



IP 骨干网流量的自动化预测

韦炬, 阮科, 黄晓莹, 陈迅, 黄灿灿

(中国电信股份有限公司研究院, 广东 广州 510630)

摘要: 高效、可靠的网络流量预测是网络规划、扩容建设的基础。互联网流量目前缺乏完备的理论模型, 行业内大多根据工程实践特点, 设计简化可操作的预测模型以满足 IP 网络规划需求。首先根据中国电信自身 IP 骨干网流量预测工作的需求及特点, 使用时间序列分析的多因子回归模型和函数自适应模型对 IP 骨干网流量进行分析和预测, 基于大量现网实际数据的仿真运算, 对比两种模型的特点、优劣和适用场景, 提出了一种预测模型选择和参数优化的原则和方法。在此基础上, 构建了可以满足百千量级时间序列要求的自动化流量预测系统, 极大简化并提升了流量预测工作的效率。最后, 展望了未来 IP 流量预测工作的延展方向和关注重点。

关键词: 时间序列; 流量预测; 预测模型

中图分类号: TN915

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2020153

Automatic prediction for IP backbone network traffic

WEI Xuan, RUAN Ke, HUANG Xiaoying, CHEN Xun, HUANG Cancan

Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China

Abstract: Efficient and reliable network traffic prediction is the basis of network planning and capacity expansion construction. Currently, there is no integral theoretical model to describe internet traffic. Most of the industry designs simplified and operable prediction models. Firstly, according to the characteristics of China Telecom's IP backbone network traffic and its planning requirements, the IP backbone network traffic was analyzed and forecasted by using the multi-factor regression model and the function adaptive mode of time series. The characteristics, advantages, disadvantages and applicable scenarios of these two models were compared based on simulation of a large number of actual network data. A set of principles and methods for selecting prediction model and optimizing parameters were proposed. Then, an automatic forecasting system with the high performance of dealing with hundreds of time series was built to greatly simplify and improve the traffic prediction efficiency. Finally, the development orientation of network capacity extension and key points of future IP traffic prediction were prospected.

Key words: time series, traffic forecast, prediction model



1 引言

近年来,随着各种互联网应用和服务的不断兴起,作为主要承载网的IP骨干网流量也持续高速增长。为了适应IP流量的快速增长,满足上层业务和应用的承载,运营商需密切监控网络流量现状、及时评估和预测规划期内流量的走向和规模,并以此作为后续网络规划的输入,对网络结构、带宽分布等按需进行优化改造或扩容建设。因此,IP骨干网流量的准确分析和预测在网络规划工作中具有至关重要的地位。

不同于传统电话网络,IP网络流量具有突发性、自相似性(self-similar)、长相关性、周期性、混沌性等特点。这些特点使IP网络根本性地区别于传统电话网,决定了传统泊松流量模型不再适用于如今的IP网络。为此,近年针对IP网络提出一些基于智能算法的预测方法,包括神经网络理论模型、模糊理论模型等,但这些模型距离实用性的要求还存在较大差距。因此在缺乏理想理论模型支持的情况下,有必要从现网实测数据的分析中提炼出适用的预测方法以弥补理论模型缺失的不足。

2 IP骨干网流量预测需求分析

对于IP骨干网级别的大型网络而言,网络流量除了具备以上IP网络共性特点外,还具有一些与某个特定业务流不一样的其他特点。因而IP骨干网对于流量预测的需求也将不同于某个具体的业务流量的预测需求。

首先,IP骨干网的流量具有高度聚合性,通常以某个省份或城市为单位进行流量的收集和分析,流量颗粒度高达几十Gbit/s甚至几百Gbit/s,实际工程中通常主要关注管道内整体流量的总体变化趋势。

其次,IP骨干网的规划周期通常以年为单位,如预测1年后、3年后规划期的流量总量。对这种

以年、季度、月份等长时间颗粒度进行观测的流量,流量长期增长的幅度远大于短期的流量突发波动,长期流量变化曲线更容易呈现规律增长变化的趋势。图1展示了短期流量(日流量)和长期流量(月流量)随时间变化的情况,可看到短期流量随时间的波动性是比较大的,而长期流量则表现出稳定的增长趋势。

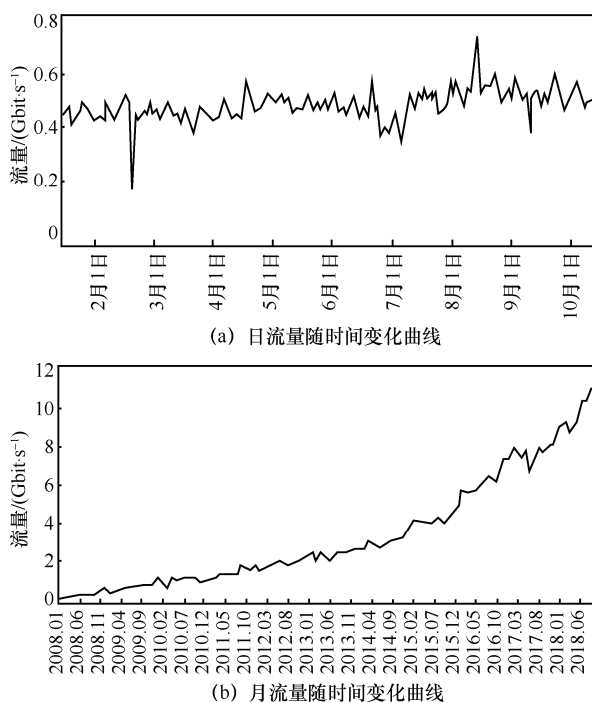


图1 短期流量(日流量)和长期流量(月流量)变化对比

对于IP骨干网的流量预测,尽管缺乏完善的理论模型,无法设计精确的数学预测模型,但存在大量的工程实践经验可供参考。

3 IP骨干网流量预测方法

基于时间序列分析的方法首先需要对时间序列进行分解,然后通过不同的模型、算法分别对各分量进行拟合、预测,最后将各分量合并得到时间序列的总体预测结果。

对于IP骨干网流量预测而言,历史流量 $y(t)$ (t 代表时间)首先分解为趋势项 $g(t)$ 、季节项 $s(t)$ 、节假日项 $h(t)$ 以及随机项 $r(t)$,如图2所示。

因骨干网流量的采集记录时间周期以月为单位, 所以图2中不存在节假日项 $h(t)$ 。如果采用加法形式, 流量的时间序列分解用计算式可表示为:

$$y(t)=g(t)+s(t)+r(t) \quad (1)$$

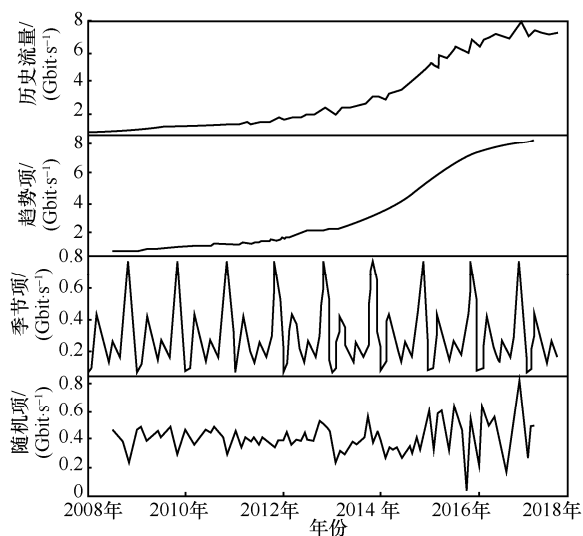


图2 历史流量的时间序列分解

历史流量分解后, 需要选择合适的模型或算法对趋势分量和季节分量分别求解并预测, 而随机分量通常是由某些不具确定性的随机因素导致的白噪声构成的, 通常无法预测量化。

3.1 多因子回归模型

多因子回归模型是一种结合时间序列趋势外推和流量变化原因分析的预测方法, 即对时间序列的长期趋势主要通过历史趋势外推法拟合和预测, 同时分析流量构成中的主要应用类型及其流量特性, 并对流量特性有重要影响的宏观或微观因素进行分析和筛选, 以用于对趋势外推法得到的预测结果进行修正和优化。

利用多因子回归模型对IP骨干网流量进行预测, 首先利用通信网络中常用的曲线方程(包括线性方程、指数方程、幂函数方程、二次曲线方程、多项式方程等)对网络流量的长期变化趋势进行曲线拟合。多项式方程一般可写作:

$y(t)=a+bt+ct^2+dt^3+\dots$, 其中 a 是截距, b 是 t 的一次项系数, c 是 t 的二次项系数, d 是 t 的三

次项系数; 其中幂函数方程、二次曲线方程都可看作多项式方程的特殊形式。根据IP骨干网流量的实际情况, 实践中多采用多项式方程来拟合流量的长期趋势, 此时可同时考虑流量增长的线性和非线性特性, 有利于适应多种场景的预测需求。

由于趋势外推曲线拟合无法反映流量发生变化的内在原因, 因此需要引入一些与流量变化密切相关的影响因子, 以此作为曲线变化的约束条件来修正流量的长期趋势。这些影响因子包括宽带用户数、平均流速、平均接入带宽以及应用类型、用户行为、内容分布、资源策略、地域经济状况等。由于涉及的影响因子很多且各因子之间具有错综复杂的关系, 因此影响因子的选取是模型的关键。第4.1节将重点分析多因子回归模型影响因子的选取问题。

3.2 函数自适应模型

函数自适应模型并非传统的时间序列模型, 而是把时间序列预测问题转变成了一个曲线实时、自适应拟合练习的过程。在本文任务中, 流量作为因变量, 是趋势项、季节项分别进行曲线拟合的加和表现, 在具体实现中需要应用Facebook的开源prophet包。函数自适应模型可调整的参数较多、灵活性较大, 但容易出现过拟合的问题, 在具体实践中应采用相关措施, 尽量避免此类问题的发生。

3.2.1 趋势项 $g(t)$ 的预测

对历史流量趋势项的分析求解, 其实可以归结为选择变点及确定流量增长率的问题, 也即求解出特定变点下时间序列的增长率。可以采用以下两种方式来实现: Logistic趋势增长、分段线性趋势增长。

(1) Logistic增长趋势 $g(t)$ 的预测

Logistic增长趋势多应用于具有非线性增长且逐渐趋向饱和的场景中。Logistic增长函数 $g(t)$ 可表示为:



$$g(t) = \frac{C(t)}{1 + \exp(-(k + a(t)^T \delta) \cdot (t - (m + a(t)^T \gamma)))} \quad (2)$$

其中, $C(t)$ 对应时间序列 $g(t)$ 的饱和值, k 是时间初始点的趋势增长率, $a(t)$ 是指示向量 $\{0, 1\}$, $k + a(t)^T \delta$ 是时间 t 处的趋势增长率, δ 是增长率的变化量, m 是偏移参数, γ 是变点调整系数。

采用 Logistics 增长趋势进行曲线拟合和预测时, 可以调整的参数主要有 3 个: 容量 Cap、变点数量以及变点权值参数 cps。其中, Cap 对应着流量时间序列的饱和值 $C(t)$, 它与带宽用户数、平均带宽、平均流速等参数密切相关, 此时可结合多因子回归模型中的影响因子对之进行合理估算。变点数量及变点权值参数与式 (2) 中的 δ 、 γ 、 m 相关联。

(2) 分段线性增长趋势 $g(t)$ 的预测

众所周知, 线性函数是 $y(t)=kt+b$, 其中 k 是斜率, b 是截距。分段线性增长函数是在每两个变点之间的子区间上, 函数都是线性函数, 但在整段区间上, 函数是一个分段的折线形式。此时, 时间序列增长趋势曲线的求解转化为分段线性函数的逼近和拟合问题。通过运算求解, 得到基于分段线性函数的增长趋势模型如下:

$$g(t) = (k + a(t)\delta) \cdot t + (m + a(t)^T \gamma) \quad (3)$$

分段线性增长趋势适用场景较为广泛, 理论上, 只要变点选择得当且数量足够多, 任何非线性增长都可以通过分段线性函数拟合。分段线性增长趋势模型可调整的参数包括: 变点数量以及变点权值参数 cps。

3.2.2 季节项 $s(t)$ 的预测

流量时间序列通常会随着天、周、月、年等时间周期变化而呈现季节性变化, 也称为周期性变化, 此时需要用周期性函数来描述时间序列的周期性变化。在数学分析中, 常用正弦、余弦函数即傅里叶级数来表示:

$$s(t) = \sum_{n=1}^N \left(a_n \cos \frac{2\pi nt}{p} + b_n \sin \frac{2\pi nt}{p} \right) \quad (4)$$

其中, P 是时间序列的周期, N 为级数, a_n 、 b_n 分别是余弦函数、正弦函数的系数。此时流量季节项 $s(t)$ 的问题转变为确定 P 、 N 参数并求解最优系数 $\beta = (a_1, b_1, \dots, a_n, b_n)^T$ 的问题。对于流量季节项 $s(t)$ 的预测, 可调整的参数主要是季节权值参数 sps。

4 海量流量预测对象的自动化预测

大型 IP 骨干网流量的预测需求包括对出省流量、IDC 流量、城域网流量、国际流量、互联互通流量等进行分析预测。以出省流量为例, 除港澳台外, 全国共有 31 个省份, 且流量分流入、流出两个方向; 以城域网为例, 全国共有将近 300 个城域网, 同时区分流入、流出方向。这些预测对象加起来达到几百甚至上千个规模, 而这么多的流量时间序列都需要逐一对其进行分析预测, 如果完全依靠手工选择模型及相应参数、调整参数, 如此巨大的工作量将使流量预测工作变成几乎不可能完成的任务。因此, 将基于以上两种流量预测模型, 提出针对海量流量预测对象的自动化预测系统, 如图 3 所示。

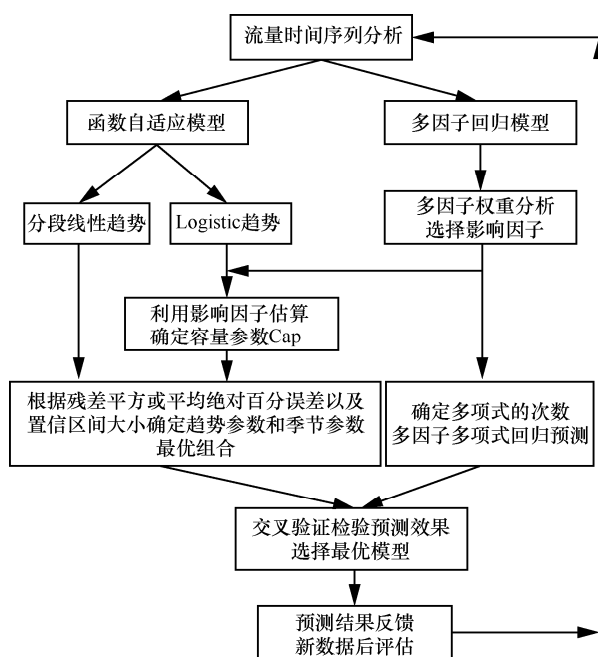


图3 自动化流量预测系统流程

自动化流量预测系统的预测流程如下：利用两种预测模型分别对流量进行预测，在同等置信区间的前提下，选取残差平方和（sum of squares due to error, SSE）最小或平均绝对百分误差（mean absolute percentage error, MAPE）最小的模型作为该流量时间序列的最终预测模型；多因子回归模型中，需要通过多因子相对重要性分析、数据来源可靠性等综合条件选取最重要的影响因子；自适应模型中，关键是在置信区间一致的前提下选取 SSE 或 MAPE 最优的预测模型及参数组合；最后将通过交叉验证检验机制来评估预测的效果，同时将预测结果反馈到流程起点，并与后面持续新采集的流量数据进行分析对比，结合目标区域网络扩容和实际运营情况分析预测流量与实际流量的差异，以此作为流量预测的后评估进一步优化模型和参数，最终形成闭环的流量预测评估调整机制。目前该系统主要应用于中国电信的大型 IP 网络省际流量、城域网流量、IDC 流量的长期预测中，极大地提高了 IP 网络规划建设的质量和效率。

4.1 多因子回归模型影响因子的选择

与 IP 骨干网流量密切相关的影响因素包括宽带用户数、流速、接入带宽以及资源策略、经济发展等，如果要求更精细的分析，还将涉及用户行为分析、应用类型分析及内容分布等。但对于骨干网出省流量这类大颗粒度、非实时流量而言，在工程实践中重点关注用户数、平均流速、平均接入带宽以及地域经济发展情况等对流量的影响就基本足够了。

为此，本文收集了各省、各地市的宽带用户数（user）、平均流速（speed）、平均接入带宽（bw）、国内生产总值（GDP）等历史相关数据，并结合时间多项式进行多因子回归分析，以寻找更适合的解释变量及回归模型。该过程主要包括两方面：多个不同解释变量下拟合判定系数 R^2 的分析对比以及解释变量相对重要性分析。

4.1.1 多个不同解释变量下拟合 R^2 的分析对比

对 IP 骨干网出省流量分别采用以下多种解释变量的组合进行多因子回归，得到各种情况下的判定系数 R^2 ， R^2 越接近 1 说明模型拟合效果越好。表 1 为各种情况下 R^2 的对比。

表 1 不同解释变量下的 R^2 对比

解释变量	判定系数 R^2
t	0.973
t +GDP	0.984
t +user	0.982
t +speed	0.976
t +bw	0.980
t +GDP+speed	0.989
t +GDP+bw	0.984
t +speed+bw	0.981
t +bw+user	0.982
t +GDP+speed+bw	0.989
t +bw+user+GDP	0.984
t +bw+user+speed	0.987
t +bw+user+speed+GDP	0.990

由表 1 可见，解释变量由只有一个时间参数 t 到最多增加了 4 个其他解释变量，尽管判定系数 R^2 有不同程度的增加，但增加最大不到 2%。由此可见回归解释变量的增加对预测效果的提升是有限的。依据奥卡姆剃刀原理，在精度相当的情况下应遵从模型简化的原则，尽量减少解释变量的个数，因此非常有必要对回归解释变量做进一步的评估和选择。

4.1.2 解释变量相对重要性分析

评价解释变量相对重要性最简单的方法是比较标准化偏回归系数，它表示当其他解释变量不变时，该解释变量的变化可引起的响应变量的预期变化（以标准差单位度量）。表 2 列出了其中几个省份进行多因子回归后的标准化偏回归系数。



表2 标准化偏回归系数

省份	截距	t	user	GDP	bw	speed
A 省	1.308 550 445 0	1.588 941 391 5	0.001 327 578 9	-0.000 949 902 5	-0.016 033 093 7	0.000 225 597 1
B 省	-0.760 127 102 3	0.279 224 397 7	0.001 632 749 8	0.000 185 975 4	0.019 669 111 7	-0.000 185 031 9
C 省	-0.219 677 824 6	0.441 768 532 4	0.002 783 861 9	-0.000 280 364 5	0.011 487 391 0	-0.000 334 667 4
D 省	-0.485 530 931 3	0.762 304 378 7	-0.000 580 710 1	0.000 187 694 3	0.010 025 748 9	-0.000 335 265 5

由表2可见,对于这些省份而言,时间参数 t 是最重要的解释变量,时间 t 一个标准差的变化将会引起0.28~1.59个标准差的流量变化。除时间参数 t 外,平均接入带宽(bw)和宽带用户数(user)是相对较重要的解释变量,而国内生产总值(GDP)及平均流速(speed)这两个解释变量对出省流量的影响非常小,可考虑在模型中去除。

另外,考虑个别解释变量的数据获取存在难度及准确性的问题,例如国民生产总值总是滞后发布,且只有分省历史年度数据,缺乏可靠的官方预测数据;各省平均流速的采集难度较大且准确性也有待验证;而各省宽带用户数(user)及平均接入带宽(bw)相对可靠和准确许多。

回归模型的选择总是涉及预测精度与模型简洁度的调和问题。因此,综合以上对 R^2 的分析对比、解释变量相对重要性分析的评估结果,并考虑解释变量数据来源的可靠性及准确性,最终建议选取时间(t)+宽带用户数(user)+平均接入带宽(bw)作为多因子回归模型的解释变量。

4.2 函数自适应模型参数的确定

当使用函数自适应模型来对IP骨干网流量进行拟合和预测时,首先需要确定趋势项和季节项的参数,即第3节中提到的容量参数Cap(仅对Logistics趋势增长)、变点数量、变点权值参数cps以及季节权值参数sps。变点数量和变点权值参数反映了拟合曲线变化的速度和灵活性,变点数量越多、变点权值越大则曲线变化越快、趋势越灵活,但过拟合的风险也越大,预测空间的不确定性也越大;季节权值参数cps反映了季节分量的

灵活度,该值越大,则季节的变化越快、越显著。

在工程实践中,对于Logistics趋势增长中的容量Cap参数,通过影响因子法来估算和验证。上述多因子模型中,已选取宽带用户数及平均带宽作为显著性影响因子;利用以上影响因子的历史数据和增长速率,结合相关业务的覆盖规模、发展规律,通过历史外推及合理估算得到预测期内流量所应到达的饱和容量Cap值。

对于模型中其余几个趋势参数和季节参数,则主要通过大量的尝试、模拟和仿真,反复修改各参数值并进行迭代运算,同时记录每种情况下SSE、MAPE以及置信区间的大小,最终选取预测3年后0.8置信区间不超过标准差(standard deviation, SD)且SSE或MAPE足够小的相关参数组合作为预测模型的最优参数。图4、图5通过迭代生成了SSE、MAPE分别随变点权值参数cps和季节权值参数sps变化的散点图。

4.3 流量预测模型的对比及选择

在同等置信区间的前提下,选取SSE或MAPE最优的模型作为流量时间序列的预测模型。下文对31个省份的出省流量分别用以上两种模型进行拟合、预测,比较两种模型下SSE的大小,以及函数自适应模型Logistic趋势和分段线性趋势下的SSE,最终选取SSE最优的模型。

4.3.1 多因子回归模型与函数自适应模型的对比

图6是多因子回归模型与函数自适应模型的比较。由图6可知,函数自适应模型SSE更优的省份占据了多数,达26个;只有5个省份多因子回归模型SSE更优,预计与这几省的流量与影响

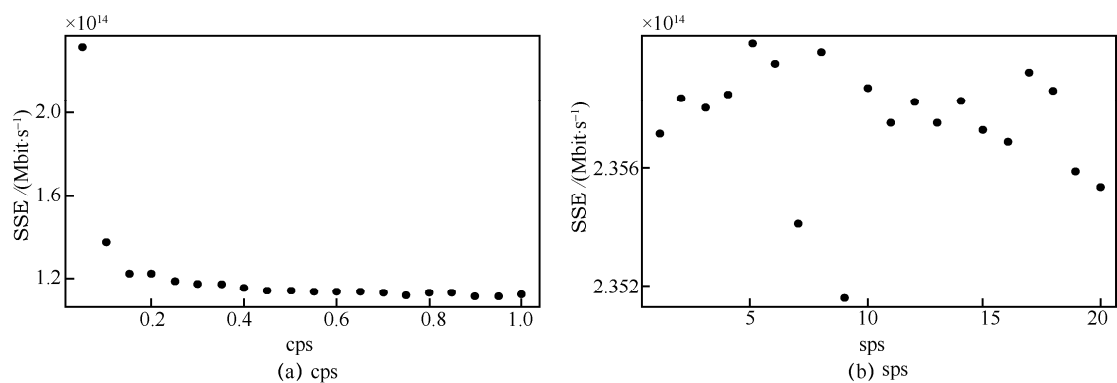


图 4 SSE 随 cps、sps 变化的散点示意图

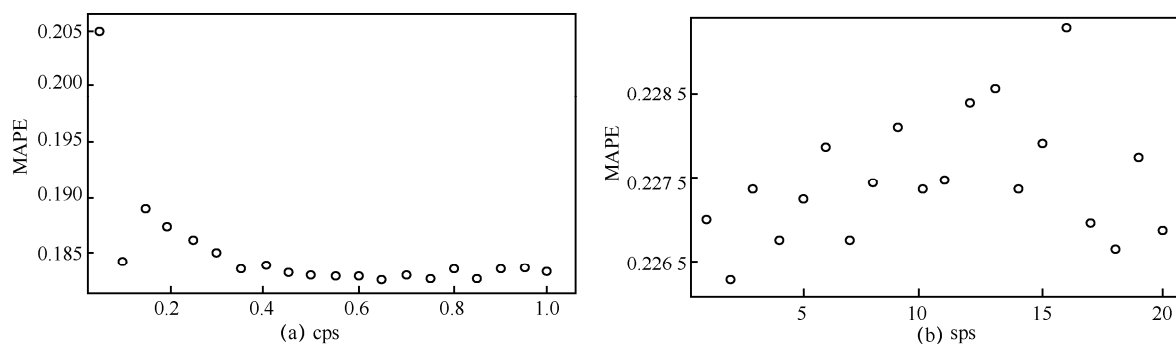


图 5 MAPE 随 cps、sps 变化的散点示意图

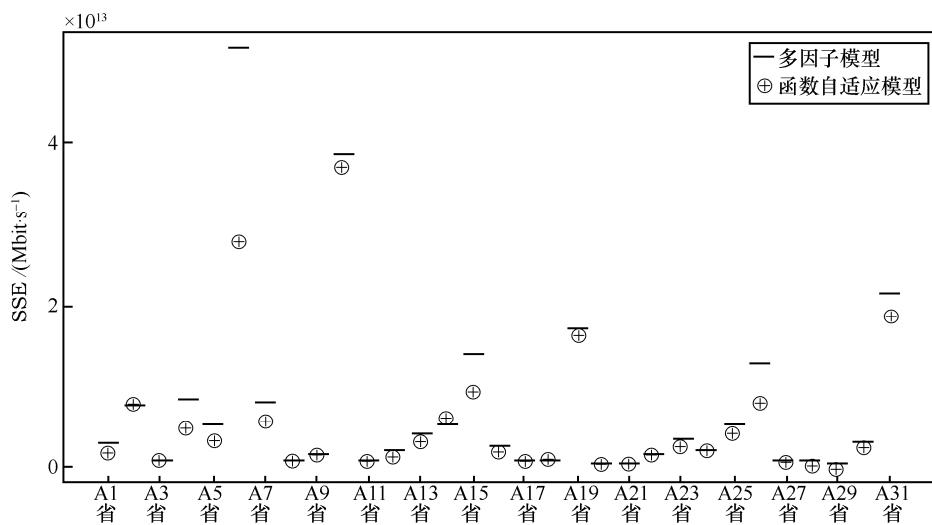


图 6 多因子回归模型与函数自适应模型 SSE 对比

因子高度相关有关。函数自适应模型由于可随着时间、变点的变动而灵活调整拟合曲线变化速率，因而表现出更优异的拟合效果。

4.3.2 Logistic 趋势与分段线性趋势对比

图 7 是函数自适应模型中两种趋势增长模型

的对比。由图 7 可见，分段线性趋势模型 SSE 更优的省份占据绝大多数，达 29 个；Logistic 趋势 SSE 占优的只有 2 个。这与这些省份的流量是否具有饱和态势有关。Logistic 趋势仅适合于具有饱和效应的增长，而分段线性趋势则具有更好的适应性。

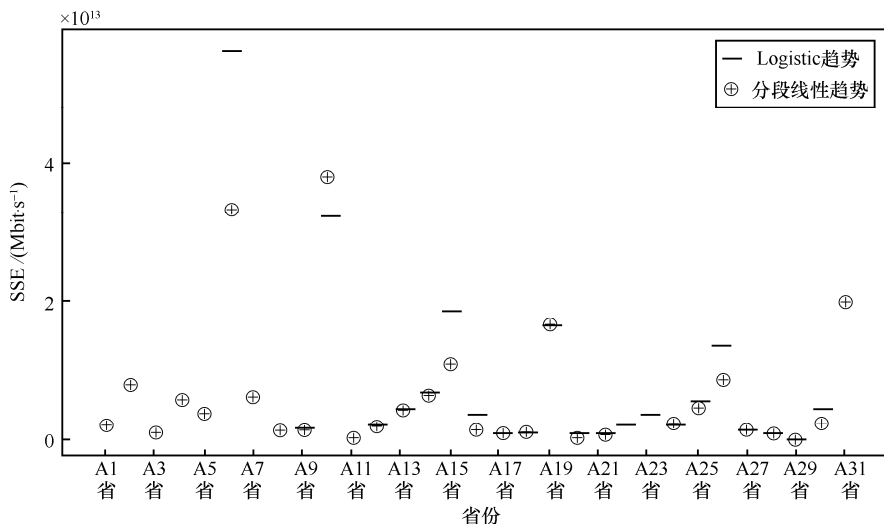


图7 Logistic趋势与分段线性趋势 SSE 对比

4.4 流量预测结果分析

利用上文提出的自动化预测系统,本文对IP骨干网的31个省份的出省流量、近300个城域网流量、直连骨干的20多个IDC流量分别进行了大量的仿真和预测,通过选取不同的影响因子、修改趋势模型、改变模型参数,对比每种情况下的残差分布、SSE、MAPE以及交叉验证的性能矩阵,最终得到每个流量时间序列对应的最优预测模型和相应参数。

下面以A省出省流量的预测结果为例,经自动化预测系统的模拟对比,A省出省流量的最优模型是函数自适应模型,趋势增长是分段线性增长,对应的变点数量为25,变点权值参数cps为0.1,季节权值参数sps为9。

A省出省流量的趋势项和季节项如图8所示。由图8可知,A省出省流量增长较平缓,同时表现出明显的季节性波动,每年1月末到2月初是流量峰值期,而9月末到10月初则处于波谷区域,与A省历史流量波动表现相吻合。

A省总体预测结果如图9所示。图9中,出省流量历史区间为2008年1月—2019年2月,由黑点表示;预测区间为2019年3月—2024年12月,预测流量曲线由灰线表示。图9中灰色曲线周围的灰色区域对应了置信度0.8的置信区间。由图9可见,拟合曲线可以较好地与原始流量吻

合,意味着该模型具有很好的SSE性能,且置信区间较小,能够满足预测期3年后置信区间不超SD的要求。另外,可看到置信区间将随着预测时长加大而变大,此时不确定性增加,预测准确性也下降,因此基于可靠性考虑不建议预测时长过长。

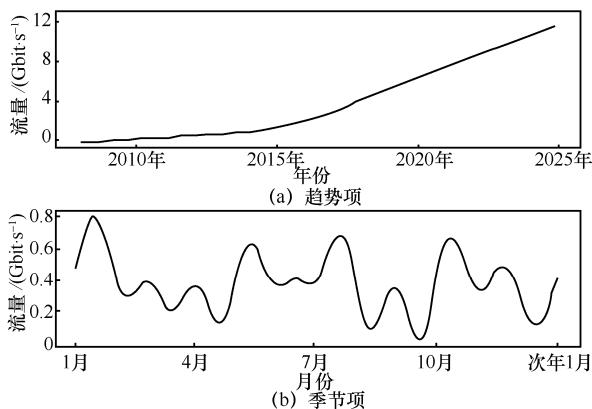


图8 A省出省流量的趋势项和季节项

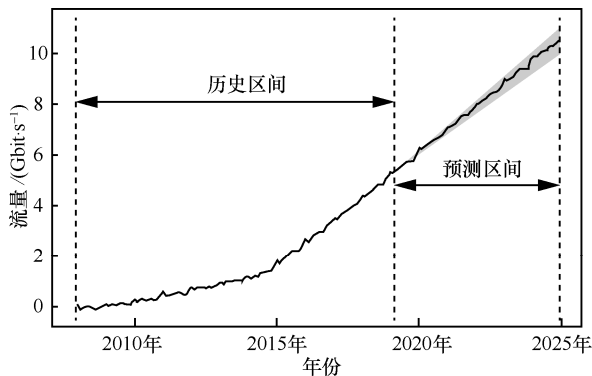


图9 A省出省流量预测结果

图 10 是实施交叉验证后得到的 MAPE 性能, 由图 10 可看到该模型的 MAPE 大多位于 10% 左右, 总体基本控制在 15% 以内。对于长期预测而言, 这样的准确度已可以满足要求。

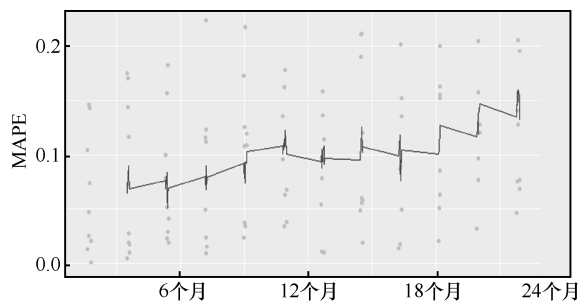


图 10 A 省出省流量预测模型的交叉验证 MAPE

5 结束语

本文针对 IP 骨干网的流量特点、预测需求及现有问题, 基于多因子回归模型和函数自适应模型, 构建了一个适用于具有大规模流量对象预测需求的自动化预测系统。该系统通过大量的仿真和运算, 最终选取出具有最优性能的预测模型及其相应参数, 在工程实践中可以方便、快速地对大型 IP 骨干网各种流量的合理、高效预测。全国 31 个省份 (不含港澳台) 以及将近 300 个城域网的流量预测结果表明, 该系统可以满足目前 IP 网络规划流量预测精确度的要求, 提高了网络规划和优化迭代的效率。

随着各种新型应用和业务的不断涌现, 未来 IP 网络流量的增长和分布仍具有很大的不确定性, 今后的流量预测和网络规划将面临更大的挑战, 特别是对 IP 网络流量流向的分析、预测需求将愈加迫切。如何结合现有流量预测模型的优点, 并开拓新的预测方法以解决流量流向的分析、预

测问题将是今后流量预测的重点工作。

参考文献:

- [1] 张宾, 杨家海, 吴建平. Internet 流量模型分析与评述[J]. 软件学报, 2011, 22(1): 115-131.
ZHANG B, YANG J H, WU J P. Survey and analysis on the internet traffic model[J]. Journal of Software, 2011, 22(1): 115-131.
- [2] 赵子岩, 樊剑辉, 金鑫, 等. 电力骨干通信网信息流模型与流量测算[J]. 电信科学, 2017, 33(5): 153-163.
ZHAO Z Y, FAN J H, JIN X, et al. Information flow model and service flow calculation of electric power backbone communication network[J]. Telecommunications Science, 2017, 33(5): 153-163.
- [3] 欧亮, 陈迅, 沈晨, 等. IP 网络流量流向分析与预测技术研究[J]. 电信科学, 2013, 29(7): 24-29.
OU L, CHEN X, SHEN C, et al. Studies on traffics analysis and prediction technology for carrier-grade IP network[J]. Telecommunications Science, 2013, 29(7): 24-29.
- [4] TAYLOR S J, LETHAM B. Forecasting at scale[Z]. 2017.

[作者简介]



韦炬 (1974-), 女, 现就职于中国电信股份有限公司研究院, 主要研究方向为 IP 通信。

阮科 (1979-), 男, 现就职于中国电信股份有限公司研究院, 主要研究方向为 IP 通信。

黄晓莹 (1980-), 女, 现就职于中国电信股份有限公司研究院, 主要研究方向为 IP 通信。

陈迅 (1981-), 女, 现就职于中国电信股份有限公司研究院, 主要研究方向为 IP 通信。

黄灿灿 (1981-), 女, 现就职于中国电信股份有限公司研究院, 主要研究方向为 IP 通信。