

灰色关联分析

Xu, X.

2025/3/23

Background

一般的抽象系统，如社会系统、经济系统、农业系统、生态系统、教育系统等都包含有许多种因素，多种因素共同作用的结果决定了该系统的发展态势。人们常常希望知道在众多的因素中，**哪些是主要因素，哪些是次要因素；哪些因素对系统发展影响大，哪些因素对系统发展影响小；哪些因素对系统发展起推动作用需强化发展，哪些因素对系统发展起阻碍作用需加以抑制**

为了实现少投入多产出，并取得良好的经济效益、社会效益和生态效益，就必须进行**系统分析**。

Background

数理统计中的回归分析、方差分析、主成分分析等都是用来进行系统分析的方法。这些方法都有下述不足之处：

- (1)要求有大量数据，数据量少就难以找出统计规律；
- (2)要求样本服从某个典型的概率分布，要求各因素数据与系统特征数据之间呈线性关系且各因素之间彼此无关，这种要求往往难以满足；
- (3)计算量大，一般要靠计算机帮助；
- (4)可能出现量化结果与定性分析结果不符的现象，导致系统的关系和规律遭到歪曲和颠倒。

Background

我国统计数据十分有限，而且现有数据灰度较大，再加上人为的原因，许多数据都出现几次大起大落，没有典型的分布规律。因此，采用数理统计方法往往难以奏效。

灰色关联分析方法弥补了采用数理统计方法作系统分析所导致的缺憾。它对样本量的多少和样本有无规律都同样适用，而且计算量小，十分方便，更不会出现量化结果与定性分析结果不符的情况。

灰色关联分析的基本思想是根据序列曲线几何形状的相似程度来判断其联系是否紧密。曲线越接近，相应序列之间的关联度就越大，反之就越小。

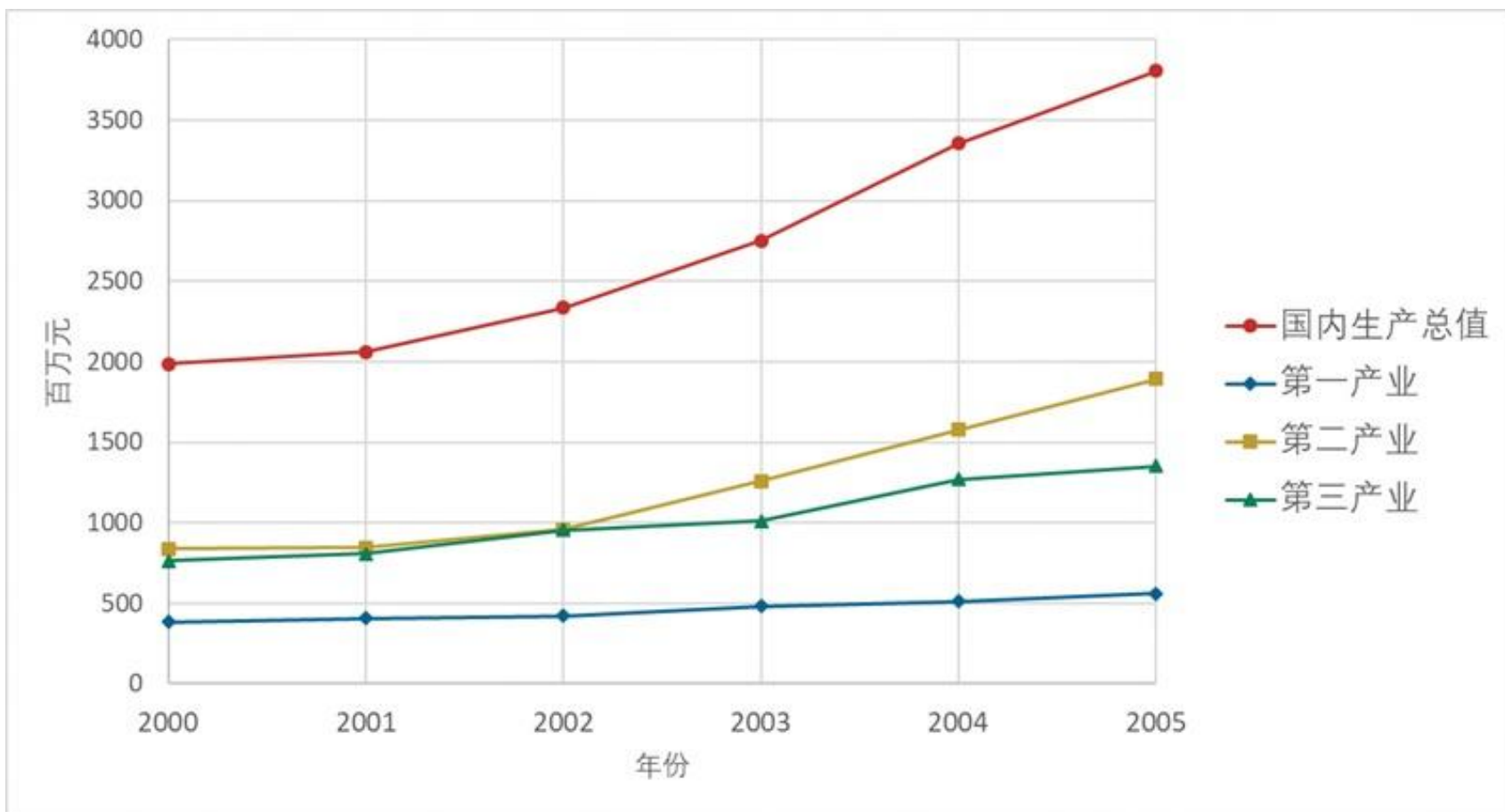
Example

下表为某地区国内生产总值的统计数据（以百万元计），问该地区从 2000 年到 2005 年之间哪一种产业对 GDP 总量影响最大。

年份	国内生产总值	第一产业	第二产业	第三产业
2000	1988	386	839	763
2001	2061	408	846	808
2002	2335	422	960	953
2003	2750	482	1258	1010
2004	3356	511	1577	1268
2005	3806	561	1893	1352

Procedure

Step1. 画统计图



画图后得配上简单的分析：

① 四个变量均呈上升的趋势

② 第二产业的增幅较为明显

③ 第二产业和第三产业的差距在后期相差更大

Procedure

Step2. 确定分析数列

(1) 母序列 (又称参考数列、母指标) : 能反映系统行为特征的数据序列。→ 类似于因变量 Y , 此处记为 X_0 。

(2) 子序列 (又称比较数列、子指标) : 影响系统行为的因素组成的数据序列。→ 类似于自变量 X , 此处记为 (X_1, X_2, \dots, X_m)

在本例中：国内生产总值就是母序列，第一、第二和第三产业就是子序列。
 X_0 X_1 X_2 X_3

Procedure

Step3. 对变量进行预处理(去量纲，缩小变量范围简便计算)

对母序列和子序列中的每个指标进行预处理：先求出每个指标的均值，再用该指标中的每个元素都除以其均值。

年份	国内生产总值	第一产业	第二产业	第三产业
2000	0.7320	0.8361	0.6828	0.7439
2001	0.7588	0.8838	0.6885	0.7878
2002	0.8597	0.9141	0.7812	0.9292
2003	1.0125	1.0440	1.0237	0.9847
2004	1.2356	1.1069	1.2833	1.2363
2005	1.4013	1.2152	1.5405	1.3182
均值	2716	461.67	1228.83	1025.67

Procedure

Step4. 计算子序列中各指标与母序列的关联系数

年份	国内生产总值 X_0	第一产业 X_1	第二产业 X_2	第三产业 X_3
2000	0.7320 $X_0(1)$	0.8361 $X_1(1)$	0.6828 $X_2(1)$	0.7439 $X_3(1)$
2001	0.7588 $X_0(2)$	0.8838 $X_1(2)$	0.6885 $X_2(2)$	0.7878 $X_3(2)$
2002	0.8597 $X_0(3)$	0.9141 $X_1(3)$	0.7812 $X_2(3)$	0.9292 $X_3(3)$
2003	1.0125 $X_0(4)$	1.0440 $X_1(4)$	1.0237 $X_2(4)$	0.9847 $X_3(4)$
2004	1.2356 $X_0(5)$	1.1069 $X_1(5)$	1.2833 $X_2(5)$	1.2363 $X_3(5)$
2005	1.4013 $X_0(6)$	1.2152 $X_1(6)$	1.5405 $X_2(6)$	1.3182 $X_3(6)$

记 $a = \min_i \min_k |x_0(k) - x_i(k)|$, $b = \max_i \max_k |x_0(k) - x_i(k)|$

↳ 两极最小差

↳ 两极最大差

	$ X_0 - X_1 $	$ X_0 - X_2 $	$ X_0 - X_3 $
(数据保留四位小数)	0.1041	0.0492	0.0119
	0.1249	0.0704	0.0289
	0.0544	0.0785	0.0694
	0.0315	0.0112	0.0278
	0.1288	0.0477	<u>0.0006</u> → a
	<u>0.1862</u> → b	0.1392	0.0832

Procedure

Step4. 计算子序列中各指标与母序列的关联系数

定义: $\gamma(x_0(k), x_i(k)) = \frac{a + \rho b}{|x_0(k) - x_i(k)| + \rho b}$ ρ : 分辨系数 (一般取 0.5) $i=1, 2, \dots, m$
 $k=1, 2, \dots, n$
 (gamma)

$$0.4751 = \frac{0.0006 + \frac{1}{2} \times 0.1862}{0.1041 + \frac{1}{2} \times 0.1862}$$

0.4751	0.6586	0.8922
0.4299	0.5733	0.7680
0.6356	0.5462	0.5766
0.7520	0.8985	0.7753
0.4224	0.6657	1.0000
0.3356	0.4035	0.5317

Procedure

Step5. 计算X0和Xi的灰色关联度

0.4751	0.6586	0.8922
0.4299	0.5733	0.7680
0.6356	0.5462	0.5766
0.7520	0.8985	0.7753
0.4224	0.6657	1.0000
0.3356	0.4035	0.5317

⑤ 定义 $\gamma(x_0, x_i) = \frac{1}{n} \sum_{k=1}^n \gamma(x_0(k), x_i(k))$ 为 x_0 和 x_i 的灰色关联度。

求平均值 $= \gamma(x_0, x_1) = 0.5084$

, $\rightarrow \gamma(x_0, x_2) = 0.6243$, $\gamma(x_0, x_3) = 0.7573$

Procedure

Step6. 得出结论

该地区在2000年至2005年间的国内生产总值受到第三产业的影响最大。(其灰色关联度最大)

Discussion

① 什么时候用标准化回归，什么时候用灰色关联分析？

当样本个数 n 较大时，一般使用标准化回归；当样本个数 n 较少时，才使用灰色关联分析。

② 如果母序列中有多个指标，应该怎么分析？

例如： Y_1 和 Y_2 是母序列， X_1, X_2, \dots, X_m 是子序列

那么我们首先计算 Y_1 和 X_1, X_2, \dots, X_m 的灰色关联度进行分析；再计算 Y_2 和 X_1, X_2, \dots, X_m 的灰色关联度进行分析。