Business Analytics Using Forecasting (Fall 2016)
Final Report

Cultivating talents for actual needs:

Forecasting numbers of patients as reference for medical profession enrollment

Team 3

Yu-Chu Shih
Tzu-Han Hung
Yi-Chun Chuang
Tonny Meng-Lun Kuo

Instructor: Galit Shmueli
Institute of Service Science
National Tsing Hua University

Executive Summary

Every year, Ministry of Health and Welfare in Taiwan (MHW) decides the capacity of medical students to ensure each division includes enough doctors in the future. However, recent report indicates the huge shortage of doctors in five demanding divisions which leads to serious problem in hospital management[1]. Deciding the capacity of medical students in specific divisions is challenging for MHW.

Fortunately, with the affordances of data forecast techniques and algorithms, more accurate forecasting methods could be used to offer a meaningful reference for decision making. Therefore, this report aims to provide Department of Medical Affairs of MHW with a reference to better distribute medical students capacity of 5 specific divisions (General Medicine, General Surgery, Pediatrics, Obstetrics and Gynecology, and Emergency Room). Using the yearly patient visits data from MHW and other external data (1998-2014), this study forecasts the future number of patient visits in 5 divisions. In particular, we forecast the three-year ahead yearly patient visits. Due to the nature of our data (with trend but no seasonality), we applied different forecasting methods such as naive, regression, moving average, and ensemble to forecast the models. By using performance evaluation measures MAPE (mean absolute percentage error) and RMSE (root-mean-square error), we identified the best model to each series.

Our findings indicates a significant increase of GS (9.10 %) and PED (11.86 %) in 2017. Additionally, there are no significant changes in OB (0.85 %) and ER (-0.29%) compared to other series. Based on our findings, we recommend MHW consider allocating more doctors capacity in GE and PED in 2017-2018. However, although more demands are expected in PED in 2017, considering the decrease of birth rate, the MHW should carefully investigate the reasons behind the visits change to develop more comprehensive medical policies. Moreover, because forecast uncertainty of ER visits (ca. 10%) is higher than other series, we suggest MHW considers upper bound rather than lower bound to avoid doctor shortage.

[1] Zheng, Y. I. & Li, S. R. (2015). Doctor shortage in five demanding divisions: 500 government-funded doctors re-open. Retrieved from http://udn.com/news/story/6886/1292392

Detailed Report

Problem description

Every year, Ministry of Health and Welfare in Taiwan (MHW) decides the capacity of medical students to ensure that each hospital division includes enough doctors in the future. However, a recent report indicates the huge shortage of doctors in five demanding divisions which may lead to serious problems in hospital management, such as work overload, doctor shortage, or low medial quality (Zheng & Li, 2015). In reality, deciding the capacity of medical students in specific divisions is challenging for MHW.

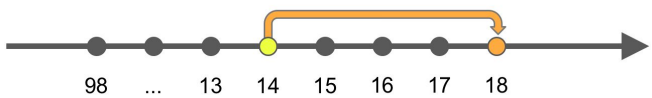
Business goal: Provide a meaningful reference to better distribute medical students capacity of 5 specific divisions (GM, GS, PED, OB, and ER) for improving medical quality.

- Client: Department of Medical Affairs (Ministry of Health and Welfare, MHW)
- Stakeholders: Patients, doctors, medical schools and hospitals
- Humanistic implication: Over-forecasting might waste financial resources because training a doctor costs a lot more than training other professions. Also, over-forecasting suppresses the capacity of other divisions. Other division might face doctor shortage. However, under-forecasting may lead to insufficient man-power, which is more serious than over-forecasting.
- Opportunity: Time-series forecasting provides more precise and data-driven forecasts which helps decision making.
- Challenges: MHW should also take other factors into account to make decisions. Patient visits forecast solely is not enough.

Forecasting goal: Forecast the future number of patient visits in five major divisions for 2017-2018

$Y_{t=1}$: denotes the number of patient visits in 1998.
F_t: denotes the forecasted number of patient visits for year ***t***.

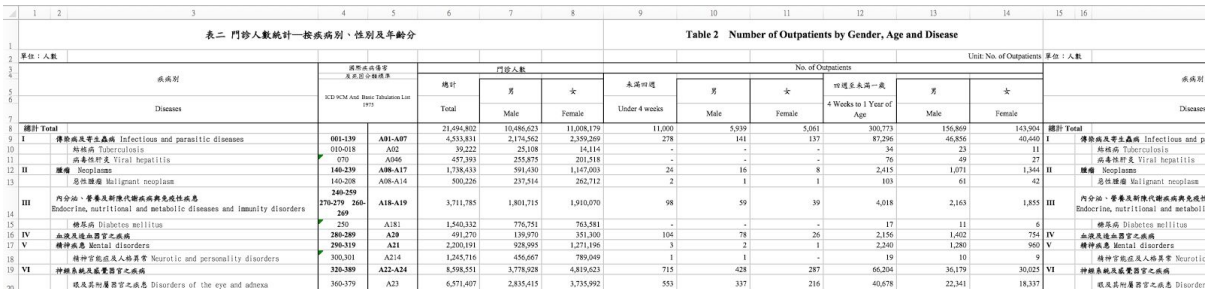
k = 1, 2, ...4. This is a 4-step-ahead forecast. [2]



Data description

We collected the Yearly Medial Report (1998-2014) from Statistics Department of Ministry of Health and Welfare (http://www.mohw.gov.tw/CHT/DOS/Statistic.aspx?f_list_no=312&fod_list_no=1604). Our data of interest is the annual number of patient visits in GM, GS, OB, ER and PED.

[2] MHW requires the next (two) year forecast to decide the doctoral capacity. However, our data is only available until 2014, to generate forecast for year 2017-2018, we need to set our forecast horizon to 4 years. In year 2016, we'll focus on the forecast in 2017, which is a 3-year ahead forecast.



表二 門診人數統計—按疾病別、性別及年齡分

Table 2 Number of Outpatients by Gender, Age and Disease

單位：人數 Unit: No. of Outpatients

疾病別 Diseases	ICD 9CM And Basic Tabulation List 1975		門診人數 No. of Outpatients 總計 Total	男 Male	女 Female	未滿四週 Under 4 weeks	男 Male	女 Female	四週至未滿一歲 4 Weeks to 1 Year of Age	男 Male	女 Female
總計 Total			21,494,802	10,486,623	11,008,179	11,000	5,939	5,061	300,773	156,869	143,904
I 傳染病及寄生蟲病 Infectious and parasitic diseases	001-139	A01-A07	4,533,831	2,174,562	2,359,269	278	141	137	87,296	46,856	40,440
結核病 Tuberculosis	010-018	A02	39,222	25,108	14,114	-	-	-	34	23	11
病毒性肝炎 Viral hepatitis	070	A066	457,393	255,875	201,518	-	-	-	76	49	27
II 腫瘤 Neoplasms	140-239	A08-A17	1,738,433	591,630	1,147,003	24	16	8	2,415	1,071	1,344
惡性腫瘤 Malignant neoplasm	140-208	A08-A14	500,226	237,514	262,712	2	1	1	103	61	42
III 內分泌、營養及新陳代謝疾病與免疫性疾患 Endocrine, nutritional and metabolic diseases and immunity disorders	240-259 270-279 280-289	A18-A19	3,711,785	1,801,715	1,910,070	98	59	39	4,018	2,163	1,855
糖尿病 Diabetes mellitus	250	A181	1,540,332	776,751	763,581	-	-	-	17	11	6
IV 血液及造血器官之疾病	280-289	A20	491,270	139,970	351,300	104	78	26	2,156	1,402	754
V 精神疾患 Mental disorders	290-319	A21	2,200,191	928,995	1,271,196	3	2	1	2,240	1,280	960
精神官能症及人格異常 Neurotic and personality disorders	300,301	A214	1,245,716	456,667	789,049	1	1	-	19	10	9
VI 神經系統及感覺器官之疾病	320-389	A22-A24	8,598,551	3,778,928	4,819,623	715	428	287	66,204	36,179	30,025
眼及其附屬器官之疾患 Disorders of the eye and adnexa	360-379	A23	6,571,407	2,835,415	3,735,992	553	337	216	40,678	22,341	18,337

Figure 1. **Screenshot of yearly medical report (raw data)**

Data preparation

The selected data of interest are from 17 reports and has been integrated into one data sheet of 5 series with 17 rows. We also include useful external data such as birth rate, death rate, population, proportion of aged population, annual average expenditure each person, expenditure on disease prevention are combined into another data sheet.

	A	B	C	D	E	F	Pop
1	Year	GM	GS	Pediatrics	OB	ER	
2	1998	8514481	3236774	3224608	3174631	4379568	
3	1999	8485101	3189756	3420196	3393981	4781756	
4	2000	8163870	3096545	3647164	3490058	4961941	
5	2001	8037544	3137589	3690612	3538893	5167025	
6	2002	8012999	3096856	3847997	3472433	5523558	
7	2003	7584501	2910109	3894194	3337983	5234851	
8	2004	7793822	3022567	4143095	3575235	5857425	
9	2005	7682190	2930573	4408299	3464324	5955062	
10	2006	7475060	2779461	4312651	3337165	5683931	
11	2007	9539080	3060094	4917917	3435142	5798377	

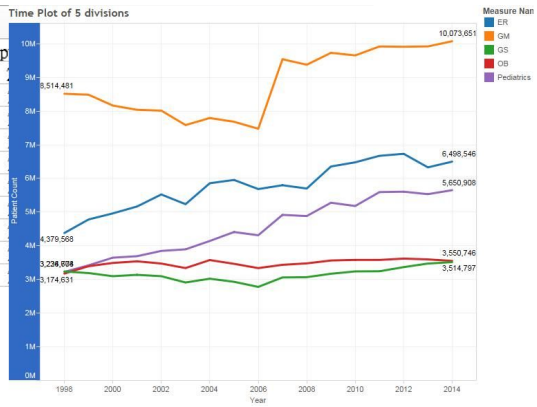


Figure 2. **The sample screenshot of 10 rows**

Figure 3. **Time plot of 5 series**

A	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
Year	OB	ER	Population_	Younth_lag	Adult_lag	Old_lag	Average an	No.of.Birth	Birth.Rate_	Male.Birth_	Female.Birt	No.Deaths_	Crude.Deat
1998	3174631	4379568	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
1999	3393981	4781756	21928591	4815400	15302960	1810231	22425	271450	12.43	141462	129988	123180	5.64
2000	3490058	4961941	22092387	4734596	15492319	1865472	24024	283661	12.89	148042	135619	126113	5.73
2001	3538893	5167025	22276672	4703093	15652271	1921308	24699	305312	13.76	159726	145586	125958	5.68
2002	3472433	5523558	22405568	4661884	15770327	1973357	25664	260354	11.65	135596	124758	127647	5.71
2003	3337983	5234851	22520776	4598892	15890584	2031300	27480	247530	11.02	129537	117993	128636	5.73
2004	3575235	5857425	22604550	4481620	16035196	2087734	29045	227070	10.06	118984	108086	130801	5.8
2005	3464324	5955062	22689122	4387082	16151565	2150475	31389	216419	9.56	113639	102780	135092	5.97
2006	3337165	5683931	22770383	4259049	16294530	2216804	32878	205854	9.06	107378	98476	139398	6.13
2007	3435142	5798377	22876527	4145631	16443867	2287029	34282	204459	8.96	106936	97523	135839	5.95

Figure 4. **The sample screenshot of external data**

Forecasting solution

Considering the nature of our data (series with trend but no seasonality), we applied the following methods to our data (see Figure 5):

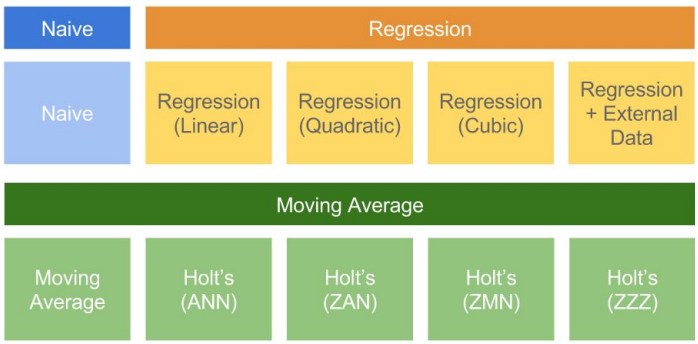


Figure 5. **Methods applied to 5 series**

Additionally, we have tested autocorrelation in our series with ACF plots, in which we found autocorrelation only in PED but not in other series. Therefore, we applied ARIMA to PED only. External data series are selected based on correlation with each of our time series, and they are lagged before applying to the regression.

To evaluate the performance of the methods, we used the roll-forward approach and compared the RMSE/ MAPE of overall, one-year, two-year, three-year and four-year ahead respectively in the validation period. We selected three best methods in each series, combined them with weighted ensemble approach, and then compared the ensemble performance with all other methods'. Finally, we found our best model for each series. Two models are reported for ER since one performs better in three-year ahead forecast and the other perform well overall. The selected best methods are listed in the table below:

Table 1. Best method for each series

GM	GS	OB	ER	ER2	PED
Holt's (ANN)	Ensemble (Quadratic regression, Holt's AAN, Regression with external data)	Naive (Holt's ANN, Naive, Regression with external data)	Holt's (ANN)	Ensemble	Holt's (ANN)

The performances of each method are shown in Appendix 1.

Future forecasts are made using the selected methods (see Table 1). Table 2 shows the forecasting results for each series in year 2015-2018 and their 90% prediction interval in 2017.

Table 2. The forecasted results of 2015-2018 and prediction interval of 2017

	Actual	Forecast				2017 Prediction Interval	
	2014	2015	2016	2017	2018	90% Lower Bound	90% Upper Bound
GM	10073651	10132464	10204426	10276387	10348349	10149397	10447575
		0.58%*	1.30%	2.01%	2.73%	0.75%	3.71%

ă

GS	3514797	3630081	3730083	3834589	3943598	3760056	3889065
		3.28%	6.13%	9.10%	12.20%	6.98%	10.65%
OB	3550746	3581080	3581080	3581080	3581080	3512857	3619477
		0.85%	0.85%	0.85%	0.85%	-1.07%	1.94%
ER	6498546	6479600	6479600	6479600	6479600	6093891	6791544
		-0.29%	-0.29%	-0.29%	-0.29%	-6.23%	4.51%
ER2	6498546	6617002	6658869	6700736	6742603	6284950	6923678
		1.82%	2.47%	3.11%	3.76%	-3.29%	6.54%
PED	5650908	5997380	6159178	6320975	6482773	5993412	6514934
		6.13%	8.99%	11.86%	14.72%	6.06%	15.29%

* Expected increasing rate: 10132464 (GM forecast for 2015) / 10073651 (GM actual value in 2014) = 0.58%

To summarize the results, GS and PED are expected to grow by 9.10% and 11.86% respectively. On the other hand, there is no significant change in GM, OB and ER.

Note. The time plots of 5 series forecasting results of are shown in Appendix 2.

Conclusion

Our study employed different forecasting methods to forecast the future patient visits in 2017-2018. After evaluating them with MAPE and RMSE, we found advanced exponential smoothing (Holt's) and ensemble best fit our series. According to our forecasting results, our recommendations show as follows:

1. Number of visits to General Surgery and Pediatrics are expected to increase significantly. MHW should consider allocating more doctors in these two divisions in the future.
2. We should carefully investigate the reasons behind the visits change to develop more comprehensive medical policies. For example, PED is expected to increase significantly, it is probably because nowadays people have only one child, and therefore, parents pay too much attention to their child, even just in small syndrome. In such cases, MHW should not significantly increase the capacity of PED division. Rather, it should place more efforts on educating the parents.
3. Forecast uncertainty of ER visits (ca. 10%) is higher than other series. We suggest MHW consider upper bound rather than lower bound to avoid doctor shortage.

Limitation:

Although this study offers data-driven findings as reference, two limitations should be pointed out. First, our business goal is to forecast patient visits for year 2017/ 2018, but due to data lag, we have to forecast three/ four-year ahead instead. However, we found forecasting result using one-year ahead performs much better in three/four-year ahead. We recommend MHW release annual data as soon as possible for better performance. Additionally, our results are based on annual reports, which aggregated data from different data from monthly data all over Taiwan. Therefore, we suggest MHW provide higher frequency data such as area-specific and month-specific data for more detailed forecasts.

Appendix 1: Best methods and their performances

Series	GM	GS	OB	ER	ER2	PED
Best Method	Holt's (ANN)	Ensemble (Quadratic regression, Holt's AAN, Regression with external data)	Naive (Holt's ANN, Naive, Regression with external data)	Holt's (ANN)	Ensemble	Holt's (ANN)
3-year ahead best method RMSE	74000	58065	38144	23220	171383	34587
3-year ahead Naive MAPE	1.33%	3.72%	1.05%	2.32%		3.90%
3-year ahead best method MAPE	0.73%	0.75%	1.05%	3.08%	2.37%	0.61%