**Statistics and Application** 统计学与应用, 2020, 9(6), 1027-1039
Published Online December 2020 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2020.96108

Hans汉斯

# 上海市浦东新区高血压发病因素分析

**赵子杰，林沛辰，兰　雪，白晓东**

大连民族大学理学院，辽宁　大连
Email: baixd518@126.com

收稿日期：2020年11月21日；录用日期：2020年12月20日；发布日期：2020年12月31日

## 摘　要

**为研究浦东新区高血压发病的因素，本文收集了2010年1月至2014年12月共60期上海市浦东新区高血压发病就诊数据，并使用ARMA模型，季节模型和Holt-Winters拟合分析并预测得到了12期数据，与实际数据对比，得出高血压发病就诊数具有明显的季节周期特征的结论，故我们认为季节变化对于高血压发病情况具有显著影响。**

## 关键词

**时间序列，季节模型，高血压，Holt-Winters指数平滑**

# Analysis of Factors of Hypertension in Shanghai Pudong New Area

**Zijie Zhao, Peichen Lin, Xue Lan, Xiaodong Bai**

Dalian Minzu University, Dalian Liaoning
Email: baixd518@126.com

Received: Nov. 21st, 2020; accepted: Dec. 20th, 2020; published: Dec. 31st, 2020

## Abstract

**In order to study the factors of the incidence of hypertension in Pudong New Area, this paper collected 60 periods of data on the incidence of hypertension in Pudong New Area, Shanghai from January 2010 to December 2014, and used ARMA model, seasonal model and Holt-Winters fitting analysis. The 12-period data is predicted and compared with the actual data. It is concluded that the number of patients with hypertension has obvious seasonal cycle characteristics. Therefore, we believe that seasonal changes have a significant impact on the incidence of hypertension.**

文章引用: 赵子杰, 林沛辰, 兰雪, 白晓东. 上海市浦东新区高血压发病因素分析[J]. 统计学与应用, 2020, 9(6): 1027-1039. DOI: 10.12677/sa.2020.96108

**Keywords**

**Time Series, Season Model, Hypertension, Holt-Winters Exponential Smoothing**



## 1. 前言

高血压作为一种慢性疾病，其原发性发病机理尚在探寻，近年来高血压患病群体趋于青年化，而与此同时人们的健康意识也在逐渐提升，对于高血压等慢性疾病的探索需求与日俱增。因此，研究高血压发病规律对于统筹医疗资源，提高卫生服务质量，探究发病机理具有重要意义。本文中，我们将从宏观的时间跨度来分析高血压可能存在的发病因素。

## 2. 问题分析

对浦东新区 2010 年 1 月至 2014 年 12 月的高血压发病就诊数据建立时间序列，若时序图具有明显趋势，则采用差分或温斯特指数平滑法提取趋势[1]，建立适当的时间序列模型，通过调整参数，显著性检验得到最优模型，并对未来 12 期高血压发病就诊数据进行预测。

## 3. 名词解释

(1) Holt 线性指数平滑：考虑时间远近对 t 时期趋势估计值的影响，对含有线性趋势的数据采用的一种加权平均的办法。

(2) Holt-Winters 指数平滑：在 Holt 线性指数平滑的基础上考虑季节变动的影响，一般来讲，对于趋势和季节的加法模型，Holt-Winters 指数平滑法[2]的公式如下

$$\begin{cases} \alpha_t = \alpha\left(x_t - s_{t-\pi}\right) + \left(1-\alpha\right)\left(\alpha_{t-1} + b_{t-1}\right); \\ b_t = \beta\left(\alpha_t - \alpha_{t-\pi}\right) + \left(1-\beta\right)b_{t-1}; \\ s_t = \gamma\left(x_t - \alpha_t\right) + \left(1-\gamma\right)s_{t-\pi} \end{cases}$$

其中，$\alpha_t$ 为该序列水平部分；$b_t$ 为该序列的趋势部分；$s_t$ 为该序列季节部分；$\pi$ 为一个季节的周期长度；$\alpha$，$\beta$，$\gamma$ 为平滑系数，介于 0 到 1 之间。

(3) 季节模型[2]：$y_t = \nabla^d \nabla_s^D x_t$，若 $\{y_t\}$ 满足季节周期为 $s$ 的 ARMA $(p,q) \times (P,Q)$模型，则称 $\{x_t\}$ 为季节周期为 $s$，非季节阶数为 $p$，$d$，$q$，季节阶数为 $P$，$D$，$Q$ 的乘积季节求和自回归移动平均模型，记作 Sarma $(p,d,q) \times (P,D,Q)$ s。

## 4. 模型假设

(1) 假设 2010 年 1 月至 2014 年 12 月之间，浦东新区没有发生导致高血压发病的重大卫生事件。

(2) 假设 2010 年 1 月至 2014 年 12 月之间，浦东新区医疗水平变化不足以导致高血压发病确诊人数出现显著变化。

(3) 假设采集的数据口径相同，无太大误差。

## 5. 时间序列模型建立

### 5.1. 序列平稳性检验

若一个随机过程的统计特性不随时间的推移而变化，则称它为平稳随机过程。平稳性是一些时间序列具有的统计特征，对数据进行平稳性检验是分析时间序列的关键步骤。平稳时间序列有两种定义，根据限制条件的严格程度，分为严平稳时间序列和宽平稳时间序列。对序列的平稳性有两种检验方法，一种是根据时序图和自相关图显示的特征做出判断的图检验方法；一种是构造检验统计量进行假设检验的方法。

#### 时序图检验

从图 1 中我们可以看出，该序列具有明显的增长趋势，为非平稳序列。根据图 2，图 3，自相关拖尾，偏自相关一阶截尾。随着人们生活条件的改善，越来越多的人饮食结构不健康，生活习惯不规律，亚健康成为当代人的常态，高血压等慢性疾病患病率持续增加，时序图反映的趋势是符合现实情况的。

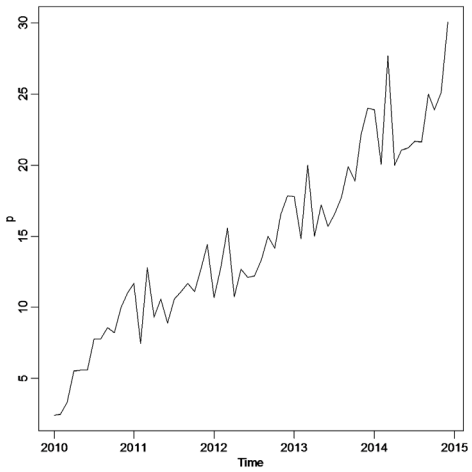


**Figure 1.** Time series of hypertension incidence in Pudong New Area from 2010 to 2014
**图 1.** 浦东新区 2010 至 2014 年高血压发病数时序图

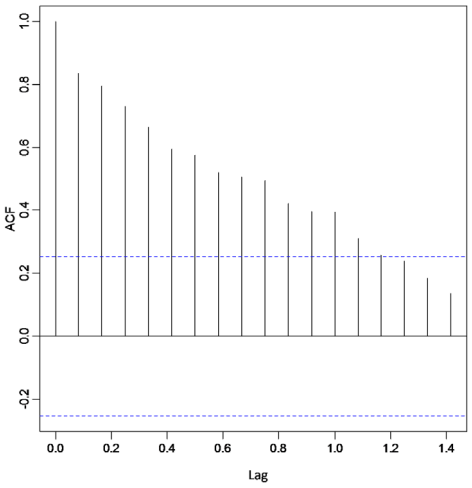


**Figure 2.** Autocorrelation graph
**图 2.** 自相关图

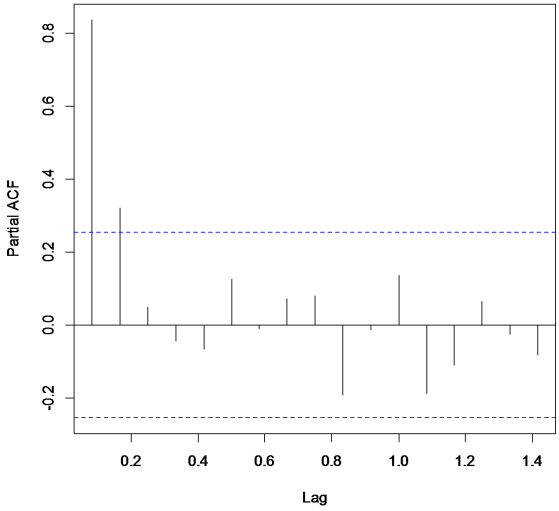


**Figure 3.** Partial autocorrelation graph
**图 3.** 偏自相关图

## 5.2. 不同模型的比较

### 5.2.1. 线性拟合[2]

原序列表现出明显线性趋势，首先对其线性拟合，图 4 中红线为拟合的线性趋势部分，去趋势后残差白噪声检验结果 p 值等于 0.3621，说明残差序列为白噪声，信息提取比较完全，同时得到趋势部分 $x_t$ = 0.3479t + 3.9784，用原序列减去趋势部分得到随机部分 $\varepsilon$ 如图 5，随机部分 $\varepsilon$ 仍然为非平稳序列，一阶差分后，使用 arima 模型进行识别，识别结果为 ARMA (0,1,2)，因此序列可表示为 $x_t$ = 0.3479t + 3.9784 + $\varepsilon$，预测结果如表 1，图 6。

### 5.2.2. 季节模型

在提取线性趋势后对随机部分建立 ARMA 模型的基础上，我们发现，序列具有明显的季节周期性特征，于是我们考虑使用季节模型对序列进行建模[3]。

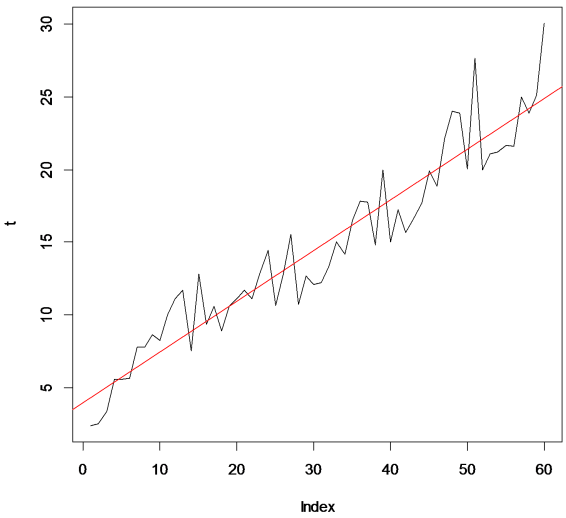


**Figure 4.** Linear trend extraction
**图 4.** 线性趋势提取

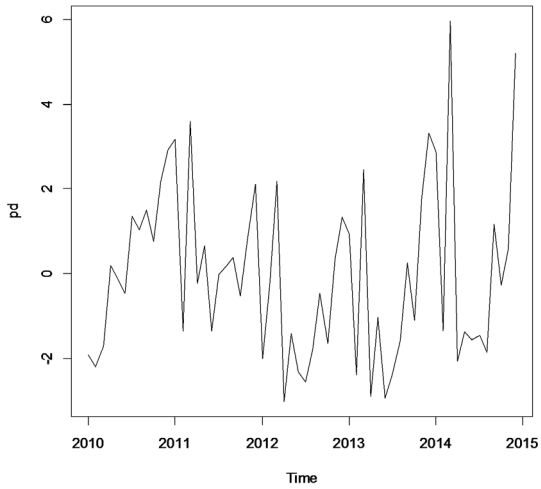


**Figure 5.** Random partial sequence diagram
**图 5.** 随机部分时序图

**Table 1.** Comparison between ARMA model prediction results and actual values
**表 1.** ARMA 模型预测结果与实际值对比

| | 一月 | 二月 | 三月 | 四月 | 五月 | 六月 |
|---|---|---|---|---|---|---|
| 预测值 | (24.05, 30) | (22.23, 28.39) | (26.39, 32.83) | (22.45, 29.17) | (23.85, 30.83) | (23.09, 30.33) |
| 实际值 | 24.87 | 22.39 | 29.69 | 20.76 | 21.77 | 21.94 |

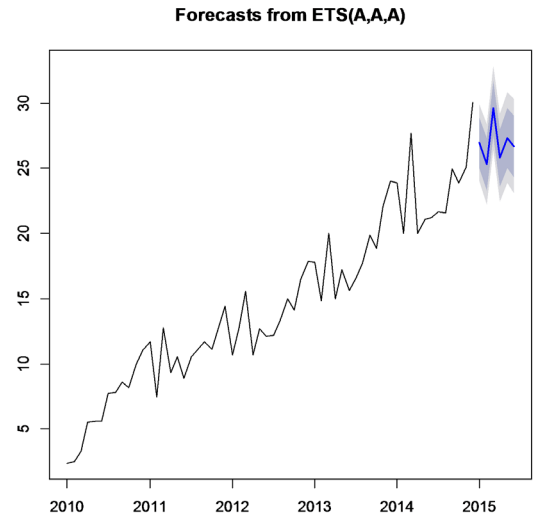


**Figure 6.** ARMA model prediction results
**图 6.** ARMA 模型预测结果

图 7 为 2010 到 2014 年间高血压就诊数的叠加曲线图，图中黑色为 2010 年数据，红色为 2011，蓝色为 2012，绿色为 2013，紫色为 2014，折线图直观表现出，所有年度曲线都位于之前年度的上方，即每一年的高血压患病数都在稳定增加，所有年份一年之中就诊数量在一月到三月之间达到一个高峰，随后大幅度下降，在十月之前保持缓慢的持续增长，十月之后开始大幅度增长，由此可以判断浦东新区高血压就诊人数具有明显季节周期特征，所以我们尝试使用季节模型对序列进行拟合。

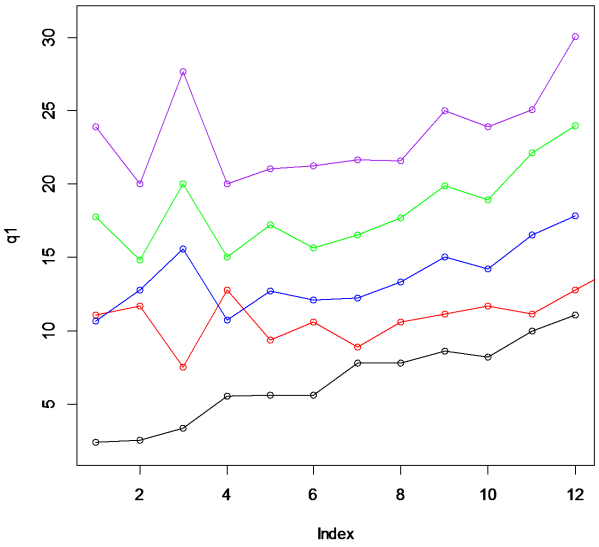


**Figure 7.** Superposition curve of hypertension incidence in 2010-2014
**图 7.** 2010~2014 年度高血压发病人数叠加曲线

一阶差分如图 8，图 9，可以看出，一阶差分序列仍具有周期特征，自相关函数一阶拖尾，偏自相关二阶拖尾。因为数据为一年中十二个月月份数据，我们考虑采用一阶十二次差分。

对原序列进行一阶十二次差分，得到序列已无明显周期特征白噪声检验LB统计量p值等于1.438e−5，小于0.05，说明该序列为非白噪声序列。

由图 10，图 11，一阶十二次差分序列自相关一阶拖尾，偏自相关一阶截尾，识别为自回归模型 arima (1,1,0) 12。

建立季节模型 Sarma (0,1,2) × (1,1,0) 12，其中原序列为 arima (0,1,2)，差分序列为 arima (1,1,0)，我们以十二个月为期，$s=12$，建立季节模型 $sarima\ (p,0,1,2,P=1,D=1,Q=0,S=12)$，并预测十二期值。

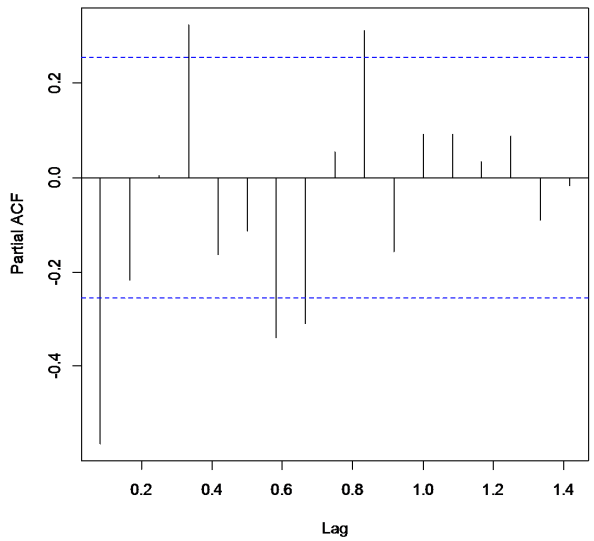


**Figure 8.** Autocorrelation graph of first order difference sequence
**图 8.** 一阶差分序列自相关图

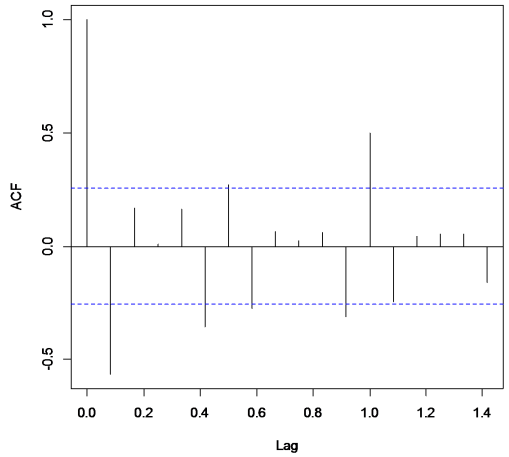


**Figure 9.** Partial autocorrelation graph of first order difference sequence
**图 9.** 一阶差分序列偏自相关图

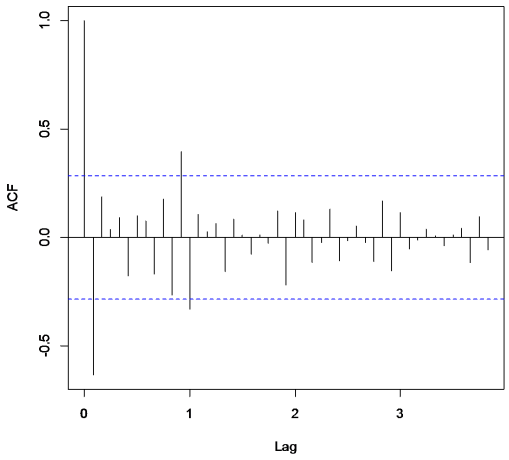


**Figure 10.** Autocorrelation graph of first order twelve order
**图 10.** 一阶十二次差分序列自相关图

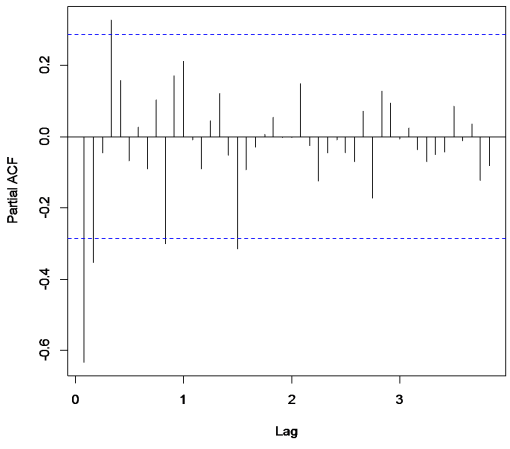


**Figure 11.** Partial autocorrelation graph of first order twelve difference sequence order difference sequence
**图 11.** 一阶十二次差分序列偏自相关图

得到季节自回归系数 sar1 = −0.338，移动平均系数 ma1 = −0.8696，ma2 = 0.5952，标准差为 1.923，AIC = 174.84，根据图 12，残差 acf 始终在二倍标准差范围内，残差白噪声检验也始终大于 p 值 0.05 所以季节模型对序列拟合程度较高，误差较小。季节模型预测结果如图 13。

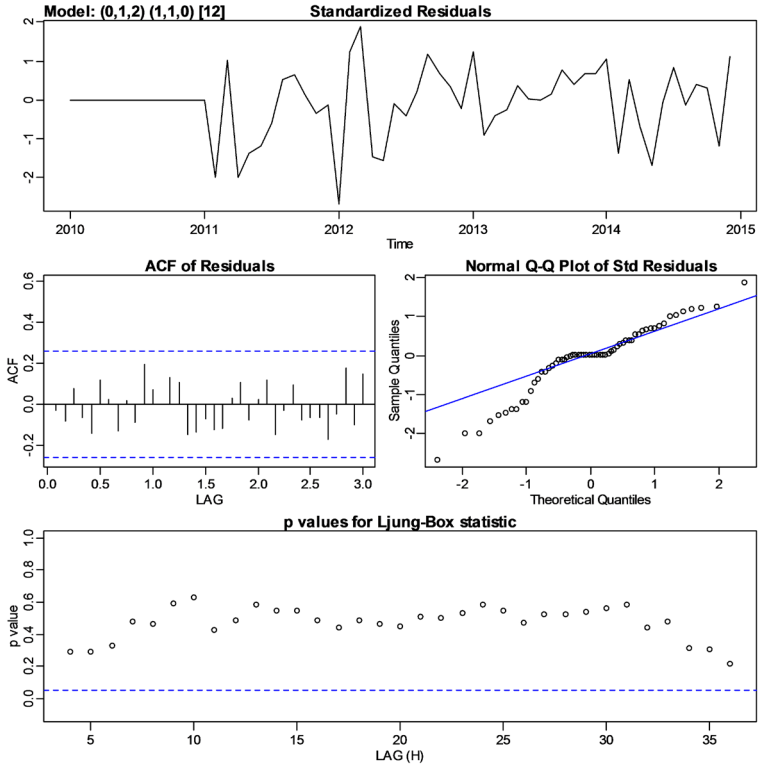


**Figure 12.** Seasonal model inspection chart
**图 12.** 季节模型检验图

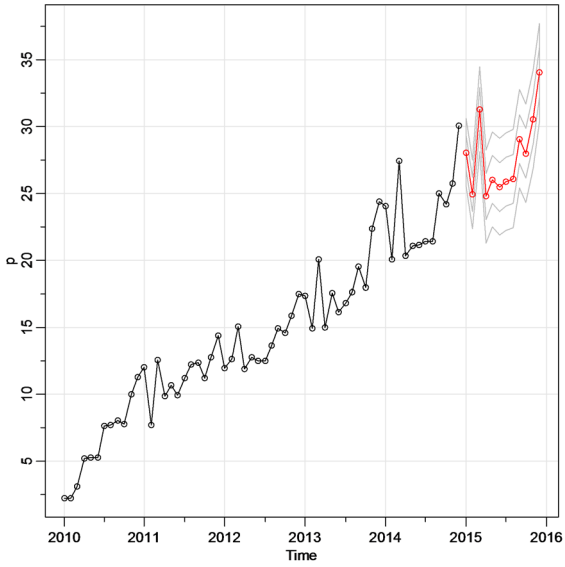


**Figure 13.** Prediction of twelve period series value of seasonal model
**图 13.** 季节模型十二期序列值预测

### 5.2.3. Holt-Winters 指数平滑[4]

Holt-Winters 指数平滑相对于简单指数平滑和 Holt 线性指数平滑考虑了季节因素的影响，相比简单指数平滑和 Holt 线性平滑更适用于高血压就诊人数序列的拟合。拟合曲线如图 14，图 14 中黑色曲线为原序列，红色曲线为指数平滑后消除趋势的拟合曲线，二者重合度较高，拟合效果较好。并在此基础上做 12 期预测。图 15 中蓝色折线为 12 期预测值，深灰色部分为百分之九十五的置信区间，浅灰色部分为百分之八十的置信区间，将 Holt-Winters 模型与季节模型对比，可以看出，Holt-Winters 模型与季节乘法模型预测结果比较相似，所谓我们可以认为这两种模型的拟合以及预测结果具有较高的置信度。

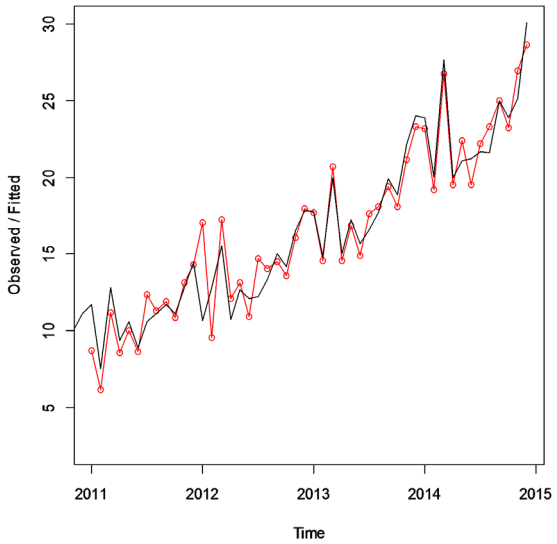


**Figure 14.** Holt winters exponential smoothing results
**图 14.** Holt-Winters 指数平滑结果

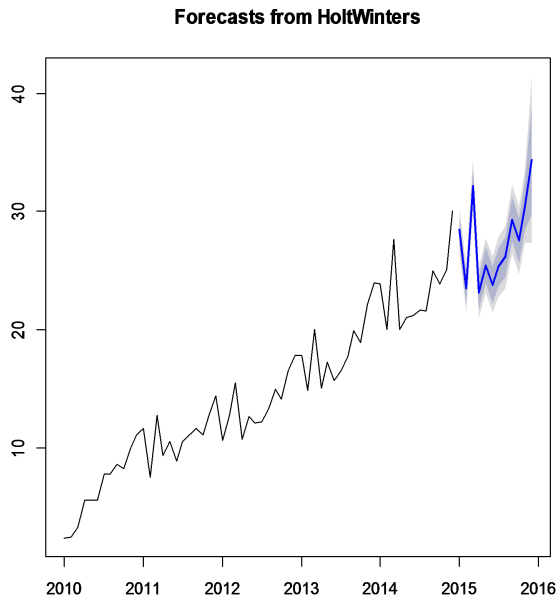


**Figure 15.** Holt winters index smoothing prediction value
**图 15.** Holt-Winters 指数平滑预测值

### 5.3. 预测分析

我们将 60 期数据以年为单位分割发现，在 12 个月的周期内，一到三月间往往是一个周期中数据峰值的位置，随后有一个明显的回落，但是根据图 16，将五个周期的序列叠加对比，一月和三月都处于相对接近峰值的位置，但二月的数据却大多出现反常的回落，主要表现为前三个月往往呈“V”字型。我们认为这与中国传统春节有关，农历新年往往在公历一，二月份之间，这期间人们的饮食往往偏向高油高盐，饮酒，吸烟行为明显大量增加，且春节期间人们常常通宵娱乐，此类行为少则持续三五天，多则半个月以上，我们认为，春节期间的大量不健康的饮食和生活行为导致了前三个月的就诊数较多，随后快速回落。其中数据在二月附近出现反常，我们认为是春节期间就诊人数减少，我们收集的医院社区有记录的就诊人数，本身就不能完全代表高血压患病真实情况，而春节期间的就诊记录数据，由于假期更无法准确代表真实情况。因此得出结论，反常回落处的数据应该与前后一个月的数据接近，总体呈增长趋势。因此，春节期间高血压患病增多是我国的特殊国情。

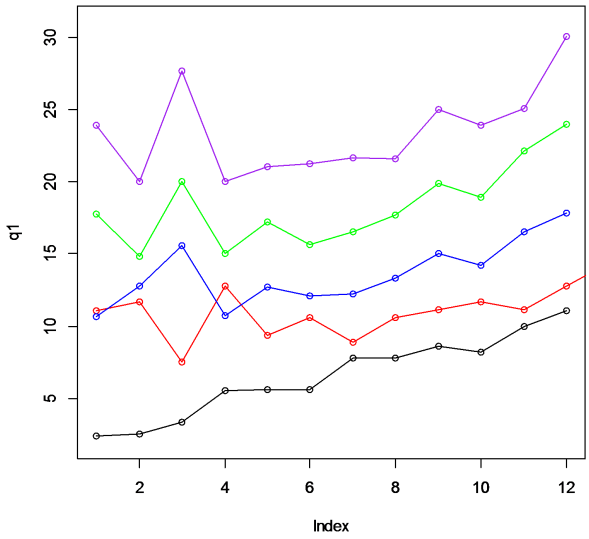


**Figure 16.** Superposition curve of hypertension incidence in 2010-2014
**图 16.** 2010~2014 年度高血压发病人数叠加曲线

综上所述，结合提取趋势后的 ARMA 模型，季节模型和 Holt-Winters 指数平滑的预测，我们认为浦东地区的高血压患病人数在未来的一段时间仍将呈线性增长趋势，周期内波动情况同前五个周期规律一致。

## 6. 模型评价

### 6.1. 模型优点

(1) 采用三种模型对同一序列进行分析，相互对照补充，使结论更加完整准确。

(2) 反馈得到春节期间医疗卫生系统信息记录存在的缺陷，对卫生系统信息工作的改进有所帮助。

(3) 分析得到高血压发病周期性变化规律，对预防高血压，以及统筹调配医疗资源，研究高血压的病理等具有现实的参考价值。

### 6.2. 模型缺点

我们将预测结果同 2015 年真实数据进行对比，发现预测值整体偏高，只有部分数据落在置信度百分

之九十五的预测区间，两种方法预测结果与实际值对比如表 2，说明模型存在部分缺陷，这可能是由于数据不完整，在处理数据时插值的问题，也表明在此基础上模型还具有改进的空间。

**Table 2.** Comparison of seasonal model and Holt winters smoothing prediction results
**表 2.** 季节模型与 Holt-Winters 平滑预测结果对比

| | 季节模型 | Holt-Winters | 实际值 |
|---|---|---|---|
| 2015.1 | (26.14, 28.9) | (26.63, 30.35) | 24.87 |
| 2015.2 | (23.67, 26.45) | (21.52, 25.42) | 22.39 |
| 2015.3 | (30.03, 33.49) | (29.92, 34.41) | 29.69 |
| 2015.4 | (23.03, 26.99) | (21.02, 25.27) | 20.76 |
| 2015.5 | (24.25, 28.67) | (23.15, 27.76) | 21.77 |
| 2015.6 | (23.63, 28.45) | (21.45, 26.13) | 21.94 |

## 7. 总结

本文收集了 2010 至 2014 年间浦东新区的高血压发病就诊月度数据共 60 期，分别采用提取趋势后的 ARMA 模型，季节模型，Holt-Winters 指数平滑进行建模，在此基础上对序列进行分析，并预测 12 期序列值，该模型很好地揭示了高血压发病的季节性规律，以及总体趋势，预测结果具有较高的置信度。

## 基金项目

国家级大创项目资助(项目编号：202012026040)。

## 参考文献

[1] 王晓丽, 施天行, 杨思睿, 陈潇雨. 温特斯指数平滑法在高血压就诊人次预测中的应用[J]. 中国卫生信息管理杂志, 2016(5): 524-52.

[2] 白晓东. 应用时间序列分析[M]. 北京: 清华大学出版社, 2017.

[3] 何书元. 应用时间序列分析[M]. 北京: 北京大学出版社, 2003.

[4] 王晓丽, 陈浩, 施天行, 杨思睿. 浦东新区高血压就诊人数预测模型的研究[J]. 中国数字医学, 2016-11-02.

# 附录

## 附录 1. 原始数据

| | 2010 年 | 2011 年 | 2012 年 | 2013 年 | 2014 年 |
|---|---|---|---|---|---|
| 一月 | 1.73 | 11.72 | 10.83 | 17.67 | 23.99 |
| 二月 | 2.02 | 7.5 | 12.97 | 14.83 | 20.01 |
| 三月 | 3.01 | 12.93 | 15.67 | 20.06 | 27.83 |
| 四月 | 5.67 | 7.27 | 11.09 | 15.02 | 20.16 |
| 五月 | 5.83 | 10.66 | 12.69 | 17.17 | 21.27 |
| 六月 | 5.93 | 8.93 | 12.00 | 16.02 | 21.33 |
| 七月 | 7.54 | 10.67 | 12.17 | 16.83 | 22.02 |
| 八月 | 7.67 | 11.03 | 13.33 | 17.63 | 21.67 |
| 九月 | 9.06 | 11.67 | 14.67 | 19.81 | 24.93 |
| 十月 | 8.07 | 10.84 | 14.06 | 18.67 | 24.47 |
| 十一月 | 9.98 | 12.33 | 16.67 | 21.43 | 27.33 |
| 十二月 | 10.67 | 14.08 | 17.99 | 24.06 | 30.06 |

## 附录 2. 程序

```
t=read.csv("E:/pudong.csv",header=F)
t
length(t)
p=ts(t,start=c(2010,1),frequency=12);p;plot(p)

Box.test(p);acf(p);pacf(p)
k=diff(p);k;plot(k)
Box.test(k)
acf(k,lag=60)
pacf(k,lag=60)
auto.arima(p)
h=diff(p,12);h;plot(h)
Box.test(h)
w=diff(h);w;plot(w);Box.test(w)
acf(w,lag=60)
pacf(w,lag=60)
w1=auto.arima(w);w1
for(i in 1:2)print(Box.test(w1$residual,lag=6*i))
c=Arima(p,order=c(1,1,0));c
d=forecast(c,h=6);d
plot(d)
```

```
sarima.for(p,6,0,1,2,P=1,D=1,Q=0,S=12)
ss=sarima(p,0,1,2,P=1,D=1,Q=0,S=12,details=T);ss
f=HoltWinters(p,seasonal="multiplicative");f
plot(f,type="o")
ff=forecast(f,h=6);ff
plot(ff)
lines(p,lty=2,col="red")
x=read.csv("E:/shanghai1.csv",header=F);x
y=ts(x,start=c(2010,1),frequency=12);y;plot(y)
x1=x[,1];x1
x2=x[,2];x2
x3=x[,3];x3
p1=ts(x1,start=c(2010,1),frequency=12);p1;plot(p1);w=diff(p2);w;plot(w)
p2=ts(x2,start=c(2010,1),frequency=12);p2;plot(p2)
p3=ts(x3,start=c(2010,1),frequency=12);p3;plot(p3)
q1=t[1:12];q1
q2=t[12:24];q2
q3=t[25:36];q3
q4=t[37:48];q4
q5=t[49:60];q5
qq1=ts(q1,start=c(2010,1),frequency=12);qq1
qq2=ts(q2,start=c(2011,1),frequency=12);qq2
qq3=ts(q3,start=c(2012,1),frequency=12);qq3
qq4=ts(q4,start=c(2013,1),frequency=12);qq4
qq5=ts(q5,start=c(2014,1),frequency=12);qq5
plot(q1,type='o',ylim=c(2,31));lines(q2,type='o',col="red");
lines(q3,type='o',col="blue");lines(q4,type='o',col="green");
lines(q5,type='o',col="purple")
c=cor(k,w);c
```