings Institution Press.

xx. Berman, G. E. (2013)"Transformational Technologies, Market Structure, and the SEC," Remarks to the SIFMA TECH Conference, New York, http://www.sec.gov/News/Speech/Detail/Speech/1365171575716.

xxi. Bernstein, P. A. and Haas, L. M. (2008) "Information Integration in the Enterprise," Communications of the ACM, 51(9), September 72- 79.

xxii. Bholat, D. (2015) "Big data and central banks," Bank of England Quarterly Bulletin, Q1, 86-93.

xxiii. Bholat, D., Hansen, S., Santos, P. and Schonhardt-Bailey, C. (2015) "Text mining for central banks," Centre for Central Banking Studies Handbook (33), eprints.lse.ac.uk/62548/1/SchonhardtBailey_text%20mining%20handbook.pdf.

xxiv. Board of Governors of the Federal Reserve (2015) "Enhancements to the Federal Reserve System's Surveillance Program," Memorandum, December, http://www.federalreserve.gov/bankinforeg/srletters/sr1516.htm.

xxv. Burdick, D., Hernandez, M., Ho, H., Koutrika, G., Krishnamurthy, R., Popa, L. C., Stanoi, I., Vaithyanathan, S. and Das, S. R. (2015)"Extracting, Linking and Integrating Data from Public Sources: A Financial Case Study," working paper, September, http://ssrn.com/abstract=2666384.

xxvi. Casey, M. (2014) "Emerging Opportunities and Challenges with Central Bank Data," presentation slides, October, www.ecb.europa.eu/events/pdf/conferences/141015/presentations/Emerging opportunties_and_chalenges_with_Central_Bank_datapresentation.pdf?6074ecbc2e58152dd41a9543b1442849.

xxvii. Cerutti, E., Claessens, S. and McGuire, P. (2014) "Systemic Risks in Global Banking: What Available Data Can Tell Us and What More Data Are Needed?" Chapter16 in: Risk Topography: Systemic Risk and Macro Modeling, Brunner Meier,

xxviii. M. K. & Krishnamurthy, A. (Eds.), University of Chicago Press, 235-260.