数据创新在政府统计中的应用

国家统计局 科研所
姜 澍

中国政府统计工作的现状

- 主要统计工作
- 统计调查方法（普查、抽样调查和全面定期统计报表等统计调查等方法）
- 统计采集方式（联网直报、电子采价器、住户记账器等电子终端、计算机辅助电话调查（CATI）系统）
- 统计数据发布

政府统计的七个流程环节

制度设计——数据采集——数据处理——数据存储——数据质量评估——数据发布与传播——数据分析

大数据是政府统计数据创新的途径之一

- **大数据的特征 6个V:**
- **Volume： 数据量大**
- **Variety： 数据类型多**
- **Velocity：处理速度快**
- **Value： 应用价值大**
- **Vender： 获取与发送灵活**
- **Veracity：真实准确性**

联合国大数据全球工作组

重点关注以下的几类数据：

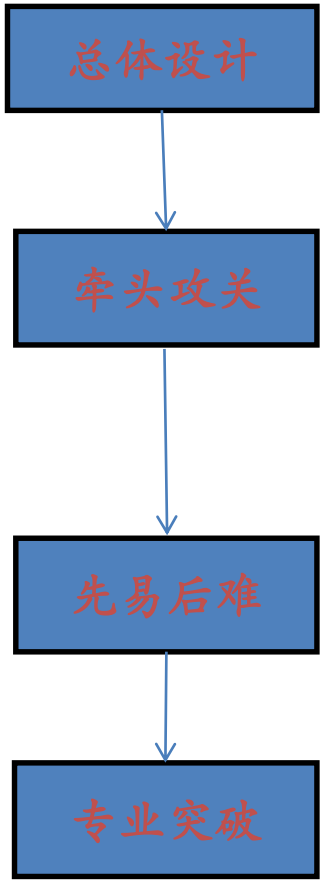
- 手机数据——人口、旅游统计
- 卫星图像和地理空间数据——农业统计等
- 扫描数据——价格统计
- 网络社交媒体数据——情绪、舆情等调查

2019年联合国第五届大数据卢旺达会议

联合国全球平台
UN Global Platform

- 分享数据、方法和技术服务
- 政府统计、企业、学术界work together, learn together

总体原则和主要目标任务



✓以互联网数据作为首要突破口

✓电商企业和数据搜集企业掌握的海量数据；通过签订战略合作协议等方式，将上述企业纳入政府统计框架；——与17家企业签订合作协议

✓与相关部门积极沟通，通过数据交换等方式，大力推动部门行政记录资料共享;——落实合作协议

✓CPI统计、贸易统计、移动手机信息、经济预测——以项目研究为依托

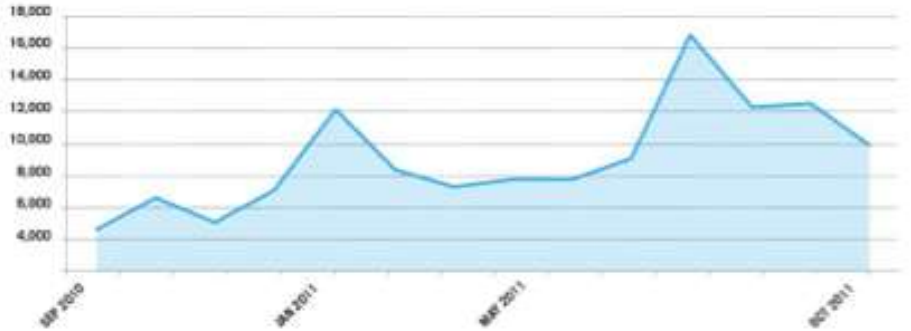
主要目标任务：

- 构建大数据统计标准和方法体系
- 扩大政府专业统计的数据源
- 改革统计调查方法和数据采集手段
- 提高数据评估、数据挖掘和数据发布水平

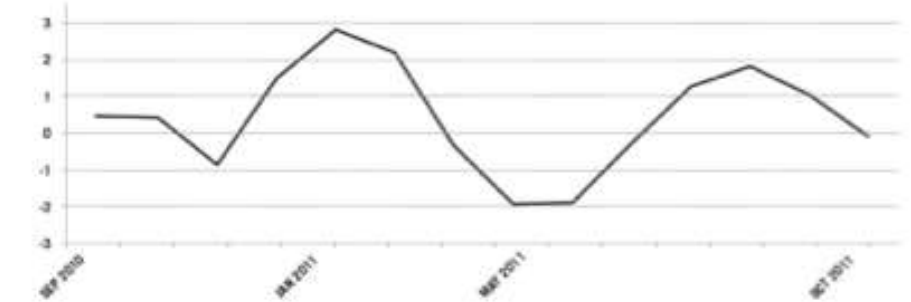
国内外政府统计基于大数据的一些实践

联合国基于Twitter数据对价格的调查

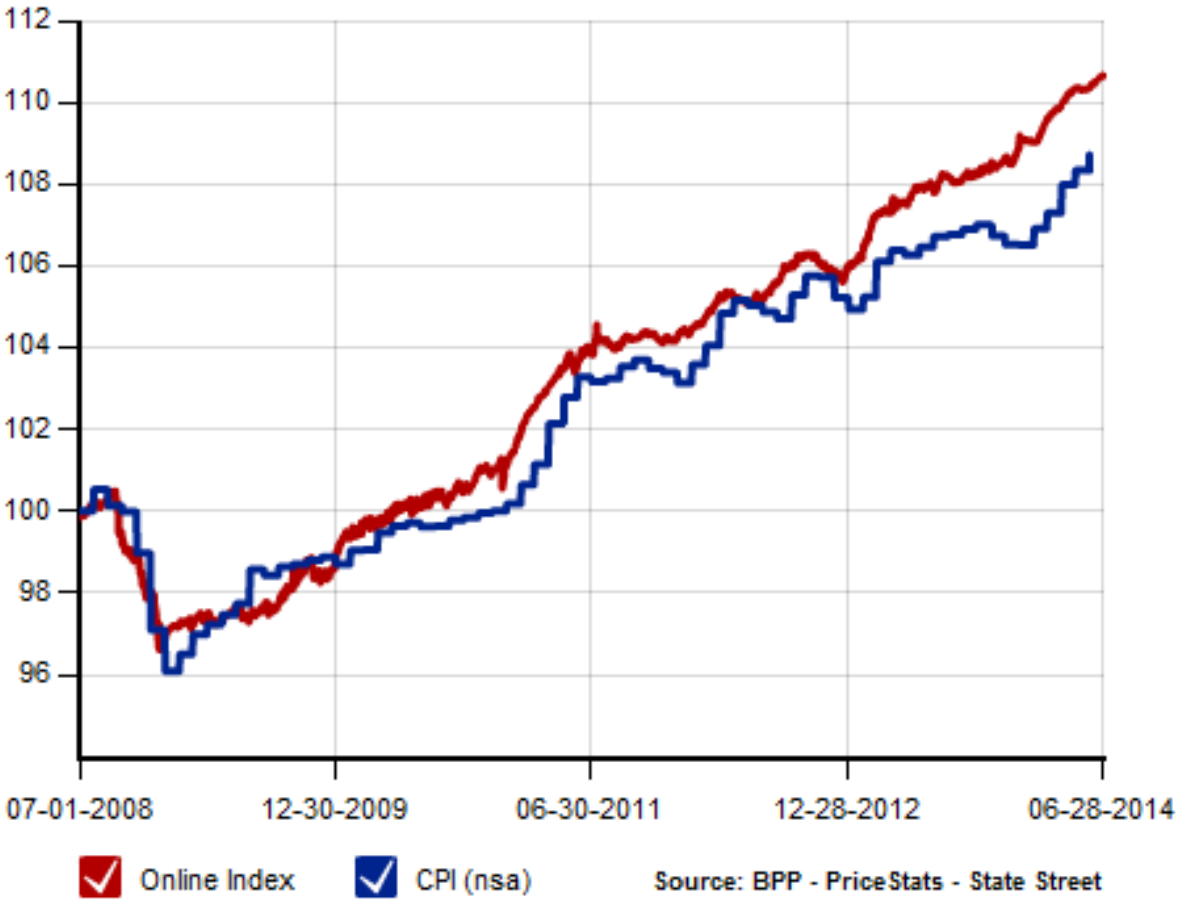
Tweets about the price of rice
(per month)



Food Price Inflation



每日网上价格指数 Daily Online Price Index



英国、荷兰、挪威等利用网络抓取数据编制CPI

图·英国食品类 CPI 价格指数和 CLIP 价格指数比较

英国利用大数据测算英国数字经济

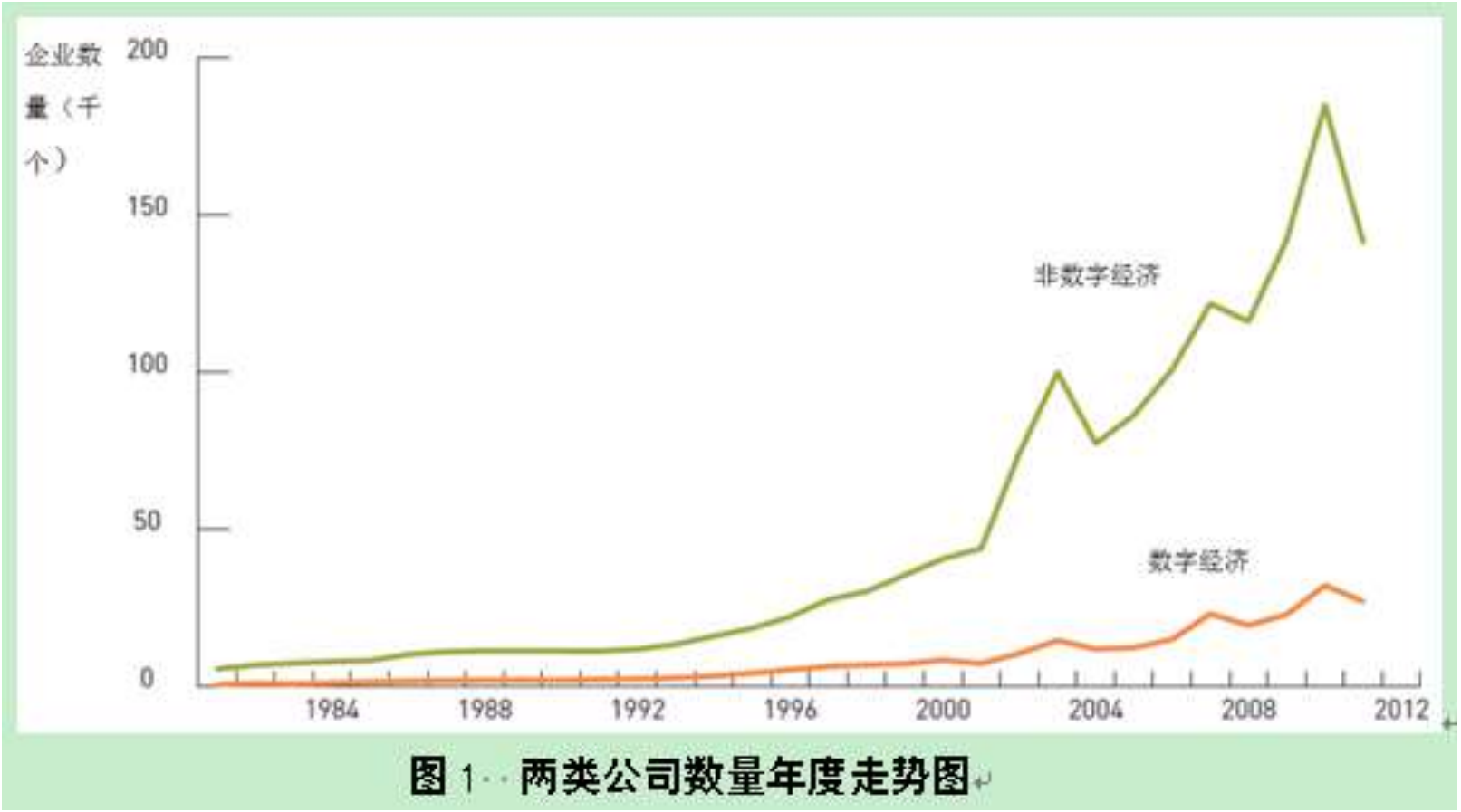
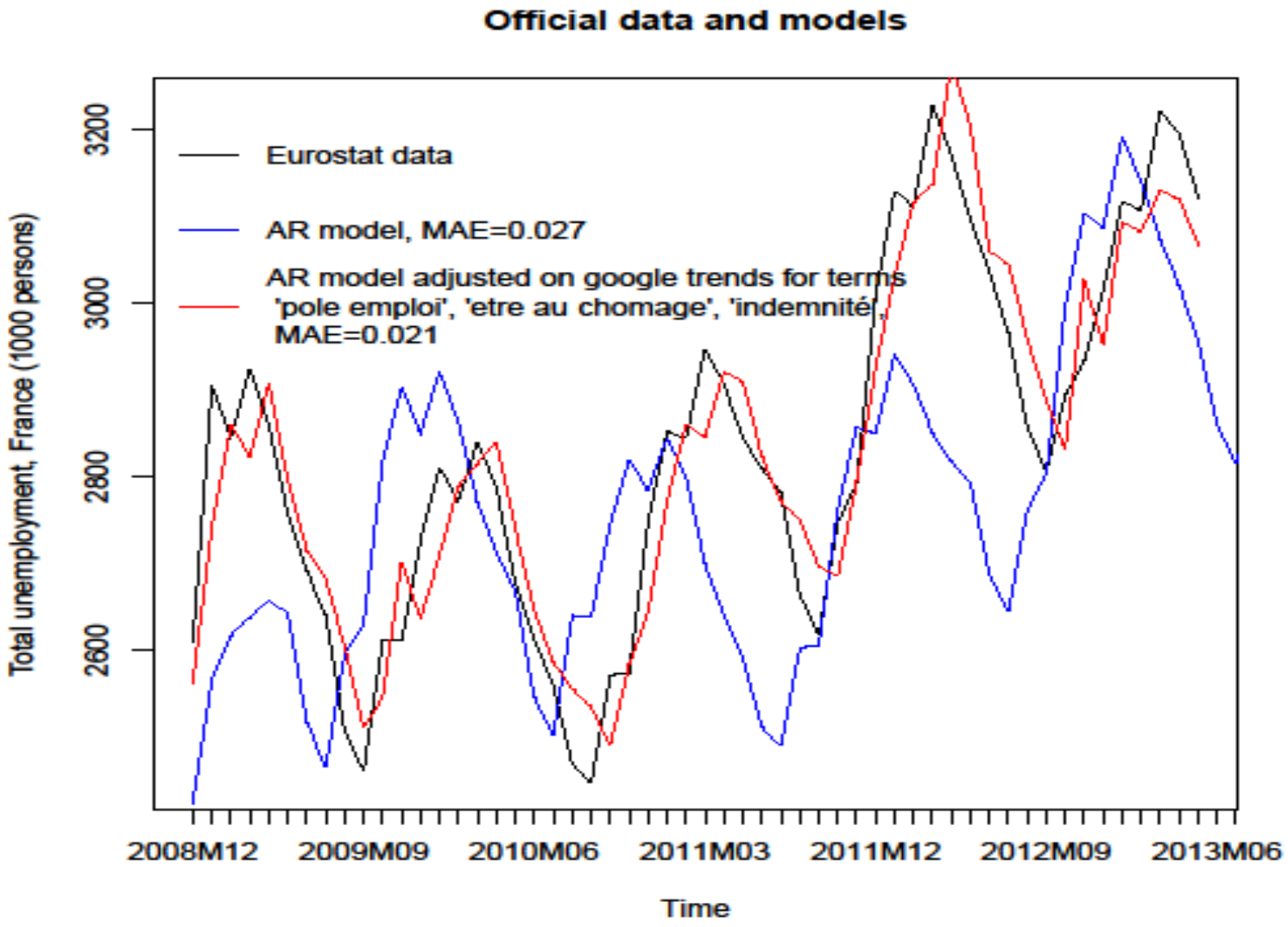


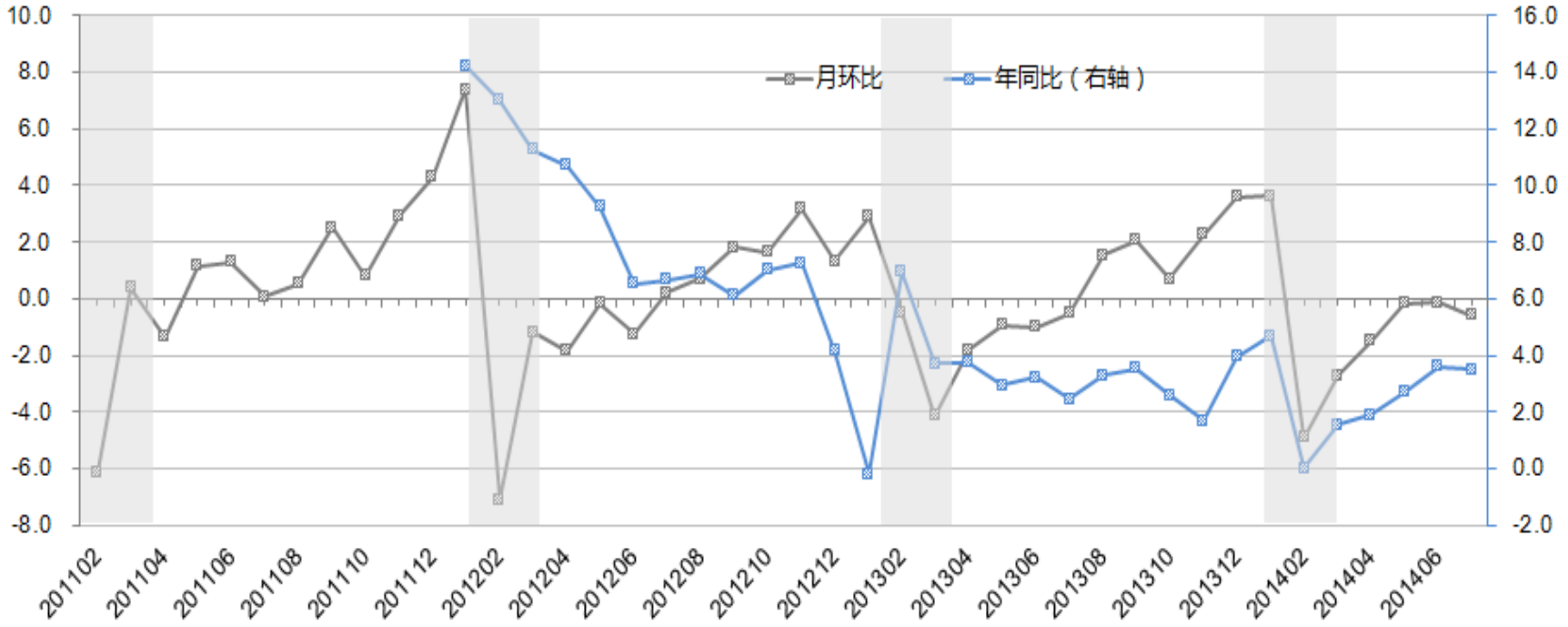
图 1 · · 两类公司数量年度走势图

欧盟利用谷歌搜索数据预测失业

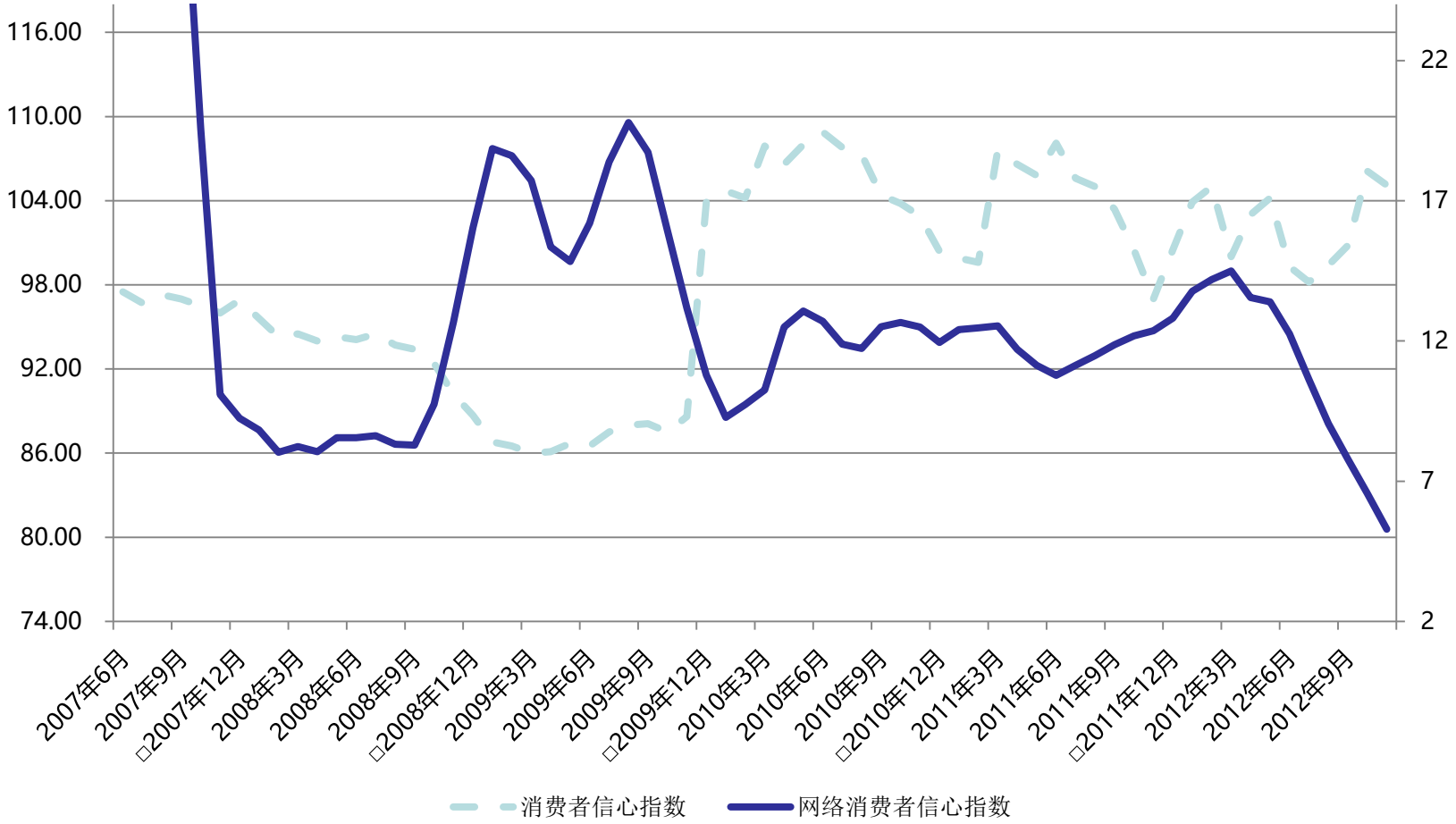


财新智库基于大数据建立新经济指数（NEI）

阿里巴巴网购全网商品价格指数 (alibaba Shopping Price Index,aSPI)



基于百度搜索数据的消费者信心指数（中科院）



利用手机移动信号进行人口统计
（北京市统计局与北京移动合作）

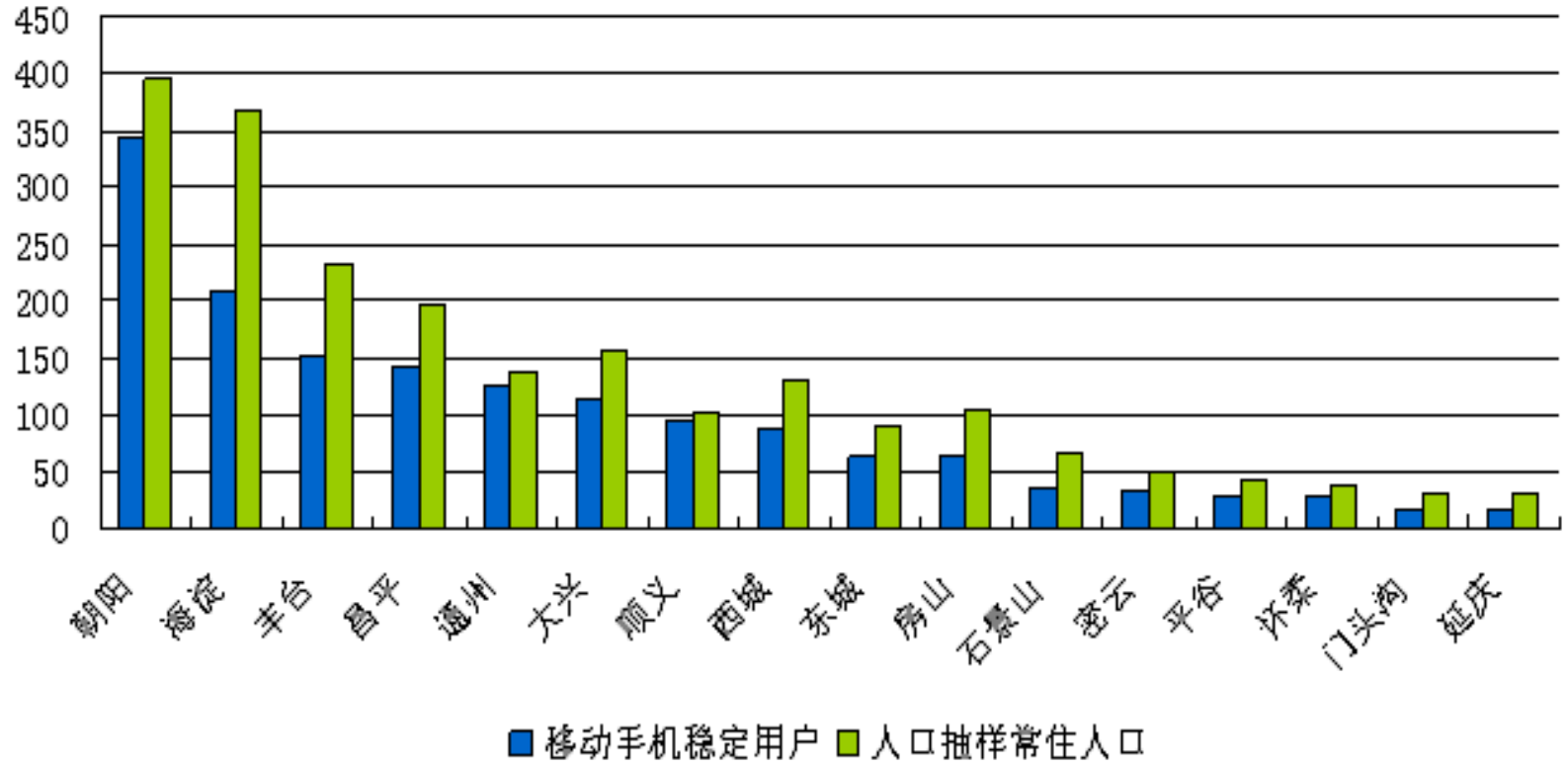
——季度常驻人口总量

——测算住职分离情况

——疏解效果情况

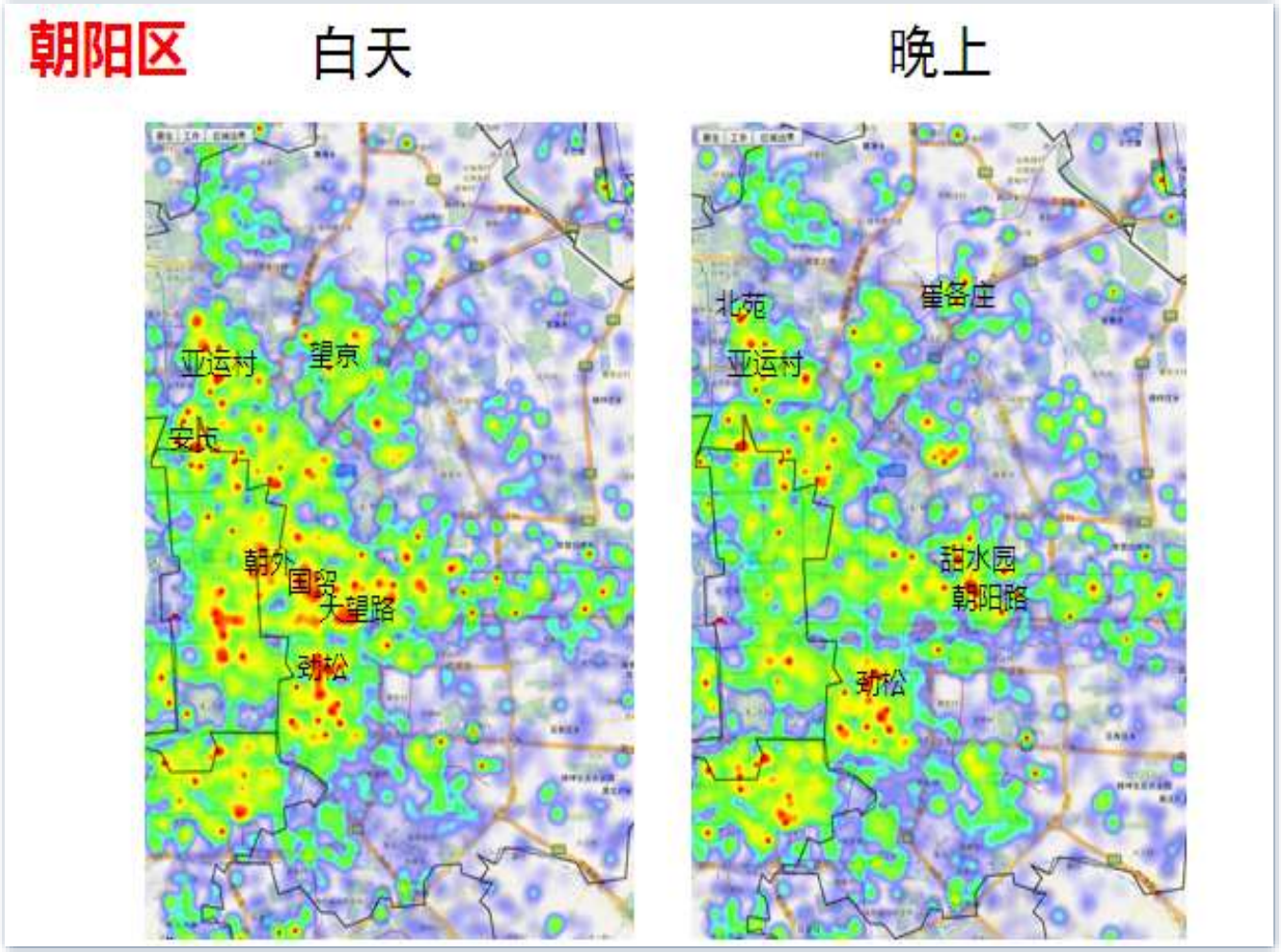
表1：各区移动手机稳定用户数与人口抽样调查常住人口数对比

单位：万人



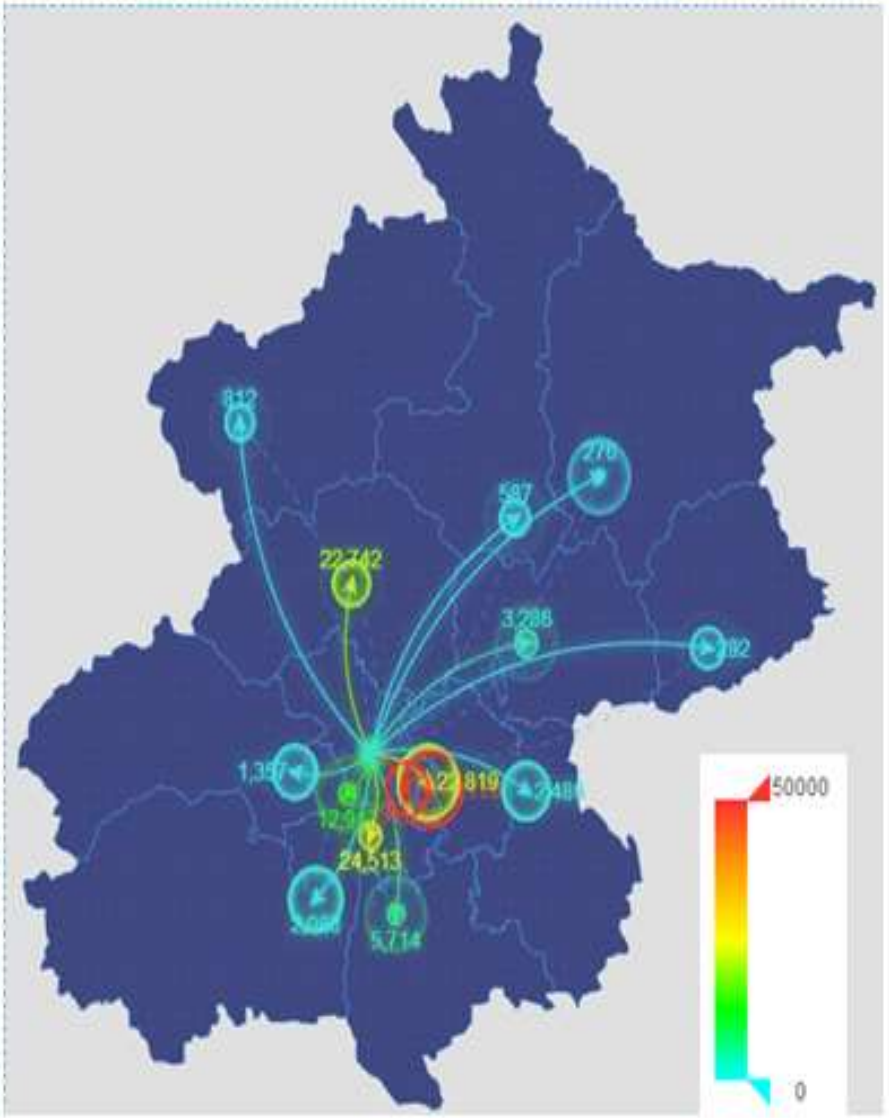
移动手机稳定用户数量与各区常住人口数不尽相同，但各区比重排名大体一致。

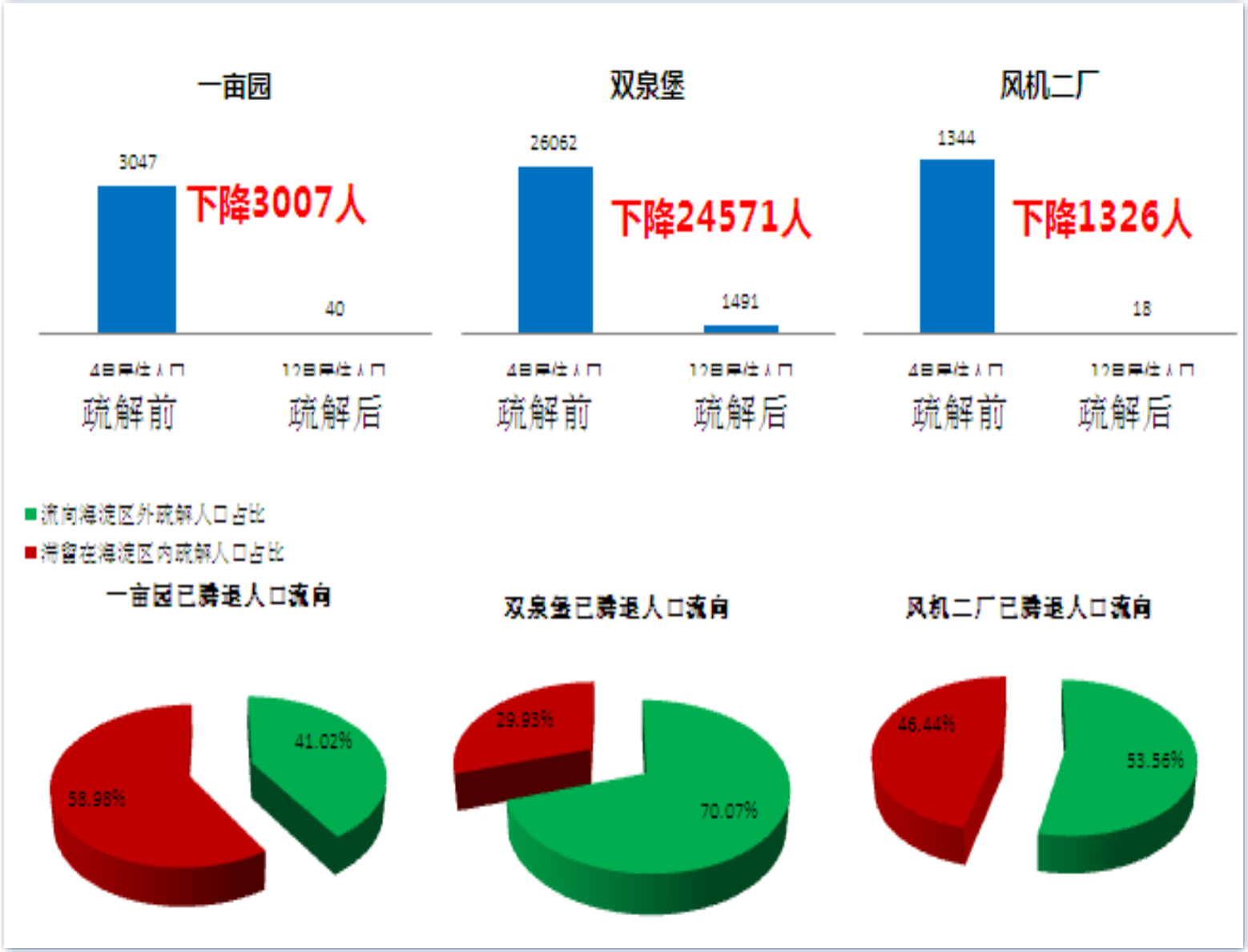
白天夜晚人群分布



海淀区夜晚用户白天通勤目的地

居住地	工作地	人数	占比
海淀	朝阳	72729	30.73%
海淀	西城	64071	27.07%
海淀	丰台	24513	10.36%
海淀	东城	22819	9.64%
海淀	昌平	22742	9.61%
海淀	石景山	12917	5.46%
海淀	大兴	5714	2.41%
海淀	顺义	3286	1.39%
海淀	通州	2486	1.05%
海淀	房山	2088	0.88%
海淀	门头沟	1357	0.57%
海淀	延庆	812	0.34%
海淀	怀柔	587	0.25%
海淀	平谷	292	0.12%
海淀	密云	276	0.12%
合计		236689	100.00%




一亩园
3047
下降3007人
40
疏解前
疏解后
双泉堡
26062
下降24571人
1491
疏解前
疏解后
风机二厂
1344
下降1326人
18
疏解前
疏解后
流向海淀区外疏解人口占比
滞留在海淀区内疏解人口占比
一亩园已劳退人口流向
41.02%
58.98%
双泉堡已劳退人口流向
29.93%
70.07%
风机二厂已劳退人口流向
46.44%
53.56%

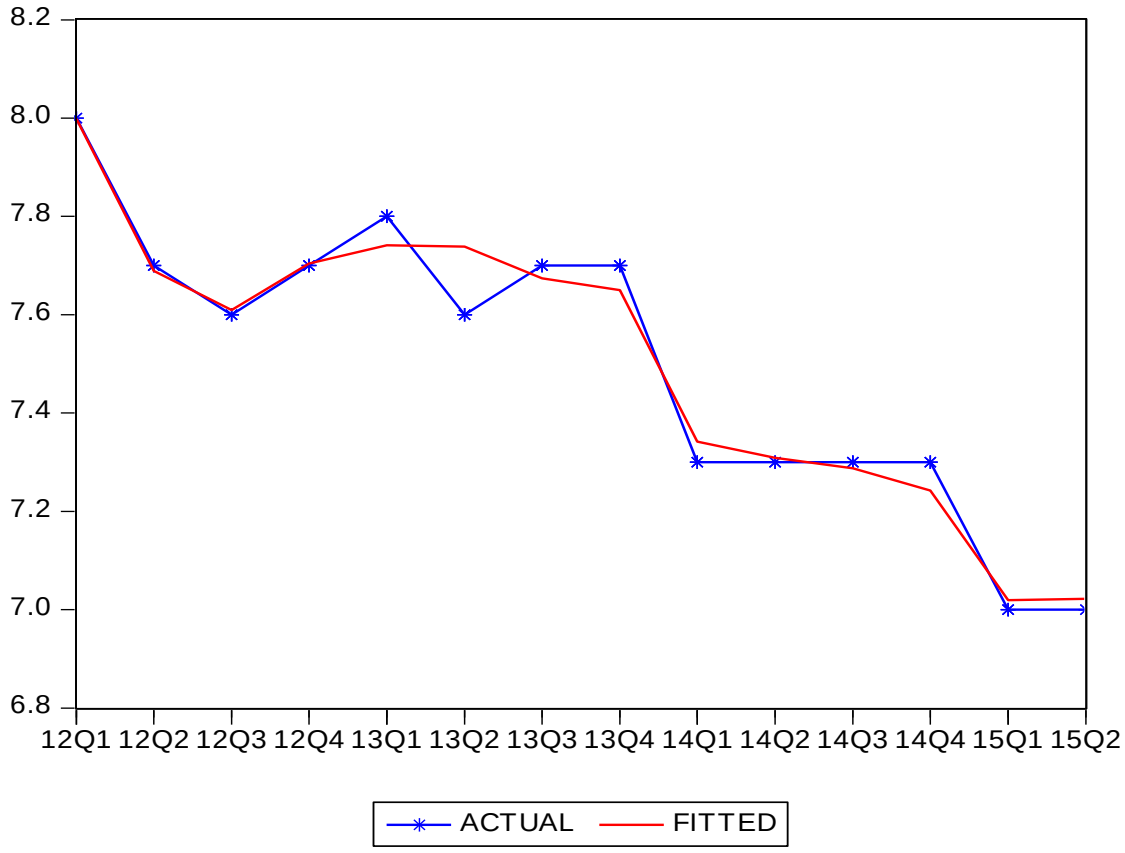
案例分析：经济雷达、新经济就业、和就业指数？

案例1: 利用百度搜索数据预测

基本方法

- 首先，遴选与季度GDP相关的关键词；
- 然后，对关键词的搜索量数据进行整合，计算出新变量的增速；
- 再对整合之后的变量，结合季度GDP累计同比增速数据，进行模型估计，选出预测效果和稳定性较好的模型；
- 最后根据遴选出的模型和新的搜索数据，对季度GDP累计同比增长速度进行预测。

考虑到变量之间可能存在的非线性关系，建立如下的自回归滑动平均模型（ARMA模型）



用百度公司的数据测算了分析了2011年第1季度至2015年第2季度的模拟预测分析。

优点：

拟合度比较理想

系数通过置信度检验

残差序列在滞后12阶之内不存在统计意义上的自相关性和偏自相关性

缺点：

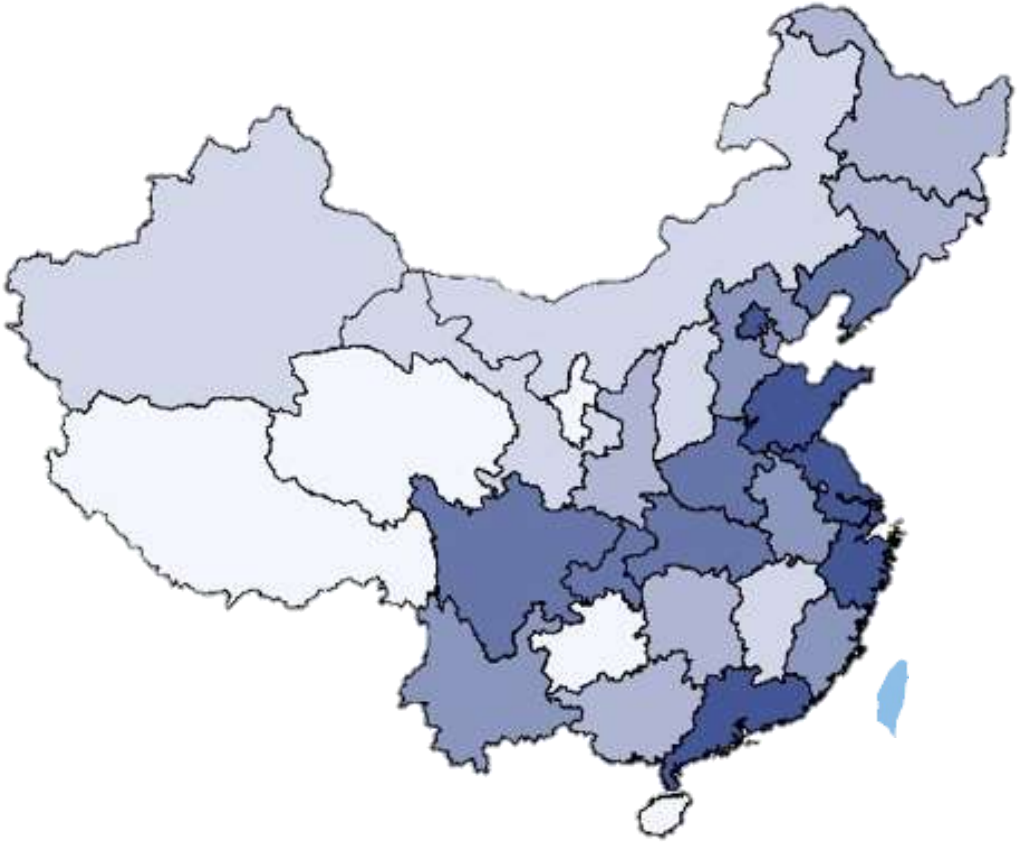
166个关键词，仅占所有288个关键词的57.64%

案例2: 新经济就业的识别和稳定性分析

中国需要积累新经济数据

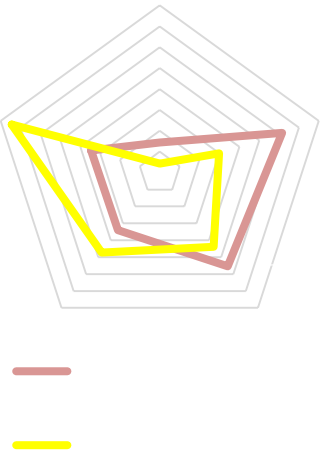
David Dollar（美国财政部前驻中国经济金融特使）

新经济就业人员地域分布



新经济就业人员年龄分布

新经济就业人员学历分布



新经济就业稳定性分析


新经济就业稳定性弱于旧经济就业，但劳动力结构在优化
城镇劳动力占比不断增加，女性和年长者在新经济部门就业中依然处于弱势地位
26-35周岁青年劳动力增加趋势明显

案例3: 判断就业景气程度

- 拟建立环比指数（重点关注的行业、主要区域）
- 用相关系数与调查失业率作验证
- 使用回归模型作就业景气预测
- 引入机器学习模型不断优化学习结果
- 持续跟踪、监测与验证结果

抓取三家网站的分职业（52个行业）、分地区每周职位招聘数量：

- 51Job
- 看准网
- 智联招聘

标准问题：职业分类不尽相同

分地区方面，有的是按省，有的是按大城市

数据问题：存在断层问题

对各专业统计应用大数据的畅想？

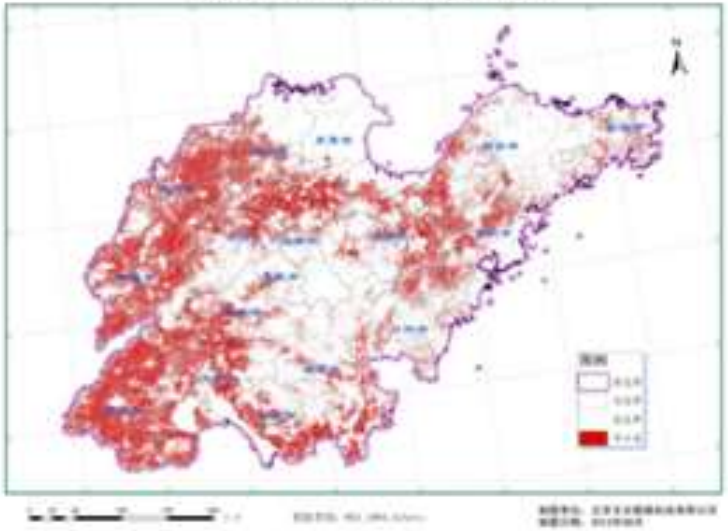
贸经统计

- 结合京东等重点网上零售交易平台，测算全国网上零售额
- 与中国银联合作，验证社会消费品零售额数据
- 开展网购用户网购行为专项调查—网购替代率
- 将网上零售额中未在库单位实物商品网上零售额纳入社会消费品零售额统计中

农业统计

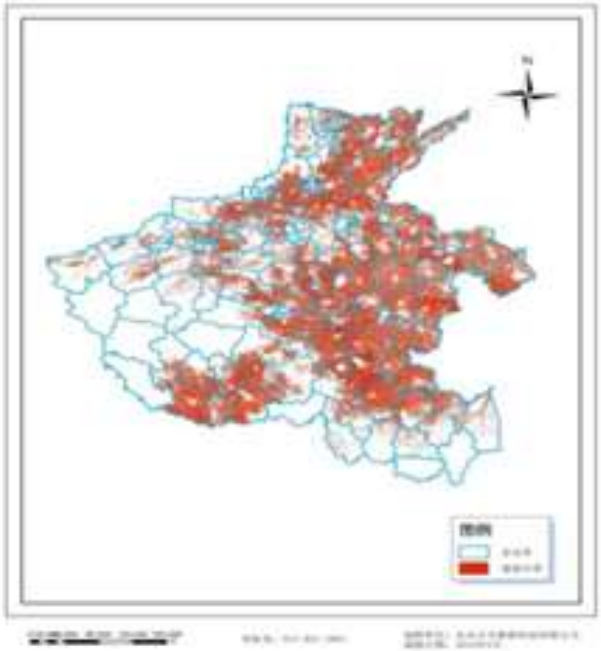
- 已在河北、江苏、河南等开展了小麦播种面积遥感测量
- 在吉林、辽宁等5省利用卫星、航空等高分影像构建空间抽样框
- 全国农业普查中，使用国产自主卫星资源及无人机技术对主要农作物面积进行遥感测量

山东省冬小麦空间分布图

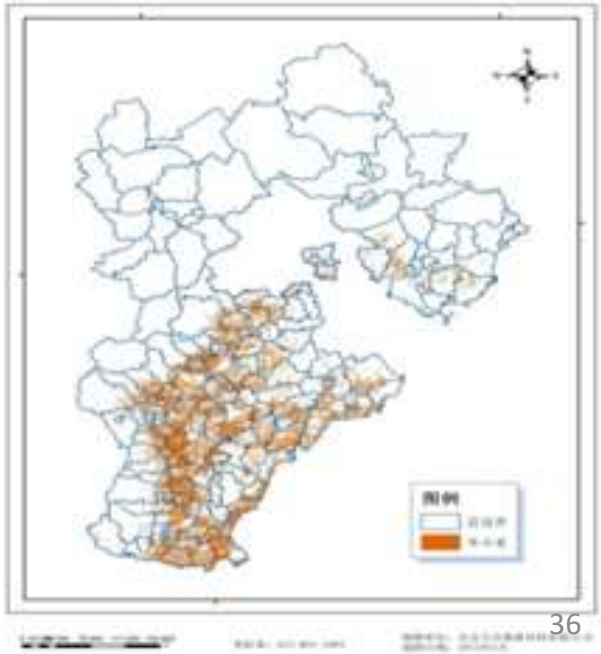


湖北省冬小麦空间分布图

河南冬小麦空间分布图



河北冬小麦空间分布图



交通运输统计

- 正在研究应用射频传感器技术对公路货运量及经济活动景气状况分析监测。
- 在河南、山西两省开展公路货运统计中应用大数据的可行性调研，摸清了高速公路大数据的采集、传输、存储等生产过程，整理生成了出口车道数据库字段名（指标）清单。

投资统计

- 每月月初收集三一重工、中联重科、徐工机械和柳工集团等四家施工机械制造企业已售设备开工率数据，计算出综合开工率和月平均工作时间两个指标
- 积极倡导“利用遥感技术动态核实投资项目”
- 与搜房网合作，利用其新房、二手房关注量等非结构化数据，以及土地交易面积、土地出让金等土地交易数据，评价房地产开发月度投资数据

人口普查和流动人口统计

案例：

北京市东城区根据城区特点实施了万米单元网格管理，运用网格地图的技术思想，将东城区所辖25.38平方公里划分为1652个网格单元，由城市管理监督员对所分管的万米单元实施全时段监控。城市网格化精准管理数据库中，拥有单元网格内的家户成员信息，具体包括姓名、性别、年龄、身体状况和户成员关系、身份证号和家庭住址等，从而为人口普查和提供了重要的信息源。

价格统计——CPI统计

- 试点手工采集和网络抓取部分商品的网络价格并测算价格指数
- 网络采价及商场超市采价
- 使用来自1号店指数等大数据企业数据，对国家CPI数据进行校验


6.00
4.00
2.00
0.00
-2.00
-4.00
-6.00
-8.00
201404
201409
201502
201507
201512
201605
1号店指数（环比）
CPI（环比）

价格统计——房地产价格统计

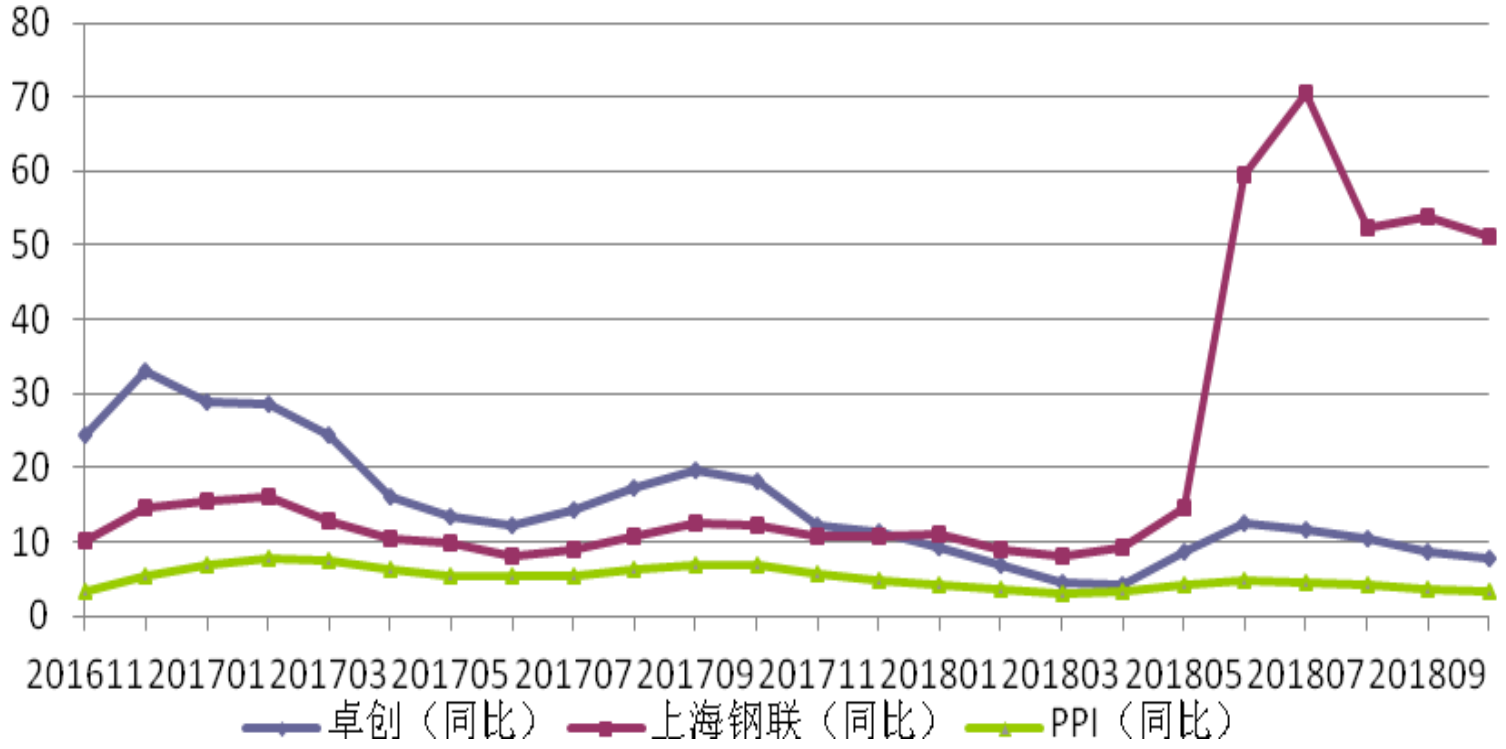
- 已在全国70个大中城市中使用网签数据计算新建住宅价格指数

- 与搜房网合作，利用其新房、二手房关注量等非结构化数据，评价房地产开发月度价格数据

价格统计——生产资料价格统计

- 与山东卓创公司开展合作，开展流通领域重要生产资料市场价格监测

- 按月搜集上海钢联大宗商品价格指数对工业出厂价格指数进行校验


80
70
60
50
40
30
20
10
0
201611 201701 201703 201705 201707 201709 201711 201801 201803 201805 201807 201809
卓创（同比）
上海钢联（同比）
PPI（同比）

价格统计——服务业价格统计

积极探索使用八爪鱼等信息抓取技术进行网络电子采价，现正积极探索使用八爪鱼等信息抓取技术进行网络电子采价，充分利用CPI现有统计成果，2013年以来按季度陆续试算了交通运输和邮政业、电信业、住宿和餐饮业分类价格指数。

价格统计——能源统计

自2009年1月起，开始收集秦皇岛港煤炭转运月度数据，包括煤炭吞吐量、港存煤及煤炭交易价格数据。用于评估能源供应及消费的进度数据。

价格统计——部门行政记录统计

- 利用部门行政记录大数据核算居民消费支出，试编全国资产负债表，以及在地区GDP统一核算中探索利用大数据评估基础数据质量。
- 积极推进统计与政府部门在电子化行政记录和统计信息的共享。在居民收入、劳动工资等统计中，通过对税务、银行等部门行政记录的分析，评估居民收入、工资、家庭资产等统计数据，特别是评估与校验高收入阶层的相关数据。

价格统计——经济预测

●与百度合作，建立相关搜索词库；

●建立相关统计分析和计量模型，辅助判断经济走势。

大数据企业统计指标月报

国际统计局-百度大数据“经济雷达”项目

●对《2013年国民经济和社会发展统计公报》的统计指标关键词搜索热度分析

●每旬对几个关键词的搜索热度动态监测，包括统计指标和重点行业

●主要经济先行指标与百度数据库关键词的匹配分析，用于判断经济走势

- **大数据真的可靠吗？**

从Google对流感趋势的预测效果开始变差说起……

大数据具有局限性：
对计量的具体内容和计算方法过于倚重

思考：数据最有价值Vs数据最容易计量

名人名言：计算得清楚的东西未必都重要，重要的东西也未必都计算的清楚。（埃尔伯特˙爱因斯坦）

应有的态度——适度谦虚，避免大数据自大症

- 政府统计是大数据的拥护者，持有条件的支持立场。
- 应该充分认识到，大数据不能取代政府统计，但是大数据可以补充和验证政府统计。
- 政府统计应更加关注大数据标准的制定。
- 需要多方合作，包括与企业、学术界。
- 依靠云计算技术进行大数据存储和计算，需要搭建类似大数据平台等应用环境。

推动“统计云”项目建设，高效整合统计业务数据、部门统计数据与大数据

一是在不改变原有业务系统的前提下，研究探索将各专业数据、普查数据、空间地理信息数据整合放到统计云上；

二是在平台上有效整合部门电子化行政记录，积极推进部门数据共享，实现数据的互联互通；

三是在云平台上对相关专业领域大数据进行收集、挖掘与分析，为政府统计提供及时有益的补充和验证，推动大数据在政府统计中的应用。

积极利用行政记录，改进现有统计方法

充分利用“五证合一、一照一码”登记制度改革的重大机遇，尽快完成统计登记模式向“五证合一、一照一码”登记模式的过渡，充分利用部门电子化单位登记资料，建立并更新基本单位名录库，推进各部门之间电子化行政记录和统计信息在专业领域的应用，力争在部门大数据应用方面取得实质性进展。

积极利用网上电子化数据，补充现有统计数据源

一方面，要与大数据企业开展实质性合作，使互联网交易信息和电子化企业生产经营资料等大数据成为政府专业统计数据源的组成部分。同时，与部分数据搜索挖掘企业开展战略合作，借助企业丰富的大数据资源和先进技术，共同开发利用大数据。另一方面，要探索改进现有数据采集手段，完善“直报+搜索”的数据采集方式。

政府统计数据创新的最大挑战

数据可获得性不高，难以接触到数据源
技术能力有限
缺少相应的人才

政府统计的创新需要加强与各方合作

谢 谢!