**飞凤可视化大赛**

**说明文档**

团队名称 团队名字不能为空！

指导教师 李扬

2018年 8 月 10 日

目录

[1、摘要 3](#_Toc521879080)

[2、数据 3](#_Toc521879081)

[2.1 环境污染数据的获取 3](#_Toc521879082)

[2.2 产业相关数据的获取 4](#_Toc521879083)

[2.3 环境污染源的分析 5](#_Toc521879084)

[3、产品 6](#_Toc521879085)

[3.1 使用工具 6](#_Toc521879086)

[4、结果 7](#_Toc521879087)

[5、展望 8](#_Toc521879088)

[6、团队 9](#_Toc521879089)

## 摘要

本作品针对无锡市公开数据和企业数据进行分析，对企业经营和环境治理关系进行分析。不同类型企业的经营，对不同的环境指标都会有相应的影响，我们挖掘了公司、企业的行业类型，经营范围和环境指标之间的关系，从而找出不同污染源的来源公司、行业以及经营内容。通过已有的数据，根据本年度的天气状况，企业经营状况，对未来一年可能发生的环境污染做出预警。

## 数据

|  |  |  |
| --- | --- | --- |
| 目标层 | 准则层 | 指标层 |
| 无锡市产业结构指标体系 | 产业产值结构 B1 | 第一产业产值占 GDP比重 |
| 第二产业产值占 GDP比重 |
| 第三产业产值占 GDP比重 |
| 产业就业结构 B2 | 第一产业就业人口占总就业人口比重 |
| 第二产业就业人口占总就业人口比重 |
| 第三产业就业人口占总就业人口比重 |

### 2.1 环境污染数据的获取

我们通过无锡市公开数据2.xlsx中的数据，获取了关于环境污染的数据。在这份数据集中，我们进行了以下处理：

1. 根据环境指标列的信息量(熵)，选取变化、波动较大的环境指标，其中，熵的步骤为：
2. 先对数据进行标准化处理，此处采用功效系数法，计算公式为：

其中，i表示年份，j表示污染指数， 表示i年第j类环境指标d的原始值，表示标准化后的值， 分别表示第j类环境污染指标的最小值和最大值。

1. 计算第j类环境污染的熵值，具体计算公式如下：

其中，且

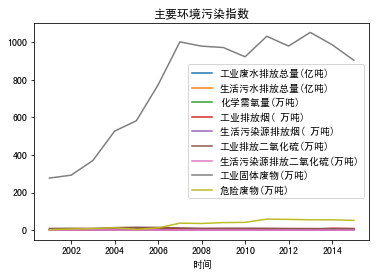
1. 选取环境指标中实际上包含年份较多的列(某些列部分年份过少)
2. 选取常识、语义上比较适合的数据(排除掉处理量、二次利用量等数据，并在废水、废气、固体废物污染指标下各选取工业和生活污染指标)
3. 选取得到：

‘工业废水排放总量(亿吨)','生活污水排放总量(亿吨)','化学需氧量(万吨)','工业排放烟( 万吨)','生活污染源排放烟( 万吨)','工业排放二氧化硫(万吨)','生活污染源排放二氧化硫(万吨)','工业固体废物(万吨)','危险废物(万吨)'

|  |  |  |
| --- | --- | --- |
| 目标层 | 准则层 | 指标层 |
| 无锡市生态环境评价指标体系 | 生态环境压力 | 工业废水排放总量(亿吨) |
| 生活污水排放总量(亿吨) |
| 化学需氧量(万吨) |
| 工业排放烟( 万吨) |
| 生活污染源排放烟( 万吨) |
| 工业排放二氧化硫(万吨) |
| 生活污染源排放二氧化硫(万吨) |
| 工业固体废物(万吨) |
| 危险废物(万吨) |
| 生态环境状态 | 生态环境状况指数(无锡市区) |
| 生态环境响应 | 综合利用工业固体废物(万吨) |
| 综合利用危险废物(万吨) |
| 一般工业固体废物处置量(万吨) |
| 危险废物处置量(万吨) |
| 医疗废物处置量(万吨) |

等列作为特征列，进行下一步的操作

1. 由于化学需氧量这一列存在着2001年到2004年的缺失数据，我们根据线性模型对其进行了补全
2. 对以上的数据进行整合，得到主要环境污染.csv(详情见附件：无锡市环境分析中间数据.zip)



### 2.2 产业相关数据的获取

我们通过政务脱敏数据-人力资源-公司基本信息获取了关于不同行业与年份之间的信息，具体如下：

1、 清洗年份不合适的数据（低于1990年和高于2018年的数据）

2、 对所有的产业类型进行排序

3、 通过百度百科对所有的产业类型进行第一第二第三产业的分配

4、 按照年份统计每一年第一第二第三产业总的就业人数

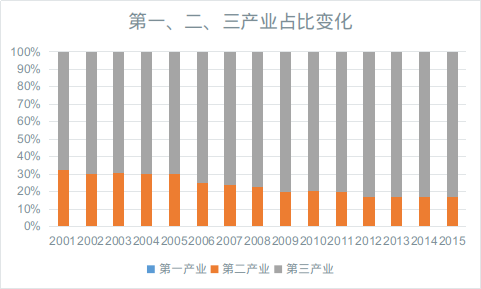
5、 分别统计各个产业每一年的就业人数

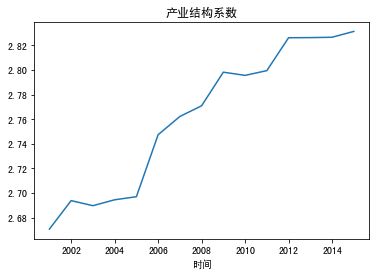
6、 用各个产业的就业人数除以总人数得到各个产业的就业率

7、 根据以上步骤获取到各个行业根据年份不同的企业数量，得到企业数量.csv(详情见插件：无锡市环境分析中间数据.zip)

8、 根据配第-克拉克定律，区域产业高级化是指：第一产业产值（或就业率）在国民经济中所占比例降低，第二产业产值（或就业率）先升后降，第三产业在国民收入（或就业）中所占比例升高。因此，我们可以根据相关文献，采取以下产业结构升级系数来测算，具体计算公式为：

其中，R表示产业结构升级系数，且其系数值的取值范围为1≤R ≤3，该洗系数若越接近哪个数字，则说明越接近第几个产业，比如，若该系数接近或者等于1，说明该系数对应的社会处于以农为主的农耕文化，而接近3说明该社会处于后工业化的信息经济社会，服务业占较大比重，若接近2，说明社会产业结构介于两者之间，正在处于工业化阶段。而其中， Y\_i为第i产业在第一二三产业总和中所占比例。





### 2.3 环境污染源的分析

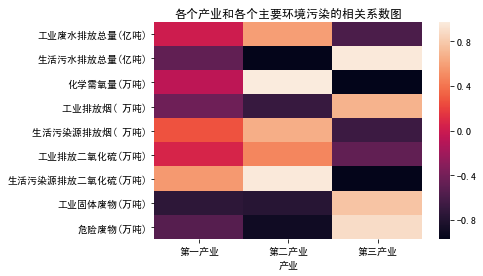
挖掘公司、企业的行业类别和环境指标之间的关系，已找到不同污染源的来源公司、行业以及经营内容。

1. 根据以上的两个步骤，获取到随着时间的变化，各个行业占比以及各个主要环境污染指标的取值(2001年到2015年)

2. 我们可以分析某个行业对某个环境污染指标的相关系数，这里我们使用了斯皮尔曼相关系数，具体计算公式如下：

其中，为在x、y两个样本集中，该样本的降序文职，也称秩

3. 两两计算完毕后，我们可以得到各个产业和各个主要环境污染的相关系数

* (产业-污染相关系数.csv)
* 

## 产品

### 3.1 使用工具

**Microsoft Excel + Microsoft CSV**

介绍：Microsoft Excel是Microsoft为使用Windows和Apple Macintosh操作系统的电脑编写的一款电子表格软件。直观的界面、出色的计算功能和图表工具，再加上成功的市场营销，使Excel成为最流行的个人计算机数据处理软件。CSV（Comma-Separated Values，CSV，有时也称为字符分隔值，因为分隔字符也可以不是逗号），其文件以纯文本形式存储表格数据（数字和文本）。

应用：本产品使用 Excel 和 CSV 中的表格筛选，表格排序，透视表的生成和应用，LEFT、RIGHT函数以及SUM等函数对文件数据进行清洗、整理、排序和统计。

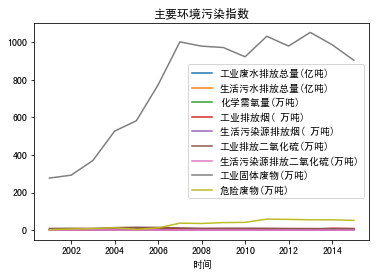
**DataV**

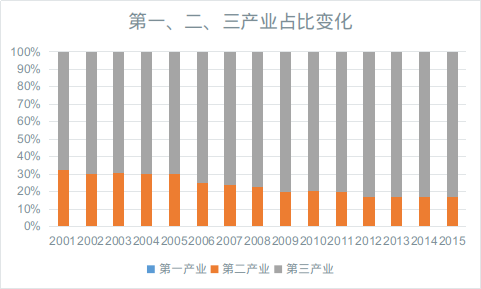
介绍：dataV是阿里云产品里用于对数据可视化的一个工具。它提供了丰富的模板以及各种可视化图表，可满足从会议记录到地理信息等多项分析展示的需求。对面临数据可视化设计无从下手，团队对于复杂数据的展现经验不足，提供了一个很好的平台。同时解决了分辨率等问题，界面美观而生动。

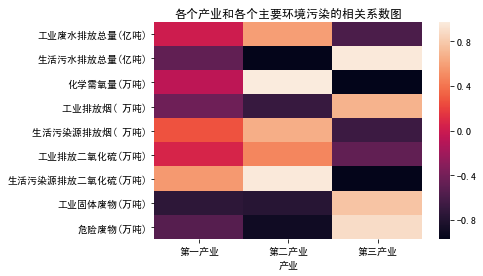
应用：本产品可供用户选择多种类型的图表，给予一定的数据格式，并提供多种数据源的接口，将数据导出成一定格式，来协助用户完成所需产品需求，进行实时数据可视化。

## 结果

通过已有的数据，结合企业经营情况，对未来一年可能发生的环境污染做出预警。



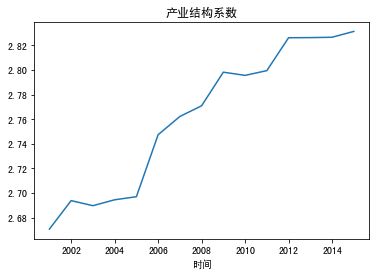




根据上文的图表，我们发现无锡市的主要产业一直是第二、第三产业，在2005年后，不断地进行快速的企业转型，第三产业的占比不断提高，然而，我们通过关联分析发现，第三产业的发展并不是百利而无一害的，生活污水排放量、工业排放烟工业固体废物、危险废物等污染指标与第三产业的发展有着强烈的正相关系数，这暗示了，当我们继续推进产业结构高级化演变时，将使得以上的污染标准（生活污水排放量、工业排放烟工业固体废物、危险废物）继续提升，所以，我们应该继续处理现有污染的同时，注重上文提到的污染标准继续提升的可能。

## 5、展望

在已有的基础上，通过关联可视分析，为无锡市的未来产业结构、经营范围调整、做出有益建议，以平衡环境污染和经济发展之间的平衡。



我们可以发现，无锡市的产业结构系数根据数据，处于2.7-2.8的水平，也就是第三产业提供的就业率是比较高的，然而，在前面的分析我们也提到了，在第三产业蓬勃发展的背后，是生活污水排放量、工业排放烟工业固体废物、危险废物等污染指标的不断提高，这说明，工业的实际污染处理水平没有根据产业结构系数的提高而提高，导致第三产业发展的同时，第二产业无法使用足够环保的原材料提供方式，导致随着产业结构系数提升，而污染指数提高的怪象，我们认为，无锡市应该敦促产业结构不断升级优化的同时，加大第二产业的污染处理水平水平优化，以顺应既要金山银山，又要绿水青山的可持续发展发展战略。

## 6、团队

团队名称：团队名字不能为空！

所在高校：广东工业大学

团队成员

高键城：来自广东工业大学计算机学院2016级软件工程专业的本科生，目前在计算机学院QG工作室担任后台组组长一职。该成员熟悉 Java 编程，熟悉Spring、MyBatis 等后台框架的使用，了解Java虚拟机以及并发编程。具有一定的开发经验和文档编写能力。

梁嘉豪：来自广东工业大学自动化学院2016级物联网专业的本科生，目前在计算机学院QG工作室担任数据挖掘组成员。该成员熟悉python、c++ 编程，熟悉python多种库的应用，熟悉地理信息的挖掘。具有一定的开发经验和文档编写能力。

伍欣怡：来自广东工业大学计算机学院2016级信 息安全专业的本科生，目前在计算机学院QG工作室担任前端组成员。该成员熟悉 html、css、Java Script ，熟悉vue、echart、webpack等前端框架的使用。具有一定的开发经验和文档编写能力。

团队分工

高键城：数据清洗和数据统计

梁嘉豪：数据分析和数据挖掘

伍欣怡：数据可视化展示