



电力数据中的幂律分布特性

王冠男,邓春宇,赵悦,郑亚芹

(北京国电通网络技术有限公司大数据业务部 北京 100070)

摘要:数据的探索性分析在数据挖掘和分析的整体过程中具有举足轻重的作用。主要针对电力数据中具有非正态分布特征的数据进行分析研究。首先,采用统计分析的方法对电力数据中的用电量、售电量和发电量几个核心指标在地域上的分布规律进行统计分析,经验证,这些指标不符合正态分布,并且得到其幂律分布特性;然后,验证了细分行业的用电量符合对数正态分布;最后,结合电力业务背景,挖掘电力数据具有幂律分布以及用电量符合对数正态分布的原因。

关键词:电力指标;幂律分布;对数正态分布

doi: 10.3969/j.issn.1000-0801.2013.11.019

Power Law Distribution in Electric Power Data

Wang Guannan, Deng Chunyu, Zhao Yue, Zheng Yaqin

(Department of Big Data, Beijing Guodiantong Network Technology Co., Ltd., Beijing 100070, China)

Abstract: The explorative analysis plays an important role in the process of data mining. The non-normally distributed data in the electric power data was analyzed. Firstly, the distribution law of generated energy, electricity sales and electricity consumption with statistical methods across the province within the business area of SGCC were analyzed, and it was found that these indices were not normally distributed but follow the power law distribution. Then the electricity consumption of subdivided trade follows the log-normal distribution was proved. At last, the reasons that these data are not normally distributed were given.

Key words: indices of electricity, power law distribution, log-normal distribution

1 引言

在自然界和社会生活中,经常存在这样的数据,数值在某个特征附近变化,并且变化较小,这一类数据被称为正态分布,具有很多优秀的特性,如图 1(a)所示。比如人的身高、学生的考试成绩、同质群体的红细胞数以及数理统计中随机误差等都符合正态分布。在传统数据挖掘方法中,正态分布被誉为概率论里最重要的分布。

然而,随着大数据时代的到来,人们有能力处理完全数据而不是样本数据,因此,现实中很多完全数据不再符合正态分布,比如个人收入的分布、变化尺度的范围很宽,

波动可以跨越很多数量级。根据这类数据可以画出一条向右偏斜并且拖着长尾巴的累积分布曲线,显然,它与正态分布曲线有显著的不同,如图 1(b)所示。这种长尾分布表明,绝大多数个体的值很小,只有少数个体的值相当大。这种现象被称为长尾效应,或者重尾法则,本文研究的幂律分布是长尾分布中重要的一种^[3,4]。

研究表明,90%的电力指标数据都不符合正态分布。确定数据是否为正态分布在数据的探索性分析中具有举足轻重的作用,因为正态分布是很多统计方法和数据挖掘方法所必需的条件,比如线性和非线性回归模型、因子分析、典型相关分析以及主成分分析等,甚至包括

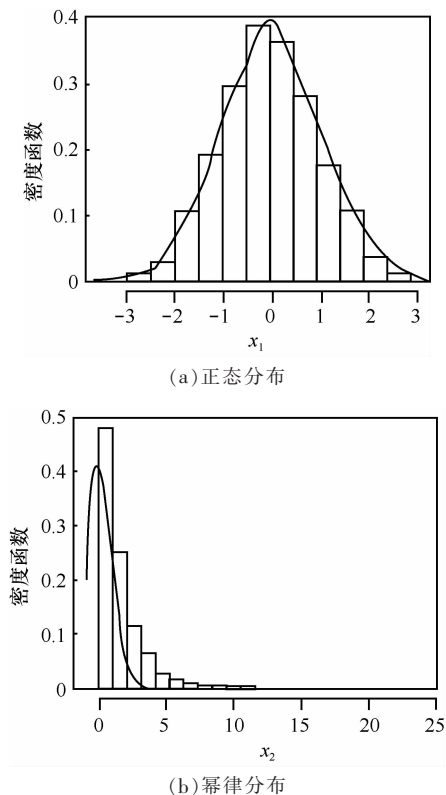


图1 正态分布和幂律分布

最基本的线性相关性系数的计算。若数据不符合正态分布,需要采取其他相应的办法,如数据转换或者更换分析方法,否则轻者导致模型计算结果的不准确性,严重者更会影响针对算法结果所做的决策。本文主要针对电力数据中的指标进行分析,研究其真正的分布规律,主要从3个方面展开,首先介绍数据源和相关指标,然后阐述电力指标数据的幂律分布特性以及细分行业用电量的对数正态分布特性,最后结合业务背景解释电力指标符合幂律分布的原因。

2 数据描述

2.1 数据源

本文选择全国发电量、国家电网公司售电量以及

全社会用电量3个指标作为分析研究的目标,选择原因如下。

- 全社会用电量与国家经济密切相关,研究表明,用电量的增长影响GDP的增长,而GDP的增长对用电量的增长影响不大,即用电量增长和GDP增长的关系是从用电量到GDP增长的单向因果关系。
- 全国发电量是直接衡量一个国家经济状况的重要指标。国家电网公司是关系国家能源安全和国民经济命脉的重要骨干企业,公司的售电量能够反映公司甚至国家的供电能力。3个电力指标具有较强的相关性。

表1阐述了所选3个电力指标的维度、时间和区域。本文所选取的时间为2011~2013年每年的2月和7月。

2.2 电力指标的非正态分布特性

本文采取密度函数曲线检验法和夏皮罗·威尔克提出的W统计量,将数据和图形相结合,共同验证全社会发电量、国家电网售电量和全国用电量的非正态分布特性。

图2展示了3个指标的密度函数曲线,由于空间限制,时间维度是2012年2月、2012年7月、2013年2月、2013年7月。由图2可知,3个电力指标大多集中在较小的数值上,只有少数分布在较大的数值上,在空间维度上不符合正态分布,并且都具有长尾分布的特征。

表2~表4分别展现了2008~2013年所有2月份和7月份中3个指标的W统计量检验结果,即P值。

由表2~表4可知,全国发电量、国家电网公司售电量和全社会用电量的W统计检验的结果都小于0.05,即有不到5%的信心认为3个电力指标符合正态分布,下文将进一步分析3个指标分别符合何种分布。

3 电力指标的幂律分布特征

3.1 幂律分布特征

长尾和幂律的区别是:幂律分布需要数据必须在统计意义上可以被幂律曲线拟合,而对于长尾分布而言,

表1 全国发电量、国家电网公司售电量、全社会用电量的数据描述

指标名称	维度	时间	区域
全国发电量	全国发电量、水电发电量、火电发电量	2011~2013年每年的2月和7月	全国范围内,共计32个区域
国家电网公司售电量	公司售电量、大工业售电量、农业售电量、居民售电量、商业售电量	2011~2013年每年的2月和7月	包括西藏电网在内的国家电网公司经营区域,共计27个区域
全社会用电量	全社会用电量、第一产业用电量、第二产业用电量、第三产业用电量、城乡居民用电量	2011~2013年每年的2月和7月	全国范围内,共计32个区域

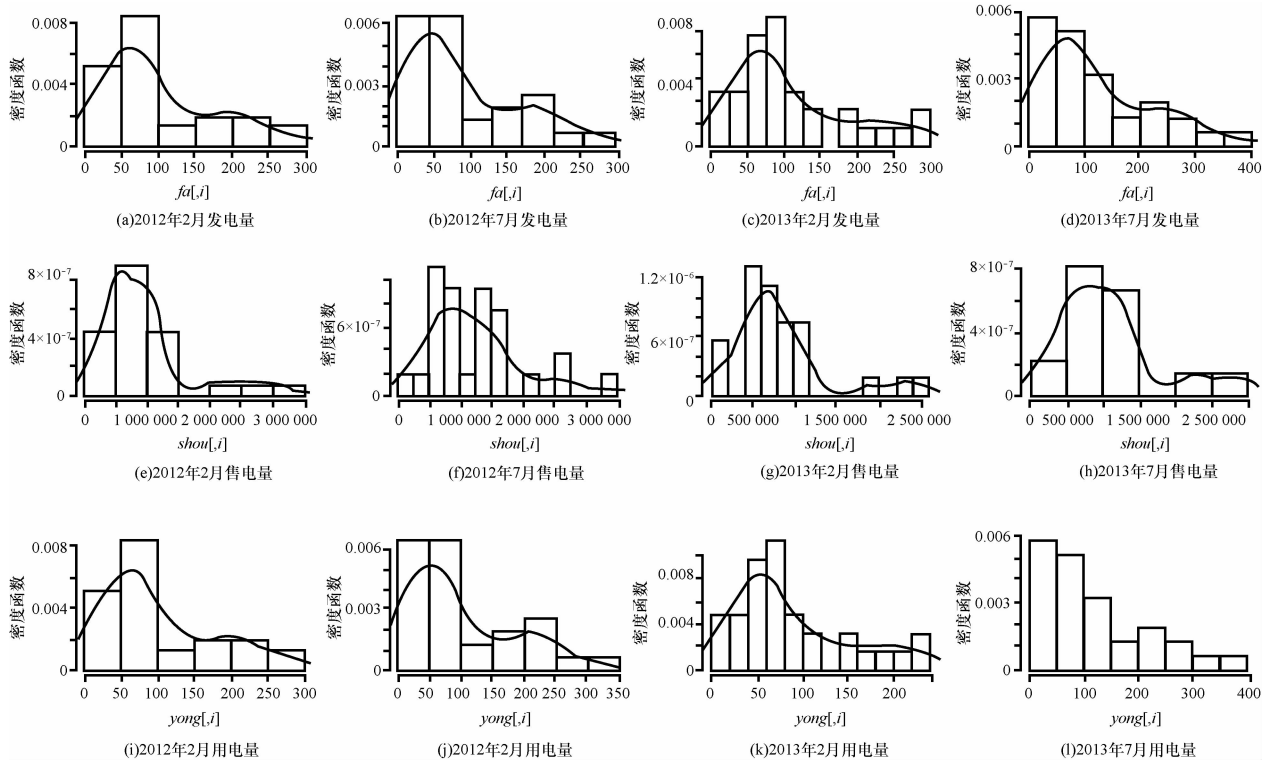


图 2 全国发电量、国家电网售电量和全社会用电量 3 个指标的密度函数曲线

表 2 全国发电量的正态分布 W 统计量检验结果

	2011 年 2 月	2011 年 7 月	2012 年 2 月	2012 年 7 月	2013 年 2 月	2013 年 7 月
全国发电量	0.004 306	0.002 193	0.005 545	0.001 644	0.003 541	0.003 022
水电发电量	9.29×10^{-11}	1.06×10^{-10}	9.42×10^{-11}	1.01×10^{-10}	9.06×10^{-11}	9.46×10^{-11}
火电发电量	1.96×10^{-11}	2.38×10^{-11}	1.99×10^{-11}	2.59×10^{-11}	1.90×10^{-11}	2.42×10^{-11}

表 3 国家电网公司售电量的正态分布 W 统计量检验结果

	2011 年 2 月	2011 年 7 月	2012 年 2 月	2012 年 7 月	2013 年 2 月	2013 年 7 月
公司售电量	0.005 381	0.000 352	0.000 304	0.024 02	0.001 11	0.005 622
大工业售电量	0.067 185	0.006 578	0.004 499	0.000 665	0.013 483	0.000 79
农业售电量	0.067 185	0.006 578	0.004 499	0.000 665	0.013 483	0.000 79
居民售电量	0.006 782	0.009 864	0.000 81	0.001 618	0.009 13	0.005 907
商业售电量	0.000 847	0.000 1	0.001 874	0.000 112	0.000 6	0.000 165

表 4 全社会用电量的正态分布 W 统计量检验结果

	2011 年 2 月	2011 年 7 月	2012 年 2 月	2012 年 7 月	2013 年 2 月	2013 年 7 月
全社会用电量	0.006 419	0.001 59	0.002 354	0.000 641	0.005 449	0.000 422
第一产业用电量	3.81×10^{-5}	0.000 35	3.61×10^{-5}	0.000 539	0.000 127	4.53×10^{-5}
第二产业用电量	0.001 954	0.000 561	0.000 619	0.000 188	0.003 673	0.000 114
第三产业用电量	0.002 149	0.024 355	0.123 449	0.016 53	0.039 999	0.028 333
城乡居民用电量	0.005 241	0.008 792	0.009 853	0.022 145	0.048 356	0.003 841

能否被幂律曲线精确地拟合并不重要,重要的是有一个长长的尾巴。已知全国用电量、国家电网公司售电量和全社会发电量属于长尾分布,但是仍需要进一步证明其为幂律分布。

幂律分布的分布特征如图 1(b)所示,通用计算式为 $y=cx^{-r}$, x 和 y 是正的随机变量, c 、 r 为常数,且大于 0。幂律分布的特点是绝大多数事件的值很小,只有少数事件的值相当大。将幂律分布公式的两边分别取对数,则 $\ln x$ 与 $\ln y$



存在以下的线性关系: $\ln y = \ln c - r \ln x$, 幂律分布在双对数坐标下表现为一条斜率为幂指数的负数的直线, 这一线性关系是判断某个随机变量是否满足幂律的依据, r 为幂律指数。

图 3 展示了 3 个电力指标的幂律分布特征, 时间维度是 2012–2013 年每年的 2 月份和 7 月份。由图 3 可知, 除了特殊原因引起的较小数值外 (西藏地区), 其他数值基本符合线性关系, 并且幂律指数为正数, 因此 3 个指标符合

幂律分布的特征, 属于幂律分布。

表 5~表 7 分别展示了 3 个电力指标在双对数坐标下的参数拟合结果, 其中, 幂律指标 r 都大于 -2, 验证了图 3 的结论, 全国发电量、国家电网公司售电量和全社会用电量符合幂律分布。

3.2 用电量的对数正态分布特征

由上文可知, 全国发电量、国家电网公司售电量和全

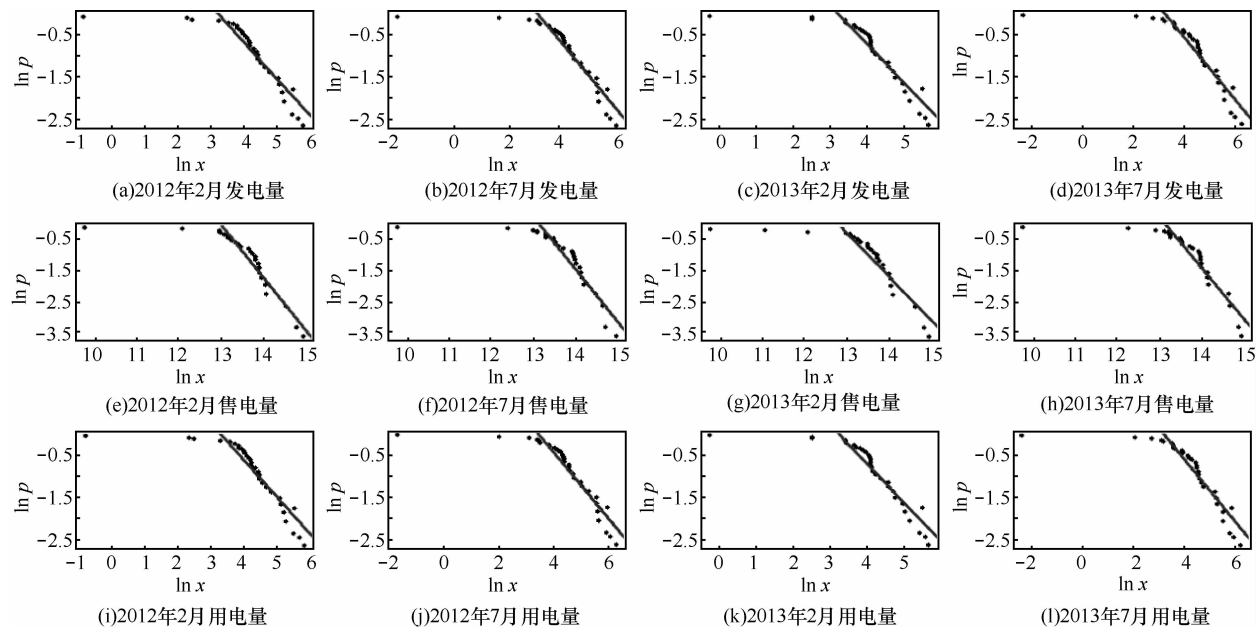


图 3 全国发电量、国家电网公司售电量和全社会用电量的幂律分布特征

表 5 全国发电量的幂律分布检验结果

	2011 年 2 月	2011 年 7 月	2012 年 2 月	2012 年 7 月	2013 年 2 月	2013 年 7 月
全国发电量	-0.871 85	-0.820 91	-0.859 67	-0.781 49	-0.930 68	-0.796 28
水电发电量	-0.345 71	-0.290 3	-0.354 26	-0.274 83	-0.344 59	-0.310 18
火电发电量	-0.791 58	-0.745 48	-0.780 58	-0.710 04	-0.844 04	-0.723 84

表 6 国家电网公司售电量的幂律分布检验结果

	2011 年 2 月	2011 年 7 月	2012 年 2 月	2012 年 7 月	2013 年 2 月	2013 年 7 月
公司售电量	-1.667 41	-1.583 4	-1.709 51	-1.780 75	-1.567 73	-1.733 2
大工业售电量	-1.557 84	-1.512 07	-1.531 38	-1.480 74	-1.361 44	-1.449 06
农业售电量	-1.557 84	-1.512 07	-1.531 38	-1.480 74	-1.361 44	-1.449 06
居民售电量	-0.898 27	-0.940 76	-0.925 42	-0.938 4	-0.913 02	-0.905 99
商业售电量	-1.021 38	-0.946 16	-1.020 74	-0.968 37	-0.944 24	-0.947 49

表 7 全社会用电量的幂律分布检验结果

	2011 年 2 月	2011 年 7 月	2012 年 2 月	2012 年 7 月	2013 年 2 月	2013 年 7 月
全社会用电量	-1.60 41	-1.494 71	-1.557 85	-1.530 02	-1.822 55	-1.523 87
第一产业用电量	-0.807 91	-0.801 95	-0.848 48	-0.780 38	-0.840 88	-0.868 36
第二产业用电量	-1.411 83	-1.405 64	-1.460 96	-1.445 14	-1.642 2	-1.413 73
第三产业用电量	-1.047 88	-0.998 7	-1.065 91	-1.003 08	-1.086 72	-1.030 81
城乡居民用电量	-0.893 28	-0.944 2	-0.904 32	-0.944 81	-0.953 62	-0.981 94

国用电量符合幂律分布的特性,与其他两个指标相比,用电量具有特殊的优良特性。

已知两个变量 X 和 Y 无论独立与否, $X+Y$ 的和仍然服从正态分布。若 X 和 Y 服从对数正态分布,则仍然服从对数正态分布,而 $X+Y$ 却不符合这一规律。通过分析,本文发现大部分细分行业的用电量服从对数正态分布,而细分行业用电量之和却不符合这一规律,比如全国用电量、第一产业用电量、第二产业用电量、第三产业用电量等。在第4节中,将详细阐述细分行业用电量服从对数正态分布的根本原因。

本文选择具有代表性的5个细分行业进行详细分析,分别是农业、轻工业、重工业、信息传输业、房地产、计算服务业。图4展示了5个行业用电量的经验累积分布函数(ECDF),应用极大似然估计的方法对分布参数进行拟合,并且将用电量与5个幂律分布进行对比,分别是指数分布、对数正态分布、逆高斯分布、帕累托分布和列维分布。由图可知,对数正态分布是最合理的拟合分布。

对数正态分布的概率密度函数为:

$$f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(\ln x - \mu)^2 / 2\sigma^2} \quad (1)$$

其中:

$$E(x) = e^{\mu + \sigma^2/2}, \text{VAR}(X) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2} \quad (2)$$

对数正态分布具有优秀的特性,当随机变量 $\log X$ 符

合正态分布时,则 X 符合对数正态分布。即对用电量进行对数转换,就可以得到正态分布的数据,因此证明用电量符合对数正态分布不仅大大减少了数据分析的工作量,还解决了用电量不是正态分布的难题。表8展示了对数转换后细分行业用电量的正态分布检验结果。

由表8可知,所有 W 统计量的检验结果大于0.05,即有95%的信心认为,细分行业的用电量经过对数转换后服从正态分布,可以直接应用以正态分布为必要条件的数据分析和挖掘方法。

4 原因分析

由上文已知,全国发电量、国家电网公司售电量和全国用电量符合幂律分布的特性,并且细分行业的用电量服从对数正态分布,本节主要结合业务背景阐述3个电力指标符合幂律分布以及细分行业用电量服从对数正态分布的根本原因。

3个电力指标符合幂律分布特性的主要原因如下。

- 由于政策、地理位置和自然环境等原因,全国经济发展严重不平衡,东部沿海地区较发达,西部等地区经济较为落后。电力和经济密不可分,是强正相关的关系,某地区的用电情况可以反映此地的经济状况,因此电力指标数据在地域上也存在不平衡的现象。
- 自然资源分布不平衡。全国的煤炭资源和水源分布

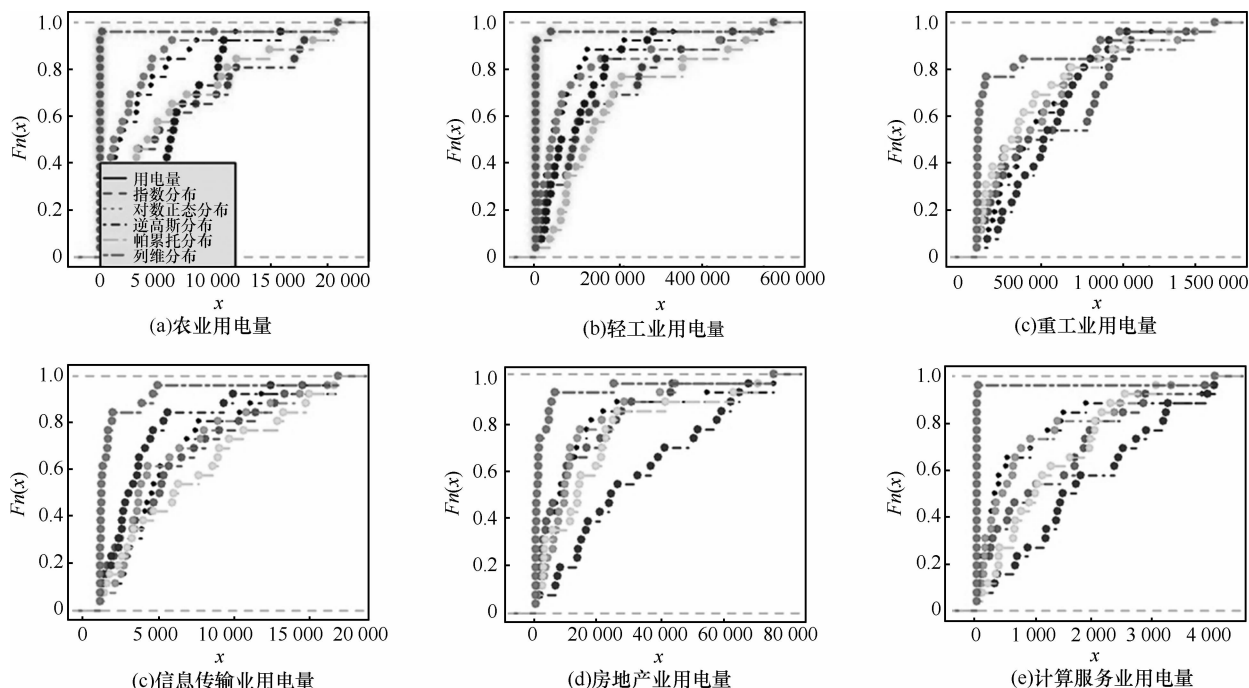


图4 细分行业用电量的ECDF



表 8 细分行行业用电量对数转换后的正态分布 W 统计量检验结果

	2011 年 2 月	2011 年 7 月	2012 年 2 月	2012 年 7 月	2013 年 2 月	2013 年 7 月
农业用电量	0.3702 86	0.722 518 6	0.894 372 9	0.743 408 6	0.831 565 3	0.714 749 6
轻工业用电量	0.918 809 3	0.947 205 5	0.948 705 1	0.945 678 4	0.647 804 6	0.912 885 3
重工业用电量	0.493 828 1	0.741 622 9	0.989 051	0.986 806	0.312 590 1	0.993 678 7
信息传输业用电量	0.702 192 8	0.685 757 3	0.808 931 3	0.511 452 1	0.500 611 7	0.666 026 5
房地产业用电量	0.702 654 5	0.858 882 1	0.849 534 7	0.836 866 8	0.883 541 3	0.836 742 8
计算机服务业用电量	0.863 242 5	0.399 651 5	0.717 818 8	0.783 987 6	0.635 606	0.675 023 1

不平衡,只有少数地区具有丰富的煤炭资源和水资源,这导致电力指标数据的曲线向左倾斜,并且具有幂律分布的特征。

因为电力指标数据在地域上分布不平衡,所以随着电力指标数值的增加, $P(Q>q)$ 逐渐降低,并且降低速度减缓,呈现幂函数的规律,这正是幂律分布的特征。

细分行行业用电量服从对数正态分布,下面运用数学理论阐述产生此种现象的根本原因。

正态分布的变量和对数正态分布的变量存在着区别和联系,就共同点来说,两种变量都是独立变量,即用户每天的用电量是独立的,符合正态分布。两种分布最大的区别在于“加”和“乘”,这种细微的不同使得数据的分布规律也具有明显差异。

首先以布朗运动为例来解释正态分布。假设某个行业在某一天的用电量为 Q_1 ,在下一天的用电量为 Q_2 ,在单位时间 Δt 内,行业用电量的差值为 ΔQ ,增加或者减少。令 Q_n 为 n 天以后某行业的累积用电量,则:

$$Q_n = Q_0 + \Delta Q (Q_{(0)} + Q_{(1)} + \cdots + Q_{(n-1)}) \quad (3)$$

其中:

$$Q_{(i)} = \begin{cases} 1, & Q_i - Q_{i-1} > 0 \\ -1, & Q_i - Q_{i-1} < 0 \end{cases} \quad (4)$$

$Q_{(i)}$ 相互独立,且 $P(Q_{(i)}=1)=P(Q_{(i)}=-1)=1/2$ 。 X_n 服从二项分布,参数为 n 和 $p=0.5$,由中心极限定理可知,当 $n \rightarrow +\infty$ 时, Q_n 近似服从正态分布。

但是在现实中,由于温度变化、某个月的假期时间长短的变化以及政策变化等原因,行业用电量的增减不是随机的,比如通过分析发现在2月和7月,细分行业的用电量更倾向于增加而不是减少,除非发生特殊情况。假设在单位时间 Δt 内,行业用电量以相同的概率增加 ΔQ_1 ,减少 ΔQ_2 , ΔQ_1 和 ΔQ_2 为已知常数,并且 $\Delta Q_1 > \Delta Q_2$ (在某些月份 $\Delta Q_1 < \Delta Q_2$,不影响最终结果)。

求解如下方程:

$$\begin{cases} Q_0 + \Delta Q_1 = Q_0 \Delta Q_m \\ Q_0 - \Delta Q_2 = Q_0 / \Delta Q_m \end{cases} \quad (5)$$

得到 $\Delta Q_m = \Delta Q_1 / \Delta Q_2$ 。则 n 天以后,行业的累积用电量为:

$$Q_n = Q_0 \prod_{i=1}^n (\Delta Q_m)^{Q_{(i)}} \quad (6)$$

其中, $Q_{(i)}$ 的含义保持不变,且 $P(Q_{(i)}=1)=P(Q_{(i)}=-1)=1/2$, $Q_{(i)}$ 相互独立,即行业在某一天用电时不会考虑下一天的用电情况。正如布朗运动中的分析, Q_n 仍然服从二项分布。由于 $\Delta Q_1 > \Delta Q_2$ (或者 $\Delta Q_1 < \Delta Q_2$),则分布曲线必定会向左偏。当 $n \rightarrow +\infty$ 时, Q_n 近似服从对数正态分布。这符合中心极限定理的多重乘法的规则,即多个独立同分布变量的乘积符合对数正态分布。因此,在同一时间段内(本文选取一个月),各个省的细分行行业用电量服从对数正态分布,且如果将细分行行业继续划分,仍然服从对数正态分布。

5 结束语

本文主要针对电力指标的分布性规律进行分析研究。经过统计分析研究发现,90%以上的电力指标不符合正态分布。经验证,全国发电量、国家电网公司售电量和全社会用电量3个电力核心指标具有幂律分布的特性,其中细分行业的用电量符合对数正态分布,并且具体阐述了产生这种现象的业务原因和数学原因。

本文对电力指标的分布规律进行了深入研究,为电力大数据的探索性分析做出了贡献,并且为数据分析和挖掘方法的选择提供了理论支撑。所得的研究成果能够为电力数据分析同行提供更多的创新思路,方法和技术也可以应用到其他电力方面的数据,不仅丰富了电力数据分析领域的技术,对很多业务问题也具有理论指导意义。不符合正态分布的电力指标数据看似毫无规律,幂律分布以简洁优雅的形式,将毫不相干的事物联系在了一起。以后也将是电力数据分析中研究的热点,未来也将对电力行业中幂律分布的产生机制进行深入研究。

(下转第121页)

软件、移动设备防病毒软件、主账号口令及移动办公使用管理制度等多重维度的手段加固系统安全,有效地保证了信息安全。

(2) 体验性

中电投集团移动办公系统以用户体验为中心进行设计,无论技术路线、产品选型或是实施方案,都以增强系统用户体验为最终目的。在设计及实施过程中摸索形成的集 VPN 连接、证书身份认证、主账号密码认证、特征码匹配认证、企业门户及公文流转系统单点登录等功能于一体的应用程序设计思路,最大程度上解决了用户体验与安全性之间的矛盾性,即在满足高安全性的前提下简化用户操作步骤,易于上手使用。在系统功能及 UI 设计上,也以用户使用方便、简洁明了及响应速度快为最优考虑因素,并全面支持了目前市场主流的手机及平板电脑,系统自正式运行起,便获得了用户的一致好评。

(3) 扩展性

中电投集团移动办公系统以扩展性为系统架构设计的核心要素,将来不仅可以在移动平台中部署其他应用系统,以满足用户更深层次的办公需求,而且系统本身也可以快速适应 iOS 及 Android 系统的新版本,自动适配尺寸更大、分辨率更高的新移动终端,且提出了应用接入规范,

制定了持续完善滚动更新的机制,确保了中电投集团移动办公系统在未来仍然可以拥有较好的兼容性和先进性。

参考文献

- 1 陈静. 移动办公与管理. 北京: 对外经济贸易大学出版社, 2012
- 2 冯登国. PKI 技术. 北京: 科学出版社, 2008
- 3 苹果公司. iPad 企业应用, 2011

[作者简介]



刘铮, 男, 中国电能成套设备有限公司工程师、信息化主管, 主要研究方向为信息化系统建设与管理。



王静, 女, 中国电力投资集团公司高级工程师、处长, 主要研究方向为大型发电集团信息化管理与建设, 包括总体规划、方案设计、应用系统开发和部署、基础设施建设、运维管理、安全体系建设等。

(收稿日期: 2013-08-01)

(上接第 114 页)

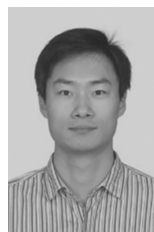
参考文献

- 1 Rhee I, Shin M, Hong S, *et al.* On the levy walk nature of human mobility. *Networking*, IEEE/ACM Transactions on, 2011(3)
- 2 Yang B, Duan W Q, Chen Z. New method to estimate scaling of power-law degree distribution and hierarchical networks. *Applied Mathematics and Mechanics*, 2006, 27(11): 1475~1479
- 3 葛玉伟, 周战, 陈万田. 洪泽地区配电网故障幂律分布特性. *现代电子技术*, 2011, 34(6): 172~174
- 4 Chris A. The Long Tail. CITIC Publishing House, 2005

[作者简介]



王冠男, 女, 现就职于北京国电通网络技术有限公司, 主要从事大数据分析和挖掘工作。



邓春宇, 男, 北京国电通网络技术有限公司工程师、部门副总经理, 主要从事大数据在电力行业中的应用与实践研究。



赵悦, 女, 现就职于北京国电通网络技术有限公司, 主要从事大数据分析和挖掘工作。

郑亚芹, 女, 现就职于北京国电通网络技术有限公司, 主要从事大数据分析和挖掘工作。

(收稿日期: 2013-08-01)