A study of forecasting cardiovascular examinations in the Abbeville health center

4/18/2020

1.0 Introduction

1.1 Study Description:

The ability to peer into the future has always been one of man kind's greatest desires – knowing what will happen allows for one to plan and prepare optimally. As the costs of health care continue to rise coupled with the increase in population that are more in need of advanced healthcare, organizations in the industry can benefit greatly from better planning and preparation via use of forecasting to help control costs and provide an overall better service to their patients.

This study will take a deep dive into developing a forecast model, using advanced techniques, to predict the number of incoming cardiovascular examination requests at the Abbeville health center, which is an entity within the Fargo Health Group organization.

1.2 Problem definition:

A current opportunity that Fargo Health Group has identified is within its Disability Compensation Benefit process. There are two challenges within this process that Fargo faces: turn-around time for referral examinations from local offices (LO) to health clinics (HC), and the reliance on out-of-network referrals for examinations.

- First, the challenge regarding turnaround time can be costly. For every overdue day in excess of 30, Fargo is required to pay the Regional Health Oversight (ROHO) office $200/day per patient in fines.

- Secondly, while some health clinics can look at patient capacity and notify the local office it cannot meet 30 days for that referral, this is not a standard service level obtained for all HC's nor is it consistent even within the same HC. While each HC has best intentions in mind, not having a centralized method for predicting demand in such a manner leads to different results, and different "sources of truth". These inconsistencies generally lead to the use of referrals to out of network outpatient clinics (OC). In this event, the average cost of an exam spikes upwards to an additional $1,250. Additionally, Fargo is unable to control out-of-network OC's to comply and meet Fargo's 30-day mandates by the ROHO, therefore increasing risk tied to the first challenge noted above. Not only do these risks involve monetary and reputational costs to Fargo, but also more importantly they negatively impact the health and wellbeing of the patients – the core of Fargo Health Group's mission.

A key solution to minimizing these two risks and associated costs lie within the ability to accurately forecast incoming demand to Fargo's clinics. Understanding the demand will allow these clinics to accurately plan, prepare, and schedule optimally, so that the experts can do what they do best – which is to provide timely point of care services to the patients of Fargo Health Group.

1.3 Stakeholder identification & Project Planning:

As with any undertaking of a project, its important to identify key stakeholder entities at the onset. In doing so, we'll ensure that as we begin developing the project charter the scope will include all appropriate actors, viewpoints, and requirements prior to execution of the project plan, thereby reducing the risk of missing information or unforeseen changes (impacting original budget, timeframe, or scope) that may need to be included in order to complete the project successfully.

While we will use various tools and techniques (such as a RACI chart) within each entity to identify individual actors and their roles, it's important to point out at this time we have identified at a high level the Local Offices (LO), Health Clinics (HC), the Regional Health Oversight (ROHO), and out-of-network Outpatient Clinics (OC) all as important entities within this project – with the Quality Assessment Office (QAO) as the primary stakeholder and project champion leading the charge from the Fargo Health side. Specifically, Mr. Jay Rubin, Director of QAO, will act as the project sponsor.

Once engaged, we will begin by developing a project plan to identify key members of a cross-functional team that is inclusive enough to represent actors uncovered via RACI charts. A project manager will help to manage and maintain scrum style meetings with this core group and keep track of progress in this plan and visually via a Gantt chart. Finally, a scope change control process will be designed with key stakeholders, as well as a completed SWOT analysis for risk identification, which will help to plan how to mitigate these identified risks (or what the alternative work around will be) – stakeholders will play a very large role in this project.

2.0 Solutions Approach

2.1 The data-driven approach:

There are many methods for forecast planning, however time and time again has proven that a data-driven approach to forecasting is far superior than it is to go on "gut-instinct" or intuition alone. Just like the weather in a certain region of the country can be predicted with a great deal of accuracy, we can apply similar tools and techniques to assess a predicted demand for incoming examinations at the Abbeville HC location.

As technology has evolved so too has the tools and techniques, therefore the focus of our study will include these advanced methods to improve accuracy, which will ultimately lend to the overall goal of reducing risk through optimal planning and preparation. While it's important to point out that these models use cutting edge technologies and scientific approaches to provide their robust forecasts, we must not forget that it is just this – a forecast, or our best (data-driven) guess of the demand. Therefore, responsible use of this information is in aid of decision making, but not to dictate it.

2.2 Time-Series & the variable of interest:

Predicting the value of a certain event or variable at a specific point in time can be thought of as a Time-Series type of problem. We call this type of data longitudinal, and we repeatedly measure the variable of interest over and over again, and by observing how it reacts over time we can learn a great deal about it.

As it relates to the project, we consider the *number of incoming examinations (that are cardiovascular related)* as our variable of interest. Reviewing measures of this variable over time on a monthly basis will allow us to use forecasting to make predictive assertions on future values of that same cadence. In this study, we will complete a 12-month forecast to predict the number of incoming cardiovascular examinations by month for the year 2014 at the Abbeville HC.

2.3 Time-Series Components & Models:

In data-driven approaches to problem solving, we build models attempt to represent these real-world scenarios. At their essence, models are nothing more than a mathematical equation (albeit complex) for which inputs are fed in, the model calculates, and outputs are returned for interpretation.

In order to make the 12-month forecast for our project, we must first build a model that "tunes in" to the existing data. This process of "tuning in", also known as "fitting", helps the model uncover hidden patterns and allows for the model to get an understanding of the data itself – thus gaining an understanding of the scenario in question. Within Time-Series data, there are three main components that these models look to understand – they are: trend, seasonality, and random error.

- **Trend** component of a Time-Series model helps to describe a particular observed increase or decrease measured in the variable of interest over a given amount of time. A business who continues to see increased sales year over year for the last 5 years may be expressed as a "positive trend" in sales. The trend component also attempts to describe the magnitude of that increase or decrease. Does the positive trend increase at a constant (linear) rate? Or is there steep pattern to the increase (exponential, etc.).

- **Seasonality** looks to describe a periodic pattern within the data. Consumer good stores generally see an uptick in sales around the end of year in November and December during the Holiday season, however these sales return to normal in the following months.

- Finally, **random-error** (sometimes known as "Irregular"), is a component of the Time-Series that describes the changes in the variable of interest that are not defined by the trend or seasonality.

The purpose to define and understand these components of Time-Series data is because it helps guide us toward appropriate model selection. Some models do not perform as well in certain scenarios, so it's important for us to complete a preliminary check process on the data to evaluate these components within the dataset. We will search for evidence of seasonality, trend, and irregular via a decomposition analysis as outlined in section 4.1 - Decomposition.

3.0 Data cleansing & imputation

3.1 Missing Values:

As with other things in the real-world, data is never in the pristine condition that is required in order to begin building the model as outlined above. A process of data cleansing and error handling is necessary to begin shaping the existing data set into something that will be consumable by the selected model – such is the case with the data set provided by Fargo. In this section we will discuss the existing data set, and what steps were taken in order to identify missing values, handle data formatting issues, and complete imputation.

To begin, we start by evaluating the Abbeville, LA worksheet with the dataset workbook. Our first step is to consistently codify missing values in the exam variable. The following entries are considered missing: Closed for Holidays, *, Entered by J.f. Williams, xx?*$?/..

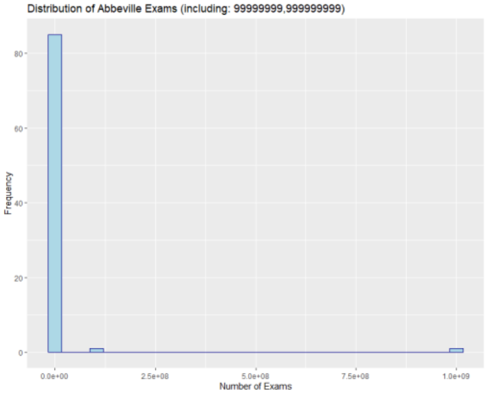


Fig. 1

After accounting for text-based missing values, we review a histogram of the Abbeville data to review its distribution (Fig. 1). An initial review of the distribution indicates an extremely right-skewed data set that points to potential outliers.

We then set out to review those data points and uncover values of 99999999 & 999999999 respectively, which explain the skew. We make the assumption that the following are also considering missing values, therefore we code them consistent to the previous set. We rerun the histogram and review the results (Fig. 2). Now we see a more accurate representation of the distribution with the missing values removed.

This set of missing values (11 in total) will be candidate for our imputation process as further described later in section 3.4 - Imputation.

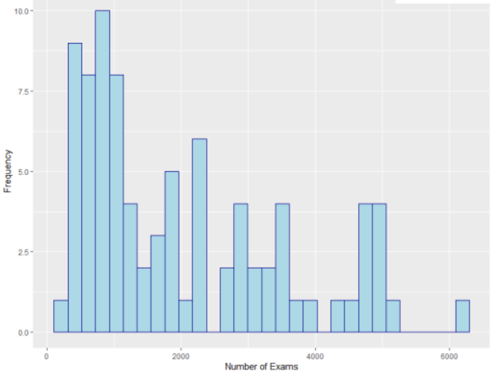


Fig. 2

3.2 Data Formatting:

Data formatting problems can stem from a myriad of underlying issues, such as: data entry methods, validations in place to handle malformed input, data transferred from one source to the next, etc. In the data set provided we uncovered inconsistencies in the formatting of date values within the various HC worksheets. For example, the Violet worksheet had entries consisting of 5/14/2007 vs. 14 May, 2007 – for which the latter format caused inconsistent results with frequency measures. The solution is simple in these cases, we apply consistent formatting across all values before analysis begins. In this case we simply converted entries similar to 14 May, 2007 to match the consistent format, 5/14/2007.

3.3 Incomplete Data:

The next consideration in the data cleansing process is to evaluate incomplete data – are there values within the data set that don't represent the true value of that particular data point? In the context of the project, do the numbers of exam counts in the Abbeville worksheet represent an accurate reflection of the true number of exams? To begin this process, we must consult with the stakeholders and subject matter experts identified in section 1.3 to gain perspective of events that may have resulted in incomplete data.

First we learn that within the data, the number of exams for May 2007 are a bit underrepresented due to a renovation project that occurred on May 2nd. To complete the true number of May 2007 exams, we filter the other HC worksheets for "Abbeville" + "5/2/2007" + [heart related exam]. This total is then added back into the Abbeville number for May 2007 – a total of 7 health related exams were discovered missing, for a total of 114 in that month.

Next we review incomplete data for the months of May, June, July of 2013. We use similar logic as we did with the previous problem, only this time we are filtering for the condition: “Abbeville”+”[total month of May]”+[heart related exam]. This is not only one date, but then entire month of May. We repeat this process for June & July and tabulate the results (Fig. 3).

	Abbeville (Original)	**Violet**	**New Orleans**	**Lafayette**	**Baton Rouge**	**Abbeville (Total)**
May 2013	4730	0	297	174	146	5347
June 2013	4729	0	0	19	4	4729
July 2013	5000	0	0	4	10	5014

Fig. 3

3.4 Data Imputation:

At its core, data imputation is a set of processes or techniques that are followed in order to derive any missing data within a data set. The process of data imputation can take on multiple forms, and with the given data set in this project we implemented a handful of these techniques. Before we dive into the imputations performed in this project, it’s important to note that we concluded the missing data is considered “Missing Completely at Random” (MCAR). This is typically the most common scenario and explains there is no systematic reason for why the data is missing – this allows us to use the standard processes and techniques to impute the missing data within this data set.

We first start out by reviewing the missing data for the December 2009 to February 2010 time period. It is noted there were a total of 5,129 cardiovascular exam requests during this time period. For this we used a standard rational approach by taking the average of 5,129 over these 3 months and using the value to populate the missing data. Last note on this is the 5,129 does not divide evenly over the 3 months, so the Jan and Feb received an additional 1 to its exam count (Fig. 4).

	Abbeville (Original)	**Abbeville (Total)**
Dec. 2009	NA	1709
Jan. 2010	NA	1710
Feb. 2010	NA	1710

Fig. 4

The next missing value we impute is the month of December 2013. This was a bit more complicated as it required evaluating and interpreting observations within the December 2013 Data worksheet. Those observations prefix with “L839” and suffixed with either “TGU3” or “ROV8” identified them as routed from Abbeville. Next we match the middle piece of the string against one of the heart-related codes via the Heart-related Condition Codes worksheet. If these components of an observation identify as a cardiovascular exam and from Abbeville then we count it into our final imputed value, which calculated as 5,933 cardiovascular exams for this month.

The remaining missing data points do not contain any other clues or rational approaches; therefore, we use an advanced method for imputation of these values (R via MICE package). This approach, called Multiple Imputation, creates multiple copies of the existing data set and then proceeds to use a statistical simulation technique called a “Monte Carlo simulation” to calculate the missing values via a “Predictive Mean Matching” process. The Monte Carlo technique allows for random error to be introduced into the process, therefore closely emulating real-world scenarios. We allowed 5 such iterations of the multiple imputation process, and for each remaining missing observation we averaged these 5 values in this simulated data set to impute the final number of cardiovascular examinations (Fig. 5).We finally have a completed data set for which we can now begin the process of model selection, fitting, and forecasting.

	Abbeville (Original)	Abbeville (Total)
Mar. 2006	NA	468
Jun. 2006	NA	481
Dec. 2008	NA	1463
May 2009	NA	1175
Jun. 2010	NA	1646
Jan. 2011	NA	2001
Dec. 2011	NA	2878

Fig. 5

4.0 Analysis & Modeling

4.1 Decomposition:

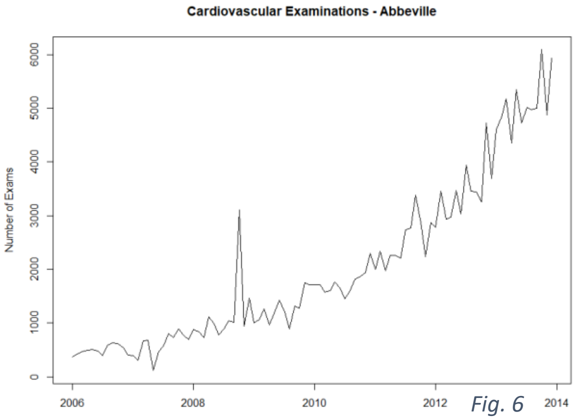


Fig. 6

To begin the process of analysis, we first plot the completed Time-Series data set and review for evidence of trend or seasonality (Fig. 6). We also look for increased variability as time increases, indicating a need for data transformation prior to model fitting – this effort of transformation helps the model perform better. Upon initial investigation, we notice evidence of a positive trend to the data and increased variability as time increases. For this, we will use a log transformation version of the dataset to fit our models, which we’ll re-transform prior to reviewing actual forecast estimates.

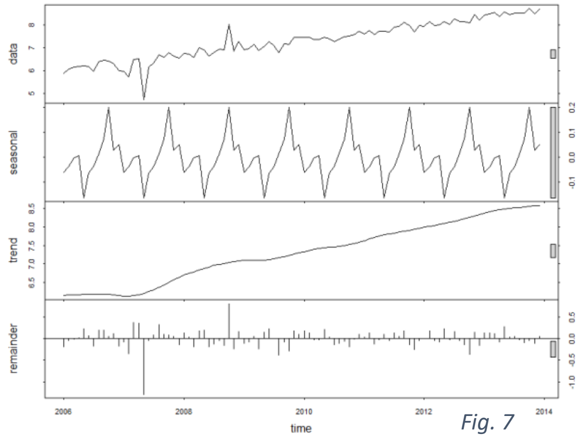


Fig. 7

Because there is evidence of trend, we’ll complete a decomposition of the logged data set to understand its influence numerically, as well as get a look at the seasonal component (as it’s difficult to see this in the original plot). The figure on the left (Fig. 7) continues supporting the evidence of our trend, and it also sheds some detail into a seasonality component of the data set.

Finally, a numerical review of the decomposed data set further validates the suspicion of seasonality and trend (fig. 8). From the figure, we can extrapolate the seasonality appears to cycle from May to October each year, with about a 16% reduction in exams in May and an increase of 22% in October. (Note: the values listed in the table under seasonal are considered a “multiplier”, so in May it’s typically 0.84 times the normal rate, vs. 1.22 times the normal rate in October). In summary, we will select models that take into account these factors of the dataset – more in the next section.

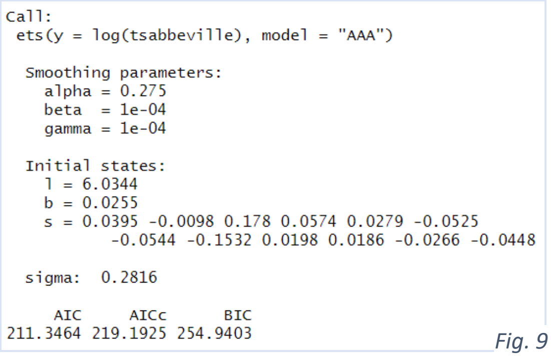
Period	seasonal	trend
May 2006	0.848337	481.0121
Oct 2006	1.224012	479.4948
May 2007	0.848337	492.7503
Oct 2007	1.224012	664.5561
May 2008	0.848337	966.7046
Oct 2008	1.224012	1133.344
May 2009	0.848337	1202.396
Oct 2009	1.224012	1381.959
May 2010	0.848337	1686.612
Oct 2010	1.224012	1854.569
May 2011	0.848337	2401.354
Oct 2011	1.224012	2725.472
May 2012	0.848337	3267.818
Oct 2012	1.224012	3836.359
May 2013	0.848337	4784.349
Oct 2013	1.224012	5205.581

Fig. 8

4.2 Holt-Winters Model:

The Holt-Winters (HW) model is an appropriate selection for our project due to the fact that we've identified both a trend and seasonal component within the data set. We fit our HW model using the log transformed dataset as discussed in the previous section and review the summary results in fig. 9.

First we review the beta and gamma values under the smoothing parameters section – these negligible values are indicating that the trend and seasonality don't change much over time, and therefore decay the upcoming forecast accordingly. Next, we review the AIC score, which is an indicator of model performance, and is a useful criterion for comparing models against one another - we note an AIC of 211.34, which will be used for comparison in the next model.



```
Call:
 ets(y = log(tsabbeville), model = "AAA")

  Smoothing parameters:
    alpha = 0.275
    beta  = 1e-04
    gamma = 1e-04

  Initial states:
    l = 6.0344
    b = 0.0255
    s = 0.0395 -0.0098 0.178 0.0574 0.0279 -0.0525
           -0.0544 -0.1532 0.0198 0.0186 -0.0266 -0.0448

  sigma:  0.2816

     AIC     AICc      BIC
211.3464 219.1925 254.9403
```

Fig. 9

Next we check the "goodness" of the model fit onto the data. While the numeric AIC score serves as one indication, we will visually check the residuals of the model fit for two assumptions: is the variance constant, and do the residuals generally follow a Gaussian (normal) distribution? Fig. 10 shows a residual plot, which generally centers around 0.0 on the y-axis with the exception of May 2007 and Oct 2008 (which were identified as outliers for clinic closure and the hurricane event respectively). Since the residuals stay close to 0.0, and there are no trends or patterns in the residuals getting larger or smaller, this indicates a good fit to the data.

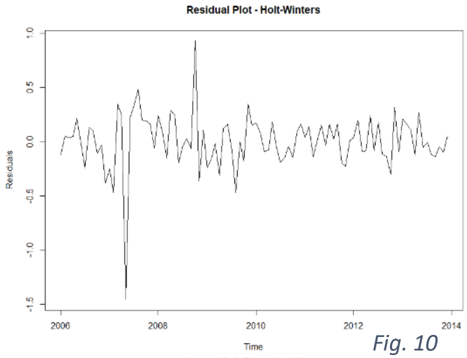


Fig. 10

Fig. 11 displays a Q-Q Plot which helps to visually describe the residuals fit against a normal distribution. This test allows us to ensure there is no inter-dependence on the residuals – again indicating there is a pattern that we may have not accounted for within our model. Reviewing the plot, we notice the majority of the residuals (black dots) fall onto the line (theoretical normal distribution). Since they align, we again have evidence of good model fit.

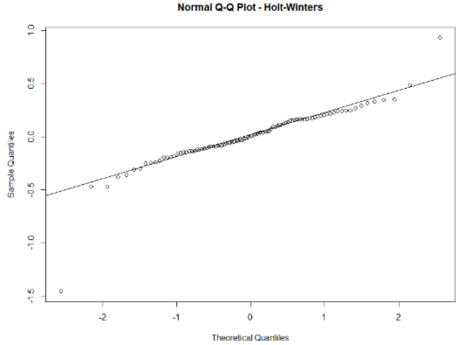
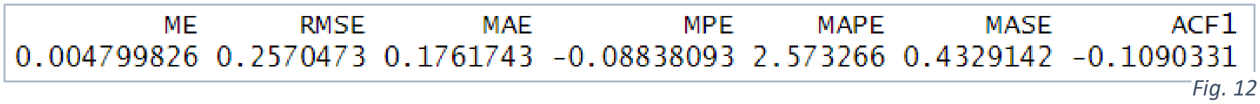


Fig. 11

Our last metric to review is the accuracy of the model. We do this by checking various predictive accuracy measures, and just like with the AIC score, we will compare the values against another candidate model. Fig. 12 displays the accuracy measures for the HW model.



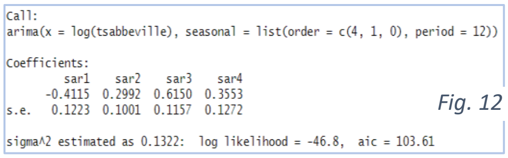
```
         ME      RMSE       MAE         MPE     MAPE      MASE       ACF1
0.004799826 0.2570473 0.1761743 -0.08838093 2.573266 0.4329142 -0.1090331
```

Fig. 12

A good comparison measure of accuracy in this project will be the Root Mean Squared Error (RMSE), which in this model is: 0.2570473 log units.

4.3 ARIMA Model:

The auto-regressive integrated moving average (ARIMA) model is the second candidate model for our project. To begin evaluating the ARIMA model, we first check our log transformed dataset for stationarity using the ndiff function in R. We determine that it requires one difference transformation to become stationary – this defines our “d” term in the ARIMA model. From here, we use the auto.arima() function in R to determine the auto-regressive (AR) and the moving average (MA) terms – this results in an ARIMA(4,0,1) model selection. Because our data has seasonality, we fit the model with an attempt to add seasonal differencing into the it by specifying the seasonal parameter (Fig. 12). Our final model is: ARIMA(4,0,1)x(0,12,0), with an AIC = 103.61.



```
Call:
arima(x = log(tsabbeville), seasonal = list(order = c(4, 1, 0), period = 12))

Coefficients:
         sar1    sar2    sar3    sar4
      -0.4115  0.2992  0.6150  0.3553
s.e.   0.1223  0.1001  0.1157  0.1272

sigma^2 estimated as 0.1322:  log likelihood = -46.8,  aic = 103.61
```

Fig. 12

Now we’ll check both the autocorrelation (Fig. 13) and partial autocorrelation (Fig. 14) plots to check stationarity of our model, and we notice the first few lags are not significant (breach the dotted lines), indicating stationarity.

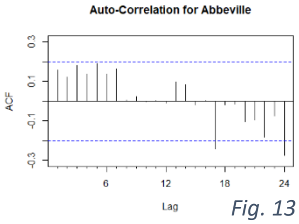


Fig. 13

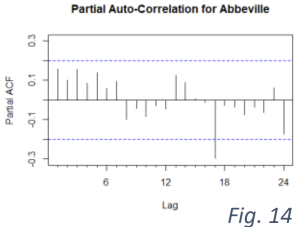


Fig. 14

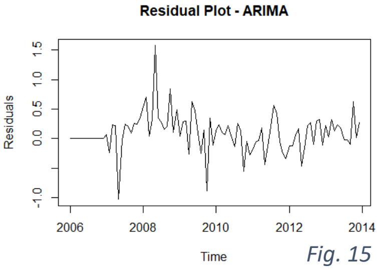


Fig. 15

As we did in the first model, we now check the “goodness” of fit onto the data. We start with a visual check of the residuals in Fig. 15 to see if there are any patterns within the error. Outside of the previously noted outliers, we see a slight pattern of oscillation, however they generally tend to stay centered around the 0.0 point.

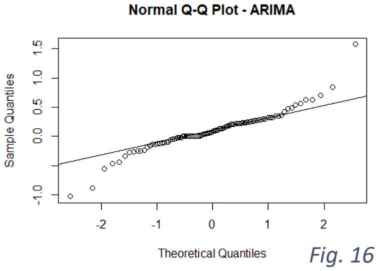
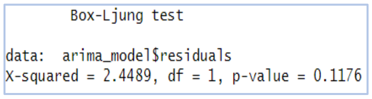


Fig. 16

Fig. 16 displays the ARIMA model’s Q-Q Plot which helps to visually describe the residuals fit against a normal distribution. Reviewing the plot, we notice the majority of the residuals fall onto the line and can be considered normally distributed, however it is noted that there are slight drifts toward the lower and higher quantiles.



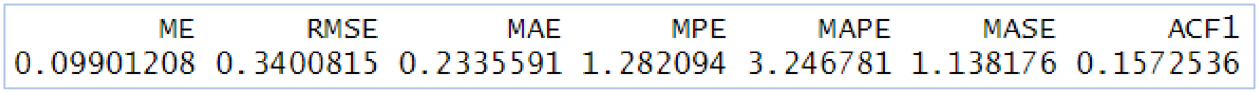
```
        Box-Ljung test

data:  arima_model$residuals
X-squared = 2.4489, df = 1, p-value = 0.1176
```

Fig. 17

The last check for fit is to complete a Ljung-Box test on the residuals, which will test if the autocorrelations are all equal to 0 (which is the null hypothesis). The results of this test (Fig. 17) shows a p-value of 0.1176 which indicates that there is not enough evidence to reject that the autocorrelations are truly equal to 0 – removing autocorrelation indicates stationarity in the dataset, and therefore indicates the model is fitting well to the data.

Our last method of evaluation to review is the accuracy of the model. We do this again by checking various predictive accuracy measures – see Fig. 18 for the ARIMA model’s outcomes.



```
         ME      RMSE       MAE      MPE     MAPE     MASE      ACF1
0.09901208 0.3400815 0.2335591 1.282094 3.246781 1.138176 0.1572536
```

Fig. 18

We note an RMSE of 0.3400815 log units, of which is higher than the Holt-Winters model described above.

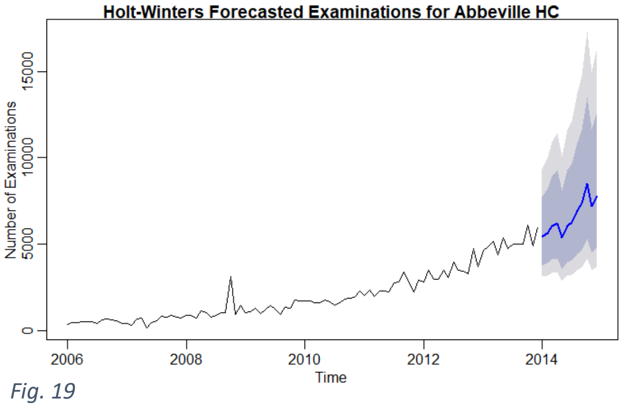
5.0 Forecasting

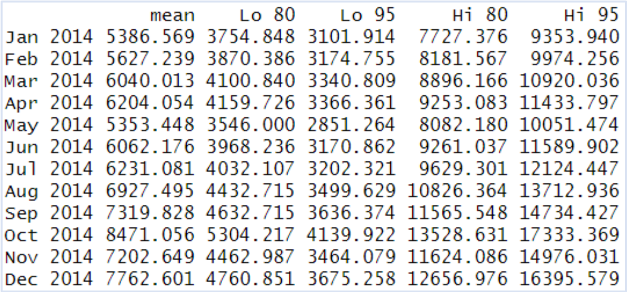
5.1 Forecast & Re-transformation:

Now that we've successfully built our models, assessed their fit, and checked various performance measures, we can begin to use them in order to make our 12-month forecasts. However, prior to proceeding forward it's important to note that we've been working with log values of the incoming examinations, therefore we will need to re-transform the predicted dataset back to the normal scale. To do this, we apply the natural exponent to the log values to bring them back to true scale.

5.2 Holt-Winters Forecast:

Our first candidate model provides the forecast displayed in Fig. 19 for the year of 2014. The graph shows a clear continuation of both the trend and seasonality (a dip in May, an increase in Oct.). The chart shows the data associated with the graph – the mean is the suggested forecast and is represented on the graph by the blue line. The blue "band" around the line demonstrates an 80% confidence interval, and the subsequent grey band shows a 95% confidence interval. This simply means the true number of incoming demand may fall within the band, for a given month, with a certain level of confidence.



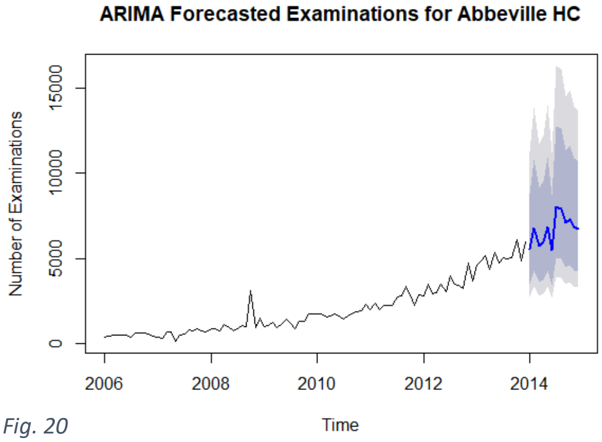


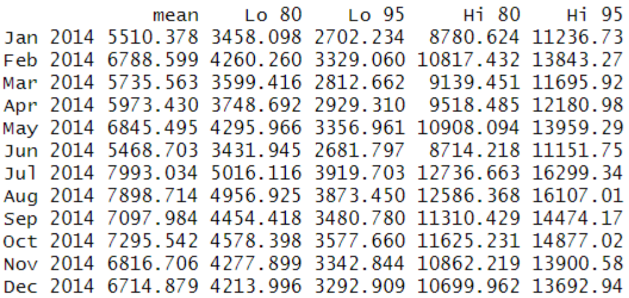
```
               mean    Lo 80    Lo 95     Hi 80     Hi 95
Jan 2014 5386.569 3754.848 3101.914  7727.376  9353.940
Feb 2014 5627.239 3870.386 3174.755  8181.567  9974.256
Mar 2014 6040.013 4100.840 3340.809  8896.166 10920.036
Apr 2014 6204.054 4159.726 3366.361  9253.083 11433.797
May 2014 5353.448 3546.000 2851.264  8082.180 10051.474
Jun 2014 6062.176 3968.236 3170.862  9261.037 11589.902
Jul 2014 6231.081 4032.107 3202.321  9629.301 12124.447
Aug 2014 6927.495 4432.715 3499.629 10826.364 13712.936
Sep 2014 7319.828 4632.715 3636.374 11565.548 14734.427
Oct 2014 8471.056 5304.217 4139.922 13528.631 17333.369
Nov 2014 7202.649 4462.987 3464.079 11624.086 14976.031
Dec 2014 7762.601 4760.851 3675.258 12656.976 16395.579
```

Fig. 19

5.3 ARIMA Forecast:

The second candidate model provides the forecast displayed in Fig. 20 for the year of 2014. The ARIMA model typically has a stronger affinity toward the trend component of the data, although it does reflect the seasonality component with our additional parameter specified. We again see the mean prediction, along with the confidence intervals at the 80 and 95 percentiles.





```
               mean    Lo 80    Lo 95     Hi 80    Hi 95
Jan 2014 5510.378 3458.098 2702.234  8780.624 11236.73
Feb 2014 6788.599 4260.260 3329.060 10817.432 13843.27
Mar 2014 5735.563 3599.416 2812.662  9139.451 11695.92
Apr 2014 5973.430 3748.692 2929.310  9518.485 12180.98
May 2014 6845.495 4295.966 3356.961 10908.094 13959.29
Jun 2014 5468.703 3431.945 2681.797  8714.218 11151.75
Jul 2014 7993.034 5016.116 3919.703 12736.663 16299.34
Aug 2014 7898.714 4956.925 3873.450 12586.368 16107.01
Sep 2014 7097.984 4454.418 3480.780 11310.429 14474.17
Oct 2014 7295.542 4578.398 3577.660 11625.231 14877.02
Nov 2014 6816.706 4277.899 3342.844 10862.219 13900.58
Dec 2014 6714.879 4213.996 3292.909 10699.962 13692.94
```

Fig. 20

6.0 Conclusion & Ethical Considerations

6.1 Summary & Results

In this project report we took a deep dive on a study involving the prediction of cardiovascular examinations at the Abbeville health center. From clearly defining the business problem and identifying stakeholders, to outlining the data-driven approach and how we dealt with data quality issues, we build two models and evaluated their effectiveness. We then made forecasts using these two models and discussed their outputs.

Although the Holt-Winters is the more parsimonious model, we feel our ARIMA model enhanced for seasonality performs better with an AIC=103.61 and RMSE=0.3400815 log units. Potential enhancements to the seasonal ARIMA model may be found by further investigating the residual and Q-Q plots, and applying further transformations to the data to enhance fit.

6.2 A Note on Ethics & Governance Considerations

As with any data-driven solution, we must take into account various ethical considerations. First, we note that the original purpose of the data set used was not to build the predictive models as outlined in this report – this is a "secondary utility" of the data set. This can represent a number of challenges for Fargo if they are not considered for further action. Specifically, when it comes to consent this can bring up privacy concerns with those individuals with whom the data represents. It would be in the best interest of Fargo to alert their patients of a new data collection policy that outlines not only the use of their de-identified information for these business purposes, but also the actions being taken to protect their data and privacy via a very heavily regulated industry by HIPPA laws.

Next we consider the assumptions made while creating the minimum viable data set to fit the models. As discussed in section 3.0, we made multiple assumptions in place of missing data, formatting issues, and imputation techniques. While these are all necessary to come up with a final solution, they do run risk of introducing bias into our solution, an example of what we considered "missing" could be different based on different approach. Or, a different technique could be used to impute this data, which may return different results in the final product.

This leads us to our last consideration – how does Fargo put a data governance policy in place that provides guidance on developing a sufficient data infrastructure as well as identify data stewards that take ownership and are held accountable to the quality, and timeliness, of data? This should not be considered a single department like the QAO's responsibility but is a joint effort at an organizational scale. Individuals at each location with various titles, from administration to providers, should all have a stake and responsibility in the data that feed these models.

6.3 Go-Forward & Future Considerations

This study provides a small sample use case for a data-driven approach to forecasting the potential demand of cardiovascular examinations at one HC within the Fargo group. Prior to deploying and scaling this solution for other locations or exam types, it's advised that Fargo continue researching the ethical challenges mentioned and adopt a data governance strategy for best results going forward.