Statistics and Application 统计学与应用, 2024, 13(3), 1010-1017
Published Online June 2024 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2024.133102

Hans汉斯

基于时间序列干预分析模型的我国铁路客运趋势预测研究

詹丹娇

华南师范大学数学科学学院，广东 广州

收稿日期：2024年6月2日；录用日期：2024年6月22日；发布日期：2024年6月30日

摘　要

为评估突发事件对我国铁路客运量的趋势影响，本文提出在识别时间序列离群值和最优ARIMA模型基础上，对干预影响序列进行基于脉冲函数或阶梯函数的虚拟变量针对性设计，确定干预过程的传递函数，与最优ARIMA模型结合，建立干预预测模型。文章选取2015年1月至2021年12月的全国铁路客运量月度序列数据建立模型，通过研究发现：疫情的影响导致原本具有周期性、增长趋势的客流序列产生离群值，并且干预发生后对滞后期仍有影响；干预分析模型拟合程度较好，预测符合实际走向，模型的均方根误差(RMAE)为2856.15949，总体上看精确度较高。

关键词

铁路客运量，突发事件，ARIMA，干预分析模型

A Study on Trend Prediction of China's Railway Passenger Transport Based on Time Series Intervention Analysis Model

Danjiao Zhan

College of Mathematical Sciences, South China Normal University, Guangzhou Guangdong

Received: Jun. 2nd, 2024; accepted: Jun. 22nd, 2024; published: Jun. 30th, 2024

Abstract

To assess the impact of emergency events on the trend of China's railway passenger volume, this paper proposes a targeted design of dummy variables for the intervention impact sequence based

文章引用: 詹丹娇. 基于时间序列干预分析模型的我国铁路客运趋势预测研究[J]. 统计学与应用, 2024, 13(3): 1010-1017. DOI: 10.12677/sa.2024.133102

on impulse functions or step functions, after identifying outliers in the time series and selecting the optimal ARIMA model. The study determines the transfer function of the intervention process and combines it with the optimal ARIMA model to establish an intervention prediction model. Using monthly series data of national railway passenger volume from January 2015 to December 2021, the research finds that the impact of the pandemic has resulted in outliers in the originally periodic and growing passenger flow sequence, and the intervention still has an effect on subsequent periods. The intervention analysis model demonstrates a good degree of fit, with predictions aligning with actual trends. The root mean square error (RMAE) of the model is 2856.15949, indicating a high overall accuracy.

Keywords

Railway Passenger Volume, Unexpected Events, ARIMA, Intervention Analysis Model

1. 引言

近年来，铁路建设不断加快，逐渐成为人们出行的最主要交通方式之一。在 2020 年之前，我国铁路客运量呈现逐年增长的趋势[1]。另外，受到大众假期集中出行的影响，铁路客运量具有明显的季节性波动。然而，2020 年初突发“新冠肺炎疫情”(以下简称新冠疫情)，对人们的生活和社会经济造成了不同程度的冲击，导致我国铁路客运量急剧下降。铁路客运量是重要的交通运输指标，能够反映突发事件对人民生活的影响。因此，研究其冲击程度和发展趋势对于预防和应对未来可能的影响至关重要。然而，突发事件如新冠疫情给铁路客运量趋势带来的影响使得预测工作更加复杂。因此，建立合适的时间序列干预预测模型分析突发事件对铁路客运量的冲击趋势至关重要。

目前，针对铁路客运量时间序列的预测研究，主要运用机器学习[2] [3]、BP 神经网络[4] [5]、灰色模型[6]、SARIMA [7] [8]模型等。夏国恩等[3]等提出了一种改进 SVR 的铁路客运量时间序列预测方法，用于预测 1980~1998 年铁路客运量；王卓等[4]采用改进的 BP 神经网络对铁路客运量时间序列进行预测，并与标准的 BP 神经网络预测结果进行对比；王慧晶[6]通过建立株洲站旅客发送量的灰色预测模型，展示了灰色模型在铁路客运量预测方面的良好精度；郝军章等[7]利用我国 2007 年至 2014 年的铁路客运量数据构建了 SARIMA 模型，对 2015 年 1 至 3 月的客运量进行了预测。可见，现有客运量预测研究侧重于对宏观的年度增减趋势进行预测，而对中观层面的月度客运量变化规律和特点的研究尚不够深入。此外，大部分预测研究大多集中在时间序列分析模型上，对于突发事件的干预因素的研究相对较少。

时间序列经常会受到特殊事件及态势的影响，这类外部事件称为干预。[9]新冠疫情等突发事件对铁路客运量的影响实际上是一个突发事件的干预问题。干预事件的发生通常会使客运量数据产生离群值，其影响过程不能忽视。为增强干预事件的针对性，早期学者们先计算真实值与基于序列构建的 ARIMA 模型拟合值之间的差异，将其作为纯干预序列。随后，估计并构建纯干预效果模型，将 ARIMA 模型和干预效果模型结合，构成时间序列的 ARIMA 干预模型，以提高对干预事件的识别能力[10] [11] [12]。

基于以上研究现状，本文在前人研究基础上主要从以下方面展开研究：考虑铁路客流的短期波动变化，以铁路月度客流量时间序列为研究对象。其次，将 ARIMA 的乘积季节模型与干预分析功能相结合，构建时间序列干预预测模型，优化前人在铁路客运量干预预测中的不足。

2. 数据与方法

2.1. 数据来源及趋势分析

本文数据来源于国家统计局网站，选取 2015 年 1 月~2021 年 12 月$\{x_t, t=0,1,2,\cdots,83\}$的全国铁路客运量月度数据为原始数据，其中 2015 年 1 月为起始时刻($t=0$)，数据无缺失值。

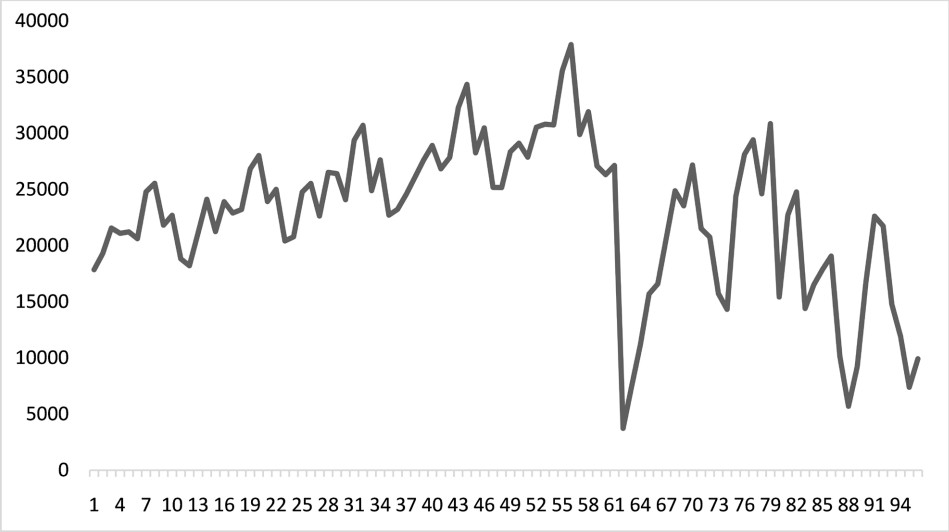

Figure 1. Time series chart of railway passenger volume from January 2015 to December 2021
图 1. 2015 年 1 月~2021 年 12 月铁路客运量时序图

通过图 1 可知，在 2015 年至 2019 年期间，中国铁路运输业客运量受气候条件、工农业生产活动、居民节假日等风俗习惯以及国民经济发展等因素的周期性影响，呈现出明显的季节性波动，周期为 12 个月，且呈现出明显的递增趋势。然而，在 2020 年 2 月，铁路客运量突然骤减至 3723 万人，相较上个月的 27126 万人，减少了 23403 万人。这一急剧下降的趋势可以推断为新冠疫情的蔓延所导致的结果。政府为了遏制疫情的蔓延，采取了一系列紧急措施，包括限制人员流动和加强卫生防护措施，这直接导致了客运量的急剧下降。

从 2020 年 3 月开始，铁路客运量呈现出逐渐递增的趋势，然而仍未恢复到疫情前的水平。表明疫情对铁路客运量产生了长期的影响。随着疫情逐渐得到控制，人们对于出行的需求可能会逐渐恢复，但疫情防控措施对于客运量的影响可能会持续一段时间。

2.2. 干预分析模型理论简介

通过数据的趋势分析可知包含季节性特征，需建立 ARIMA 模型的季节乘积模型 ARIMA (p, d, q) (P, D, Q)^s。该模型是在 ARIMA 的基础上加入了季节性部分。(P, D, Q)^s 表示季节部分，其中 s 表示季节性频率。

干预分析模型是在对事件建立时间序列模型的基础上，将干预事件纳入考虑范围，对模型进行相应调整和整合，以建立更精确、完善的预测模型。

干预变量有两种表示形式：

(1) 持续性干预变量，表示时间序列在 T 时刻受到干预，干预发生后序列受到持续影响，此时干预变量用阶跃函数表示：

$$S_t^T = \begin{cases} 0, \text{干预事件发生之前} \\ 1, \text{干预事件发生之后} \end{cases} \tag{1}$$

(2) 短暂性的干预变量，表示时间序列在某一时刻受到干预，干预与仅在该时刻产生影响，此时干预变量用脉冲函数表示：

$$P_t^T = \begin{cases} 0, \text{其它事件} \\ 1, \text{干预事件发生时} \end{cases} \tag{2}$$

在研究中必须选择恰当的干预形式，以反映序列中紧急变量的波动情况干预事件的形式。

根据前文的分析可知，2020 年 1 月我国铁路客运量的数值为 27126 万人，而 2 月为 3723 万人，由此可知疫情冲击从 2020 年 2 月开始的，因此本文按照外部事件的发生分将数据为两个时期：第一个时期为 2015 年 1 月至 2020 年 1 月 $\{x_t, t = 0,1,2,\cdots,61\}$，该时期的铁路客运量未受到疫情影响；第二个时期为 2020 年 2 月至 2021 年 12 月 $\{x_t, t = 62,63,64,\cdots,83\}$。由于疫情防控对铁路客运量的影响是在 2020 年 2 月份突然发生，并且该影响会产生持续作用，故设定干预变量 $I_t^T = S_t^T$，干预影响为：

$$Z_t = \frac{\omega(B)}{\delta(B)} S_t^T + \varepsilon_t, S_t^T = \begin{cases} 0, t < T = 62 \\ 1, t \geq T = 62 \end{cases} \tag{3}$$

其中$\{\varepsilon_t\}$为状态零均值白噪音。当$\delta(B)=1$时，退化为 $Z_t = \omega(B) S_t^T$，意味着干预事件的影响突然开始，定性分析不能判断显示模型的准确形式[13]。

干预分析模型建模的具体步骤如下[11]：

(1) 首先，利用未发生干预事件前的序列数据进行建模，创建单变量时间序列模型。通过该模型对一定时间段内的序列进行外推预测，从而获得在没有干预事件发生时的预测数值。

(2) 然后，对干预的影响进行量化分析。通过实际观测值与预测值之间的差异来确定干预事件对原始序列的影响程度，进而推导出干预模型的相关参数。

(3) 接着，计算并获得排除了干预影响后的数据，建立基于这些数据的单变量时间序列模型，以便进一步分析和预测序列的变化趋势。

(4) 综合以上步骤，构建完整的干预分析模型。

3. 铁路客运量干预分析模型的构建

3.1. 第一个时期序列模型构建

经过自相关函数和偏自相关函数分析，在通过残差白噪音的基础上，采用最小信息准则(AIC)对模型的阶数进行判定。最后综合考虑预测效果及参数显著性，建立最优模型为 $\text{ARIMA}((1,2),0)\times(0,1,1)^{12}$ 模型的参数估计如下：

Table 1. $\text{ARIMA}((1,2),0)\times(0,1,1)^{12}$ model parameter testing

表 1. $\text{ARIMA}((1,2),0)\times(0,1,1)^{12}$ 模型参数检验

参数	估计	标准误差
MA1,1	0.43895^{**}	0.17137
AR1,1	-0.63232^{***}	0.14063
AR1,2	-0.43735^{***}	0.14709

注：***，**，*分别表示通过 1%，5%，10%显著性检验

结果表明，以上参数估计值具有统计意义。模型 $\mathrm{ARMA}(3,1,0)\times(0,1,1)^{12}$ 为：

$$\nabla\nabla_{12}x_t=\frac{1-0.43895B^{12}}{1+0.63232B+0.43735B^2}\varepsilon_t \tag{4}$$

接着利用 $\mathrm{ARMA}(3,1,0)\times(0,1,1)^{12}$ 模型预测未来 23 期，即 2020 年 2 月至 2021 年 12 月在无干预影响下我国铁路客运量的预测值。

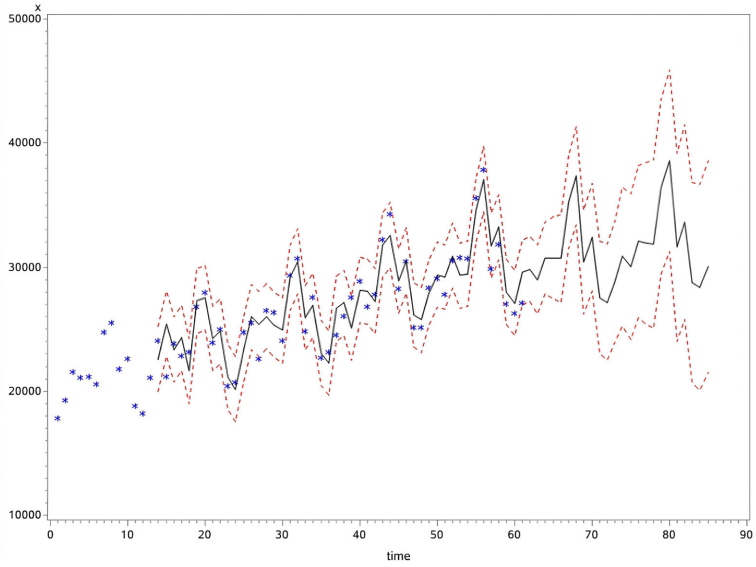

Figure 2. $\mathrm{ARIMA}((1,2),0)\times(0,1,1)^{12}$ model prediction and fitting chart

图 2. $\mathrm{ARIMA}((1,2),0)\times(0,1,1)^{12}$ 模型预测拟合图

图 2 表示的是 2015 年 1 月至 2020 年 1 月的实际值与预测值，其中，虚线是置信水平 95%的预测置信区间，中间的实线是预测值。*表示真实值

3.2. 分离干预影响

本研究已运用建立的第一个时期的 ARIMA 季节乘积模型外推预测了 2020 年 2 月至 2021 年 12 月 $\{x_t,t=62,63,64,\cdots,83\}$ 没有干预作用时的人数预测值，用实际值减去预测值，得到的差值解释疫情产生的影响程度[14]，记为 Z_t，具体数值如下表所示

Table 2. Impact degree table of the pandemic
表 2. 疫情产生的影响程度表

t	Z_t	t	Z_t	t	Z_t
		69	−5258.4242	77	−7246.5107
62	−21487.1357	70	−6056.7759	78	−5606.9468
63	−19561.3673	71	−6432.4511	79	−23150.6459
64	−15087.9374	72	−13130.9031	80	−8894.7581
65	−14102.51	73	−16536.7957	81	−8871.1528
66	−14475.1144	74	−5693.8737	82	−14360.2018

续表

67	−12509.1397	75	−3960.8446	83	−11871.9902
68	−6882.182	76	−2498.8559		

Table 3. Parameter significance test for the $\mathrm{ARIMA}(8,2,1)\times(1,1,0)^{12}$ model

表 3. $\mathrm{ARIMA}(8,2,1)\times(1,1,0)^{12}$ 模型的参数显著性检验

t	Z_t	t	Z_t	t	Z_t
61	−26119.6386	69	−5258.4242	77	−7246.5107
62	−21487.1357	70	−6056.7759	78	−5606.9468
63	−19561.3673	71	−6432.4511	79	−23150.6459
64	−15087.9374	72	−13130.9031	80	−8894.7581
65	−14102.51	73	−16536.7957	81	−8871.1528
66	−14475.1144	74	−5693.8737	82	−14360.2018
67	−12509.1397	75	−3960.8446	83	−11871.9902
68	−6882.182	76	−2498.8559		

由于 Z_t 序列无趋势性及周期性，且白噪音检验表明 Z_t 序列相关性不显著($p > 0.1$)，任何 ARMA 模型系数的显著性检验均未通过。因此，干预影响退化为 $Z_t = \omega S_t^T + \varepsilon$，$t > 61$ [14]。参数点估计 $\omega^* = -11730.26763$ 为序列 Z_t 的均值，随机误差项标准差估计为 $\sigma_\varepsilon^* = 6510.564367$。这里表明，干预作用在统计意义上表现出了突发性和持续性。这一结论与我们对疫情的认知相一致。

3.3. 净化序列的计算及建模

净化序列指的是消除了干预影响的序列，由实际值 x_t 减去干预影响值 Z_t 得到：

$$y_t = x_t - Z_t^* = x_t - \omega^* S_t^T \tag{5}$$

式中，y_t 代表净化值，

$$S_t^T = \begin{cases} 0, t < T = 62 \\ 1, t \geq T = 62 \end{cases} \tag{6}$$

对净化序列 y_t 建立拟合模型时，仍然选择 ARIMA 季节乘积模型，利用 SAS 软件反复调试后，最优模型为 $\mathrm{ARIMA}(8,2,1)\times(1,1,0)^{12}$。

$$\nabla\nabla_{12} y_t = \frac{1 + 0.88211B}{\left(1 + 0.49007B^8\right)\times\left(1 + 0.58595B^{12}\right)}\varepsilon_t \tag{7}$$

3.4. 组建干预分析模型

利用上文过程和参数估计结果，最终确定干预分析模型形式如下：

$$\hat{x}_t = \frac{1 + 0.88211B}{\left(1 + 0.429007B^8\right)\times\left(1 + 0.58595B^{12}\right)}\varepsilon_t + \omega^* S_t^T \tag{8}$$

式中，$S_t^T = \begin{cases} 0, (t < 62) \\ 1, (t \geq 62) \end{cases}$。

为检验模型的建立及拟合成果，绘制拟合效果图如下：

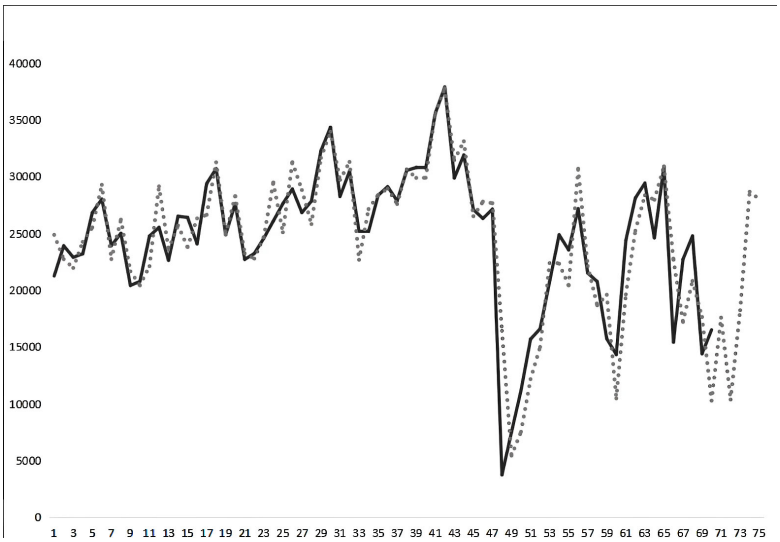

Figure3. Fitting effect chart of the intervention analysis model
图 3. 干预分析模型的拟合效果图

根据图 3 可观察到，在干预事件发生之前，大部分拟合线(虚线)与原始数据线(实线)较为吻合。对于我国铁路客运量未来 5 期的预测，呈现出波动上升的趋势。在没有其他干预事件的情况下，可以判断该预测符合实际走向。使用均方根误差 RMSE 来衡量精确预测精度，计算结果为 2856.15949，预测精度较高。

4. 结论

为了对我国铁路客运量的趋势进行预测研究，本文使用来源于国家统计局的月度铁路客运量数据集，对数据进行趋势分析后，发现由于疫情的影响导致原本具有周期性、增长趋势的客流序列产生离群值，并且干预发生后对滞后期仍有影响。基于数据的特点，本文将 ARIMA 的乘积季节模型与干预分析功能相结合，构建时间序列干预预测模型，结果表明，净化序列的最优模型 $\mathrm{ARIMA}(8,2,1)\times(1,1,0)^{12}$ 与干预影响程度 Z_t 组合而成的干预分析模型拟合程度较好，预测符合实际走向，模型的均方根误差(RMAE)为 2856.15949，总体上看精确度较高。

干预影响程度 Z_t 的计算结果揭示了 2020 年 2 月，新冠疫情的影响对我国铁路客运量的负冲击为 −26119.6386，平均负冲击为−11730.26763，这暗示着，在当今铁路高速发展的背景下，新冠疫情对铁路的影响是长期的，预计在未来的一段时间内，铁路客运量仍将受到一定程度的影响。

基于模型的预测结果和干预影响程度的计算，铁路客运部门可以制定合适的运营策略和调整运力安排，以适应突发事件对铁路客流量的影响。例如，如果新冠疫情导致客运量减少了 50%，铁路部门就可以考虑减少班次、缩短运营时间或增加运力来应对。未来，铁路部门需要密切关注新冠疫情的发展趋势，及时调整运营策略和运力安排。如果疫情得到有效控制，客运量逐渐恢复正常，便可以考虑通过增加班次或优化服务等措施来提高客运量。

5. 研究展望

在时间和水平等方面的问题上，本论文的研究还存在着不足，有待进一步的研究，包括以下几点：

(1) 本文在构建干预影响程度 Z_t 的模型时，由于数据特点，干预影响程度的参数 ω 采用的是影响程度的均值，可能会加大拟合误差。收集更多处于干预影响期间的数据有助于捕获干预影响程度 Z_t 的变化

趋势，以构建更科学的模型。

(2) 本文只讨论了单因素时间序列在客运量预测中的应用，而实际生活中，客运量的变化往往受到多种因素的共同影响，如经济波动、政策变更、气候变化等。为了更加贴近实际情况，需要对各种因素进行综合分析，权衡它们之间的作用，以更精确地预测客运量。因此，在未来的研究中，可以逐步拓展时间序列分析的领域，深入研究多种因素共同作用下的客运量预测问题，为交通规划、物流运输等领域的决策提供更加科学、可靠的依据。

(3) 为提高本文模型的普适性和推广性，未来可以考虑进行跨国家的实证研究，比较不同国家的客运量预测模型的适用性和可行性，以此进一步优化本文的预测模型以适应不同地区的实际情况。

致　谢

感谢老师和同学一路上的指导及帮助。

参考文献

[1] 俞金梦. 基于时空特征分析的铁路客运量预测模型研究[D]: [硕士学位论文]. 北京: 北京交通大学, 2021.

[2] 王艳辉, 王卓, 贾利民, 等. 铁路客运量数据挖掘预测方法及应用研究[J]. 铁道学报, 2004(5): 1-7.

[3] 夏国恩, 金炜东, 张葛祥. 改进 SVR 及其在铁路客运量预测中的应用[J]. 西南交通大学学报, 2007(4): 494-498.

[4] 王卓, 王艳辉, 贾利民, 等. 改进的 BP 神经网络在铁路客运量时间序列预测中的应用[J]. 中国铁道科学, 2005(2): 130-134.

[5] 谢小山. 基于遗传算法和 BP 神经网络的铁路客运量预测研究[D]: [硕士学位论文]. 成都: 西南交通大学, 2010.

[6] 王慧晶. 基于灰色预测模型的铁路客运量预测研究[J]. 铁道运输与经济, 2006(6): 79-81.

[7] 郝军章, 崔玉杰, 韩江雪. 基于 SARIMA 模型在我国铁路客运量中的预测[J]. 数学的实践与认识, 2015, 45(18): 95-104.

[8] 钱名军, 李引珍, 阿茹娜. 基于季节分解和 SARIMA-GARCH 模型的铁路月度客运量预测方法[J]. 铁道学报, 2020, 42(6): 25-34.

[9] 高旭, 李兴东. 干预分析模型在时间序列中的应用[J]. 科学技术创新, 2021(29): 26−28.

[10] Sangkwon, L., *et al*. (2005) Estimating the Impact of the September 11 Terrorist Attacks on the US Air Transport Passenger Demand Using Intervention Analysis. *Tourism Analysis*, **9**, 355-361. https://doi.org/10.3727/108354205789807238

[11] 蒋铁军, 周成杰, 张怀强. 事件对复杂经济时间序列影响的多尺度分析方法[J]. 统计与决策, 2019, 35(19): 10-14. https://doi.org/10.13546/j.cnki.tjyjc.2019.19.002

[12] 薛芳静, 黄灏, 许碧云, 等. 时间序列数据中的干预分析模型及 SAS 实现[J]. 中国卫生统计, 2017, 34(3): 509-511+514.

[13] 杨楠, 邢力聪. 干预分析模型在房价指数预测中的应用[J]. 统计与决策, 2005(21): 51-52.

[14] 王志坚, 宁哲源. 基于时间序列干预模型的我国就业趋势预测研究[J]. 统计与管理, 2023, 38(6): 13-23. https://doi.org/10.16722/j.issn.1674−537x.2023.06.008