Modeling and Simulation 建模与仿真, 2022, 11(5), 1352-1357
Published Online September 2022 in Hans. http://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2022.115127

Hans汉斯

函数型数据方法在水质分析中的应用

吴尚文

北京建筑大学理学院，北京

收稿日期：2022年3月21日；录用日期：2022年9月16日；发布日期：2022年9月27日

摘　要

研究水质变化趋势是水质监测的重要内容。水质变化过程是一个连续的过程，只是我们监测到的数据是离散的。由于水质监测数据具有不等时间观测、非线性变化的特点以及其数据内部表现出的函数性特征，考虑采用函数型数据分析方法进行研究。在本文中，我们在对样本数据进行函数化处理的基础上，本文将函数型回归模型应用于松花江肇源段的水质分析中，预测效果良好，为该地区的水质监测提供参考。

关键词

函数型数据，多元回归模型，水质分析

The Application of the Functional Data Method in Water Quality Analysis

Shangwen Wu

School of Science, Beijing University of Civil Engineering and Architecture, Beijing

Received: Mar. 21st, 2022; accepted: Sep. 16th, 2022; published: Sep. 27th, 2022

Abstract

Studying the variation trend of water quality is an important part of water quality monitoring. Though the data we got via monitoring is discrete, the variation process of water quality is continuous. Considering the monitoring data of water quality has the characteristics of unequal time observation, nonlinear change and functional feature, we selected functional data analysis. Based on the functional processing of sample data, we used functional multiple regression method to predict water quality of Zhaoyuan section of Songhua River. And the cluster was carried out according to principal components scores. The results show that the functional data analysis method is effective. This method provides a reference for water quality monitoring in this area.

文章引用: 吴尚文. 函数型数据方法在水质分析中的应用[J]. 建模与仿真, 2022, 11(5): 1352-1357.
DOI: 10.12677/mos.2022.115127

Keywords

Functional Data, Multiple Regression Model, Water Quality Analysis

1. 引言

大多数水质变化过程是一个连续的过程，这个过程生成的统计数据可以用一个类似函数的特征表达式来描述，而通过现有的统计手段所获取的信息往往是一个不连续的、片段的、离散的有界、有序的数列。函数型数据分析方法则能较好地处理这一类的数据：它将观测数据的产生当作一个函数过程，认为样本数据之中存在着某种函数型特征，采用连续函数的方法将原本的离散数据有效地联系起来，更好地探究了数据本身的非线性变化趋势[1]；针对函数数据的研究的目的与其他传统的统计学一样：在统计学思想和分析的指导下阐述问题；研究能够凸显数据重要特征的表现方式；为观测得到的数据建立统计模型等等[2] [3] [4] [5]。水质预测是一个经典问题，使用的方法很多，回归分析是其中最常用的方法之一。但这些方法都是基于离散数据的方法，忽略了数据背后隐藏的连续性，采用函数型数据分析可以挖掘数据的隐藏信息。实际上，函数数据分析中的各个方法比如回归分析、聚类分析等在水质数据分析、空气质量数据分析等领域得到较为广泛地应用[6] [7] [8] [9]。

本文将应用函数型数据分析方法进行黑龙江肇源水质数据的预测问题，实现更有效地监测水质的目的。

2. 函数型数据及函数型多元回归模型

2.1. 函数型数据

函数型数据(Function Data)是指一个集合，该集合中的元素均为定义在某个连续区间上的函数。

$$F\left(x_1,x_2,\cdots,x_p\right)=\begin{pmatrix} f_1\left(x_1,x_2,\cdots,x_p\right) \\ f_2\left(x_1,x_2,\cdots,x_p\right) \\ \vdots \\ f_n\left(x_1,x_2,\cdots,x_p\right) \end{pmatrix}, \tag{1}$$

$$x_j\in\left(-\infty,+\infty\right),\ j=1,2,\cdots,p$$

函数型数据是连续的数据，而一般的数据采集都是离散的。因为通过观察得到的原始数据通常以表格形式存储，这可以理解为对函数数据的截取，是自变量取一些特定的值所对应的函数值，他们是函数离散化的记录形式。显然，Ramsay 所提出的函数型数据的定义域是整个区间，所以函数型数据可以包含的信息，比常见的以数据表形式出现的离散数据所包含的信息更多。

从离散的观测数据中提取连续的函数数据，我们可以用基函数法、小波变换、核函数等[10]。本文中，我们介绍基函数法，其核心是用离散的观测值来估计其函数模型，可用下列公式表示

$$\hat{x}\left(t\right)=\sum_{k=1}^{K}c_k\varphi_k\left(t\right) \tag{2}$$

其中基函数 $\varphi_k\left(t\right)\left(k=1,2,3,\cdots,K\right)$ 的选择和系数向量 $c=\left(c_1,\cdots,c_k\right)'$ 的估计是两个难点。我们可以选择傅里

叶变换和样条插值作为基函数，前者针对周期性数据，后者针对非周期性数据，系数向量则通过最小二乘法来解决。

2.2. 函数型多元回归模型

多元线性回归分析是一个被广泛应用的重要方法。在自然科学和社会科学的诸多领域都有很好的应用[11] [12]。根据回归模型的不同，多元回归分析又可以分为线性回归和非线性回归两个大类。其中线性回归模型最为成熟，其应用也是最为广泛；此外，许多非线性回归模型也可以转化为线性回归模型来求解。

$$Y(t)=\beta_1 X_1(t)+\beta_2 X_2(t)+\cdots+\beta_p X_p(t)+\varepsilon(t) \tag{3}$$

在函数型数据分析领域内，与普通离散数据的多元线性回归分析技术相对应的就是函数型数据的常系数多元线性回归问题。与普通的多元线性回归分析的不同之处在于其因变量 Y 和自变量 $X_1,\cdots,X_p$，以及随机误差项 ε 都是以函数曲线形式存在的函数数据。而函数数据线性回归分析的目的就是通过研究因变量 Y 的曲线形态与自变量 $X_1,X_2,\cdots,X_p$ 的曲线形态之间的关系，建立因变量曲线与自变量曲线的线性回归模型[13]。

函数型数据的常系数多元回归模型的建立步骤如下：

首先，我们给出点积的定义。对于普通的离散数据而言，两个 p 维变量 x 和 y 的点积定义为：$\langle x,y\rangle=\sum_{i=1}^{p}x_i\cdot y_i$。显然，因为函数型数据的特殊性，这样的定义形式不适合定义数据型数据的点积，特此引用以下积分形式定义函数数据的点积，如下：

在函数数据空间中，对于 $x(t)\in L^2[a,b]$，$y(t)\in L^2[a,b]$，$x(t)$ 和 $y(t)$ 的点积定义为：

$$\langle x(t),y(t)\rangle=\int_a^b x(t)y(t)\mathrm{d}t \tag{4}$$

简记为：

$$\langle x(t),y(t)\rangle=\int x(t)y(t)\mathrm{d}t \tag{5}$$

另外，称 $I(t)=1,\forall t\in[a,b]$ 为单位函数。

假设因变量 $y(t)$ 和 p 个自变量 $x_1(t),x_2(t),\cdots,x_p(t)$ 均属于 $L^2[a,b]$。若观测了 n 个样本，则多元线性回归总体模型可以写成：

$$y_i(t)=\beta_0 I(t)+\beta_1 x_{i1}(t)+\beta_2 x_{i2}(t)+\cdots+\beta_p x_{ip}(t)+\varepsilon_i(t),\ i=1,2,\cdots,n \tag{6}$$

其中，$\varepsilon_i(t)$ 为随机误差项，并且 $\varepsilon(t)\sim N(0,\sigma^2),\forall t\in[a,b]$；$\beta_0,\beta_1,\cdots,\beta_p$ 是模型的待估参数，记其估计值为 $\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p$，则多元线性回归模型为：

$$\hat{y}_i(t)=\hat{\beta}_0 I(t)+\hat{\beta}_1 x_{i1}(t)+\hat{\beta}_2 x_{i2}(t)+\cdots+\hat{\beta}_p x_{ip}(t) \tag{7}$$

根据前面函数数据的点积定义，多元线性回归模型的残差平方和 SSE 可以由下列公式表示：

$$\mathrm{SSE}=\sum_{i=1}^{n}\left\|y_i(t)-\hat{\beta}_0 I(t)-\sum_{j=1}^{p}\hat{\beta}_j x_{ij}(t)\right\|^2=\sum_{i=1}^{n}\int\left[y_i(t)-\hat{\beta}_0 I(t)-\sum_{j=1}^{p}\hat{\beta}_j x_{ij}(t)\right]^2\mathrm{d}t \tag{8}$$

根据最小二乘原则，对上式求偏导，得：

$$\frac{\partial \mathrm{SSE}}{\partial\hat{\beta}_0}=-2\sum_{i=1}^{n}\int I(t)\left[y_i(t)-\hat{\beta}_0 I(t)-\sum_{j=1}^{p}\hat{\beta}_j x_{ij}(t)\right]\mathrm{d}t=0 \tag{9}$$

$$\frac{\partial \mathrm{SSE}}{\partial \hat{\beta}_k}=-2\sum_{i=1}^{n}\int x_{ik}(t)\left[y_i(t)-\hat{\beta}_0 I(t)-\sum_{j=1}^{p}\hat{\beta}_j x_{ij}(t)\right]\mathrm{d}t=0,\ \ k=1,2,\cdots,p \tag{10}$$

整理后得到正则方程如下：

$$\hat{\beta}_0\sum_{i=1}^{n}\int I^2(t)\mathrm{d}t+\sum_{j=1}^{p}\hat{\beta}_j\sum_{i=1}^{n}\int I(t)x_{ij}(t)\mathrm{d}t=\sum_{i=1}^{n}\int I(t)y_i(t)\mathrm{d}t \tag{11}$$

$$\hat{\beta}_0\sum_{i=1}^{n}\int x_{ik}(t)I(t)\mathrm{d}t+\sum_{j=1}^{p}\hat{\beta}_j\sum_{i=1}^{n}\int x_{ik}(t)x_{ij}(t)\mathrm{d}t=\sum_{i=1}^{n}\int x_{ik}(t)y_i(t)\mathrm{d}t,\ \ k=1,2,\cdots,p \tag{12}$$

用矩阵表示上述方程，有，

$$\begin{bmatrix}\sum_{i=1}^{n}\int I^2(t)\mathrm{d}t & \sum_{i=1}^{n}\int x_{i1}(t)I(t)\mathrm{d}t & \cdots & \sum_{i=1}^{n}\int x_{ip}(t)I(t)\mathrm{d}t\\ \sum_{i=1}^{n}\int x_{i1}(t)I(t)\mathrm{d}t & \sum_{i=1}^{n}\int x_{i1}^2(t)\mathrm{d}t & \cdots & \sum_{i=1}^{n}\int x_{ip}(t)x_{i1}(t)\mathrm{d}t\\ \vdots & \vdots & & \vdots\\ \sum_{i=1}^{n}\int x_{ip}(t)I(t)\mathrm{d}t & \sum_{i=1}^{n}\int x_{ip}(t)x_{i1}(t)\mathrm{d}t & \cdots & \sum_{i=1}^{n}\int x_{ip}^2(t)\mathrm{d}t\end{bmatrix}\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \vdots\\ \hat{\beta}_p\end{bmatrix}=\begin{bmatrix}\sum_{i=1}^{n}\int I(t)y_i(t)\mathrm{d}t\\ \sum_{i=1}^{n}\int x_{i1}(t)y_i(t)\mathrm{d}t\\ \vdots\\ \sum_{i=1}^{n}\int x_{ip}(t)y_i(t)\mathrm{d}t\end{bmatrix} \tag{13}$$

这是一个典型的线性方程组问题，可以很方便地采用经典的高斯消元法来求解，从而得到回归系数 $\hat{\beta}=\begin{bmatrix}\hat{\beta}_0 & \hat{\beta}_1 & \cdots & \hat{\beta}_p\end{bmatrix}'$。

回归模型即为：

$$\hat{y}(t)=\hat{\beta}'\cdot\begin{pmatrix}I(t)\\ x_1(t)\\ x_2(t)\\ \vdots\\ x_p(t)\end{pmatrix} \tag{14}$$

特别的，若函数数据 $x_i(t)$ 和 $y(t)$ 是由标准正交基 φ 展开得到的，那我们的计算将会得到很大程度的简化，因为：

$$\begin{aligned}\langle x(t),y(t)\rangle&=\int x(t)y(t)\mathrm{d}t\\ &=\int\left(a_1\varphi_q+a_2\varphi_2+\ldots+a_p\varphi_p\right)\cdot\left(b_1\varphi_1+b_2\varphi_2+\cdots+b_p\varphi_p\right)\mathrm{d}t\\ &=\int\left(a_1b_1\varphi_1^2+a_2b_2\varphi_2^2+\cdots+a_pb_p\varphi_p^2+a_1b_2\varphi_1\varphi_2+\cdots+a_pb_{p-1}\varphi_p\varphi_{p-1}\right)\mathrm{d}t\\ &=\sum_{i=1}^{p}a_ib_i\end{aligned} \tag{15}$$

3. 案例研究

3.1. 研究区域及数据

松花江肇源江段位于黑龙江省西南部的松嫩平原第二松花江与嫩江汇合口以下，是松花江哈尔滨江段的门户。松花江肇源江段的地理位置使水环境对其影响很大。松花江上游的吉林化工企业及嫩江流域齐齐哈尔等工业城市的废水排放量很大，大庆油田的废水经古恰闸门也排入松花江，同时，吉林油田，大庆油田沿江在泛洪区和江中岛上的采油作业也对肇源地区的水环境产生很大影响，肇源江段水环境问题是松花江重要的环境问题之一。并在松花江污染防治中占有重要作用。水质数据选取 4 个变量，y——PH，x_1——溶解氧(mg/L)，x_2——化学需氧量(mg/L)，x_3——氨氮(mg/L)，原始数据见表 1。

我们选取黑龙江肇源 2006~2011 年水质监测的数据，如下表所示：

Table 1. Water quality monitoring data of Zhaoyuan section of Songhua River from 2006 to 2011
表 1. 松花江肇源段 2006~2011 年水质监测数据

PH	DO	COD	NH_3-N	年份
7.27	7.17	8.6	1.04	2011
7.54	9.03	7.2	0.83	2011
7.52	7.95	7.9	0.7	2011
7.66	4.2	7.5	0.92	2011
7.53	7.41	7.6	0.61	2011
7.74	9.62	7	0.33	2011
……	……	……	……	……
7.13	5.83	8.4	0.12	2006
7.24	6.08	8.9	0.1	2006
7.17	6.1	9.1	0.11	2006
7.08	6.11	8	0.1	2006
N = 271				

3.2. 水质预测模型

3.2.1. 模型建立

对这 4 个变量进行 8 项傅里叶级数曲线拟合。然后，按照函数型数据多元回归建模方法实现下列线性方程组的计算：

$$\begin{bmatrix} \sum_{i=1}^{n}\int I^2(t)\mathrm{d}t & \sum_{i=1}^{n}\int x_{i1}(t)I(t)\mathrm{d}t & \cdots & \sum_{i=1}^{n}\int x_{ip}(t)I(t)\mathrm{d}t \\ \sum_{i=1}^{n}\int x_{i1}(t)I(t)\mathrm{d}t & \sum_{i=1}^{n}\int x_{i1}^2(t)\mathrm{d}t & \cdots & \sum_{i=1}^{n}\int x_{ip}(t)x_{i1}(t)\mathrm{d}t \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n}\int x_{ip}(t)I(t)\mathrm{d}t & \sum_{i=1}^{n}\int x_{ip}(t)x_{i1}(t)\mathrm{d}t & \cdots & \sum_{i=1}^{n}\int x_{ip}^2(t)\mathrm{d}t \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n}\int I(t)y_i(t)\mathrm{d}t \\ \sum_{i=1}^{n}\int x_{i1}(t)y_i(t)\mathrm{d}t \\ \vdots \\ \sum_{i=1}^{n}\int x_{ip}(t)y_i(t)\mathrm{d}t \end{bmatrix} \tag{16}$$

得到的结果是：

$$\begin{bmatrix} 271 & 2054 & 1645 & 170 \\ 2054 & 16014 & 12217 & 1360 \\ 1645 & 12217 & 10446 & 983 \\ 170 & 1360 & 983 & 137 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \end{bmatrix} = \begin{bmatrix} 2020 \\ 15324 \\ 12238 \\ 1276 \end{bmatrix}$$

计算得到：

$$\hat{\beta} = \begin{pmatrix} 7.8367 & -0.0454 & -0.0407 & 0.3327 \end{pmatrix}'$$

所以最终得到的回归模型是：

$$y = 7.8367 - 0.0454x_1 - 0.0407x_2 + 0.3227x_3$$

3.2.2. 模型拟合优度

拟合优度 R^2 是检验回归模型优劣的重要指标，是检验模型拟合实际数据的接近程度。该模型的拟合优度为(R^2 越接近 1，效果越好)：

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 0.9089$$

从理论上讲，模型考虑了数据的连续性，能挖掘数据的隐藏信息，实际计算的拟合优度 $R^2 > 0.9$，说明模型的效果优良，可以放入更大型的数据中进行水质数据的预测工作，为水质监测工作提供参考。

4. 结论

水质数据本质上是连续性数据，本文利用函数型数据中的回归方法进行黑龙江省松花江肇源段水质数据的预测工作。该方法具有理论优势，模型精确度也较高。该方法的引入有利于对水质监测工作进行科学合理地改进。

参考文献

[1] Ramsey, J.O. (1982) When the Data Are Functions. *Psychometrika*, **47**, 379-396. https://doi.org/10.1007/BF02293704

[2] 米子川, 赵丽琴. 函数型数据分析的研究进展和技术框架[J]. 统计与信息论坛, 2012, 27(6): 13-20.

[3] 靳雪晴. 函数型数据分析若干方法[J]. 现代计算机, 2021, 27(34): 77-80.

[4] 靳刘蕊. 函数性数据分析方法及其应用研究[D]: [博士学位论文]. 厦门: 厦门大学, 2008.

[5] 严明义. 函数性数据的统计分析: 思想、方法和应用[J]. 统计研究, 2007, 24(2): 87-94.

[6] Henderson, B. (2005) Exploring between Site Differences in Water Quality Trends: A Functional Data Analysis Approach. *Environmetrics*, **17**, 65-80. https://doi.org/10.1002/env.750

[7] 刘阳, 王欢, 唐萍, 余晓美. 环巢湖河流水环境质量的时空变化分析[J]. 安徽农业科学, 2021, 49(14): 72-75.

[8] 余晓美, 沈永昌. 中国环境保护重点城市空气质量的动态特征分析[J]. 统计与决策, 2019, 35(11): 91-94.

[9] 朱佳. 基于函数型数据分析和广义分位数回归的 PM2.5 数据探究[D]: [硕士学位论文]. 厦门: 厦门大学, 2018.

[10] 王劼. 函数型数据的分类方法研究及其应用[D]: [硕士学位论文]. 北京: 北京航空航天大学, 2009.

[11] 高惠璇. 应用多元统计分析[M]. 北京: 北京大学出版社, 2005.

[12] 郑明, 陈子毅, 汪嘉冈. 数理统计讲义[M]. 上海: 复旦大学出版社, 2006.

[13] 丁辉, 许文超, 朱汉兵, 王国长, 张涛, 张日权. 函数型数据回归分析综述[J]. 应用概率统计, 2018, 34(6): 630-654.