# 摘要

目录

[摘要 1](#_Toc29299)

[1.问题重述： 3](#_Toc30915)

[问题1分析 3](#_Toc22669)

[问题2分析 3](#_Toc31407)

[问题3分析 3](#_Toc17081)

[2.模型假设: 4](#_Toc18042)

[3.符号说明： 4](#_Toc12808)

[4.海区聚类分析 5](#_Toc11164)

[4.1 K-MEANS聚类分析概述 5](#_Toc26643)

[4.2 建立模型 5](#_Toc23875)

[4.3轮廓系数求解K 6](#_Toc19302)

[4.3 聚类结果 7](#_Toc28565)

[5.海温模型建立 8](#_Toc28716)

[5.1 同一类观测点温度曲线 8](#_Toc15406)

[5.2月平均温度 9](#_Toc28489)

[5.3温度变化趋势 12](#_Toc20295)

[5.3.1线性回归法 12](#_Toc18251)

[5.3.2 低通滤波法 13](#_Toc19314)

[参考文献 16](#_Toc7639)

[附录 17](#_Toc24522)

# 1.问题重述：

海洋表面温度变化能够从分体现环境的变化情况，海洋表面污染的增多和全球的环境的污染以及二氧化碳的排放致使气温的升高，都会引起海洋表面温度的升高，同时海洋温度同时又有四季的周期变化的特点，本题给出几个海区的温度10年的日观测数据，通过数据建模回答一下问题。

1通过每个海区的观测的观测数据，对观测点进行聚类分析

2对同一类观察点对于给定数据建立反映温度变化规律的数学模型，依据建立的模型对该区域的海洋表面温度的变化特点进行说明（比如季节，温度的变化趋势）。

3综合几个海区的情况对全球海洋表面温度变化特点进行描述，评估其变化趋势对人类的生活影响，同时为了不出现不良影响，人类的生活应该如何改变。

## 问题1分析

问题要求对四个海区，东海、杭州湾、南海、台湾海峡的观测点分别进行聚类分析。

每个观测点有9年3287天每一天的温度测量信息，考虑使用K-MEANS聚类方法对观测点进行聚类分析，K-MEANS聚类方法的时间复杂度较低，但对于分类数K的选取没有明确的方法，引入指标衡量选取的K值的聚类效果。同时，每两个观测点的距离设置为每天温差的平方和开根号。

## 问题2分析

问题2要求对同一类观察点对于给定数据建立反映温度变化规律的数学模型，依据建立的模型对该区域的海洋表面温度的变化特点进行说明。

考虑海温随时间的变化曲线，具有非常明显的周期性，可以使用信号处理中的傅里叶级数展开，将其分离成直流分量、基波、谐波叠加的形式，求出其主周期后，很容易利用信号傅里叶级数模型重建信号。

其次，从月平均温度变化来描述其季节性，从其直流分量的变化来描述总体变化趋势。

## 问题3分析

# 2.模型假设:

在对问题进行建模前，针对问题实际情况提出一些假设条件按

1. 假设数据真实有效，不存在错记录、漏记录
2. 假设对于问题2，3，4，任一选取的海区的分析方法可以直接套用到其他海区
3. 假设我的滤波器通接着电源

# 3.符号说明：

表2：本文出现的符号及文字说明

|  |  |
| --- | --- |
| 符号 | 意义 |
| t | 天数 |
| n | 观测点数 |
| p | 观测点 |
| d(x,y) | 点x与点y的欧式距离 |
| data(p,t) | 观测点p第t天的温度 |
|  | 第i簇的聚类中心 |
| Total(i) | 第i簇的点数 |
| k | 聚类分类数 |

# 4.海区聚类分析

## 4.1 K-MEANS聚类分析概述

K-means算法是很典型的基于距离的聚类算法，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大。该算法认为簇是由距离靠近的对象组成的，因此把得到紧凑且独立的簇作为最终目标。

k个初始类聚类中心点的选取对聚类结果具有较大的影响，因为在该算法第一步中是随机的选取任意k个对象作为初始聚类的中心，初始地代表一个簇。该算法在每次迭代中对数据集中剩余的每个对象，根据其与各个簇中心的距离将每个对象重新赋给最近的簇。当考察完所有数据对象后，一次迭代运算完成，新的聚类中心被计算出来。如果在一次迭代前后，分类没有发生变化，说明算法已经收敛。

K-Means聚类算法的优点主要集中在:

1.算法快速、简单;

2.对大数据集有较高的效率并且是可伸缩性的;

3.时间复杂度近于线性，而且适合挖掘大规模数据集。K-Means聚类算法的时间复杂度是O(nkt) ,其中n代表数据集中对象的数量，t代表着算法迭代的次数，k代表着簇的数目。

本题中的数据数量为576\*3287，数据较多，为再较短时间内实现聚类，故选用K-MEANS聚类算法。

## 4.2 建立模型

以杭州湾观测点数据为例，建立K-MEANS聚类算法模型。

1. 设定好的分类数k
2. 选取k个观测点，作为初始的聚类中心。
3. 对于每一个观测点p，计算其与每一个聚类中心的欧式距离



找到最小的d和，并将p归类到以为中心的簇

1. 重新计算每簇的聚类中心



1. 重复3，4两步，直到分类不再改变为止。

记录每簇的中心和类内点数

该算法中，初始k个聚类中心选取对分类效果的影响很大，理论上应选取几个相距较远的点作为初始聚类中心，避免出现类间距离太小的情况。

## 4.3轮廓系数求解K

对于K-MEANS算法来说，K值的选取是自定义的。但在实际实现过程中，K值的选取直接影响到聚类效果。为此给出以下定义：

1.类间距离

已知一个含有个数据的数据集，将其划分为类。簇类间距离即为每个簇类中心到全域中心的距离之和。类间距离越大，说明各个类之间的差别越大，聚类效果越好；如果类间距离越小，则说明不同类之间的差别很小，相似度较高，聚类效果不明显。

2.类内距离

类内距离又称类内误差平方和准则，是计算同一簇类内所有数据到该簇类中心的距离的和，如果类内距离越小，说明同一类中数据距离越相近，相似性程度越高，类别凝聚力越大，聚类效果越好；如果类内距离越大，则说明同一类中数据之间距离较远，相似性程度较低，类别凝聚力小，聚类效果不明显，有可能存在孤立点或噪声点。

聚类内部距离是指同一族类内数据对象的平均距离，它反映了聚类的紧凑性和算法的有效性；而类间距离是指簇类之间所有数据的平均距离，它反映了聚类的清晰性和算法的有效性。较优的聚类算法应该是具有类内部距离较小，类间距离较远。

查阅资料，轮廓系数可以量化类间距离和类内距离对分类效果的影响，

[轮廓系数(silhouette coefficient)](http://en.wikipedia.org/wiki/Silhouette_(clustering)) 结合了凝聚度和分离度，其计算步骤如下：

1）对于第 i 个对象，计算它到所属簇中所有其他对象的平均距离，记 （体现凝聚度）

2）对于第 i 个对象和不包含该对象的任意簇，计算该对象到给定簇中所有对象的平均距离，记 （体现分离度）

3）第 i 个对象的轮廓系数为 

从上面可以看出，轮廓系数取值为[-1, 1]，其值越大越好，且当值为负时，表明 ，样本被分配到错误的簇中，聚类结果不可接受。对于接近0的结果，则表明聚类结果有重叠的情况。

## 4.3 聚类结果

论文中以杭州湾观测点数据为例，得到K值分别为3到17情况下的轮廓系数

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K | 3 | 4 | 5 | 6 | 7 | 8 |
| 轮廓系数 | 0.2096 | 0.2187 | 0.1932 | 0.1932 | 0.1852 | 0.1776 |
| K | 9 | 10 | 11 | 12 | 13 | 14 |
| 轮廓系数 | 0.1798 | 0.1713 | 0.1845 | 0.1793 | 0.1800 | 0.1860 |
| K | 15 | 16 | 17 |  |  |  |
| 轮廓系数 | 0.1783 | 0.1917 | 0.1859 |  |  |  |

分析表中数据可以得到：

1. 对于该海区来说，当K=4时，分类效果最好
2. 轮廓系数在4到17之间差别并不大

按照上述理论，最终确定K值为4，并给出最终分类情况表

|  |  |
| --- | --- |
| 簇 | 簇内点数 |
| i=1 | 64 |
| i=2 | 142 |
| i=3 | 139 |
| i=4 | 231 |

# 5.海温模型建立

## 5.1 同一类观测点温度曲线

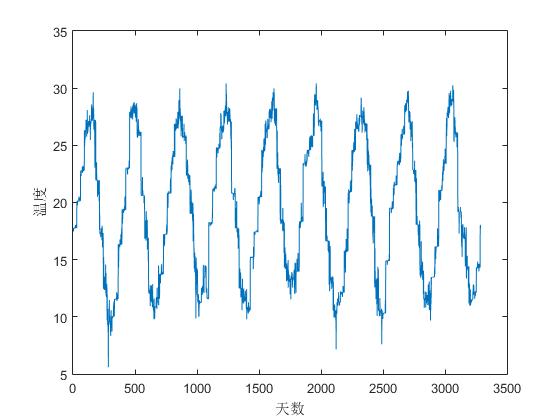
以杭州湾温度数据为例，根据问题1中的分类，取第3簇作分析。

为方便起见，一年记作365天，去除数据中两个闰年产生的2月29号数据，标号分别为425和1886。

对于同一类观测点，其温度与温度变化趋势应相近，取同一类中各点的每天温度平均值作为该类该天温度值



作出第3簇观测点的平均温度图



该曲线具有明显的周期性，考虑使用傅里叶级数模型描述该曲线



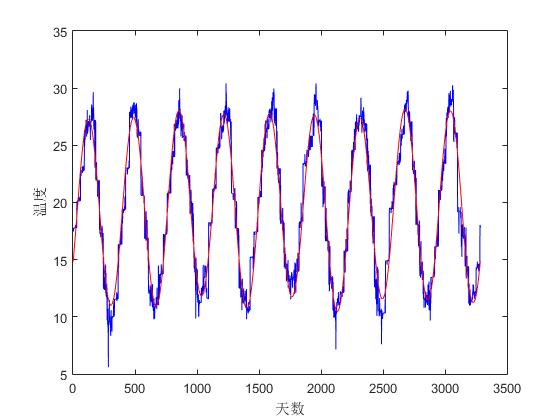
上式中d表示主周期个数，表示第i个谐波对应的周期，表示参数，表示噪声。

根据FFT变换，容易得到其主周期为365，760，1095，1460。拟合目标曲线为

共有9个参数，给出拟合结果

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |
| 8.24 | -0.56 | 0.40 | -1.07 | -0.42 | 0.52 | -0.37 | -0.50 | 19.46 |

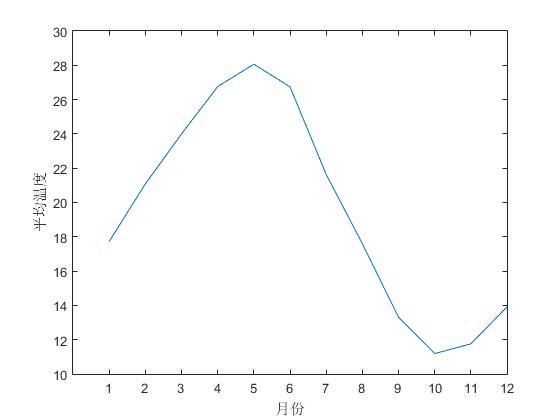
给出拟合结果



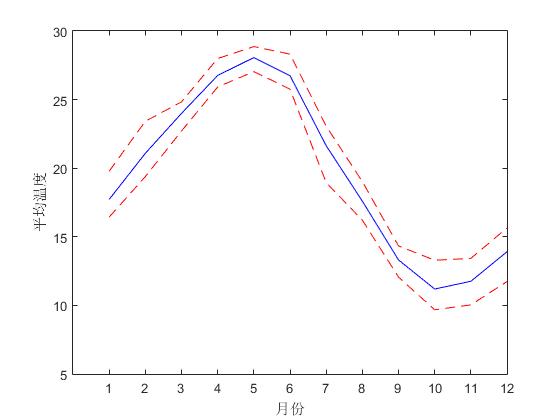
图中红色部分为拟合结果。查阅资料知，剩下的噪声序列为白噪声序列。

## 5.2月平均温度

首先分析该类观测点9年的平均每月温度，将9年数据按月份逐年相加并取平均得到月平均温度图



记录9年每月平均温度的最大和最小值，作图



具体数据见表中

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 月份 | 1 | 2 | 3 | 4 | 5 | 6 |
| 月平均温度/℃ | 17.7 | 21.0 | 23.9 | 26.7 | 28.0 | 26.7 |
| 最高月平均温度之差/℃ | +2.0 | +2.3 | +0.8 | +1.2 | +0.8 | +1.5 |
| 最低月平均温度之差/℃ | -1.2 | -1.6 | -1.2 | -.0.8 | -1.0 | -0.9 |
| 月份 | 7 | 8 | 9 | 10 | 11 | 12 |
| 月平均温度/℃ | 21.6 | 17.5 | 13.3 | 11.1 | 11.7 | 13.9 |
| 最高月平均温度之差/℃ | +1.4 | +1.4 | +1.0 | +2.1 | +1.6 | +1.7 |
| 最低月平均温度之差/℃ | -2.6 | -1.3 | -1.2 | -1.5 | -1.7 | -2.1 |

分析图和表可得如下结论：

1.月平均温度与最高月平均温度、最低月平均温度相差不大

2.月平均温度差最大为2月，最高2月平均温度与最低2月平均温度差为3.9℃

3.以春（3、4、5月）、夏（6、7、8月）、秋（9、10、11月）、冬（12、1、2月）来划分，则按季节变化的趋势为：春夏之交时，海水温度到达最高点，随后开始下降，至秋末到达最低点，再缓缓回升。

## 5.3温度变化趋势

### 5.3.1线性回归法

首先考虑用线性回归的方法，求出9年杭州湾温度变化的线性趋势。

线性回归是利用数理统计中回归分析，来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法，运用十分广泛。回归分析中，只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示，这种回归分析称为一元线性回归分析。

给出一元线性回归公式：



式中为回归系数，相关系数r的求解方法为



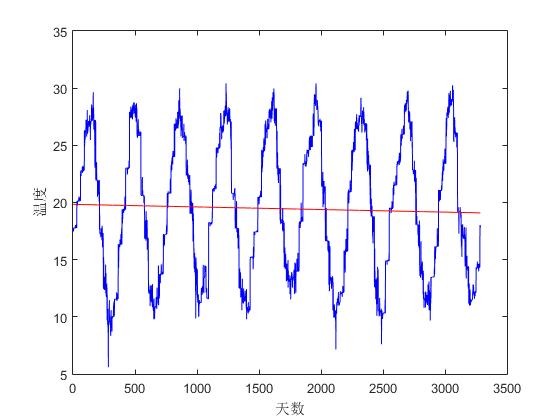
相关系数衡量回归方程与真实序列的吻合程度。

由上述公式得到







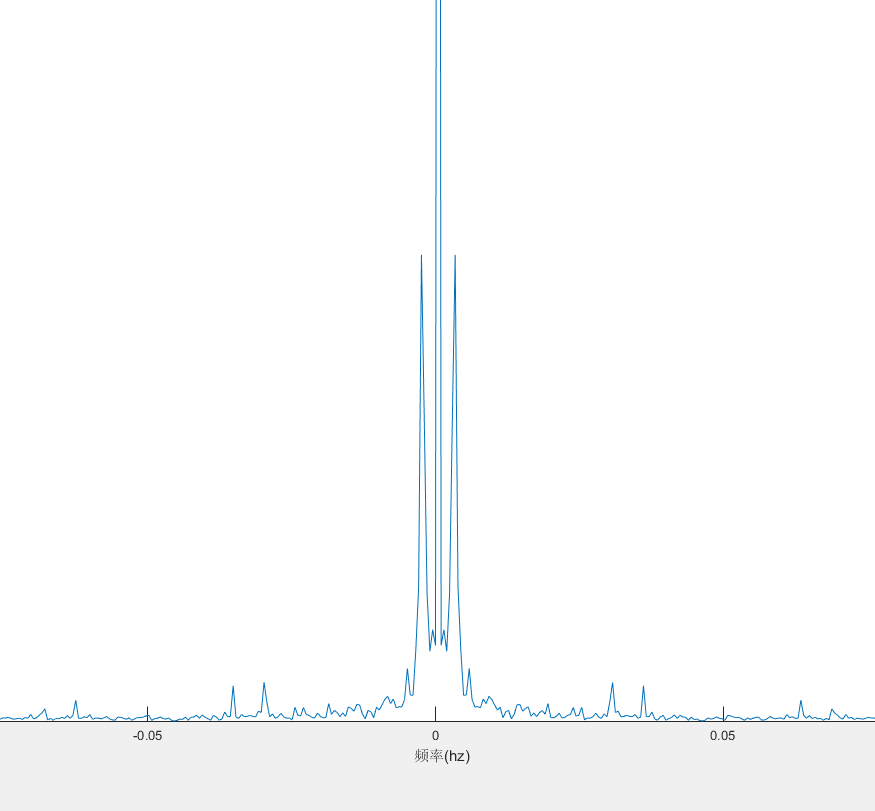


图中红色直线为回归方程，得到温度变化的线性趋势为-0.8272℃/10 a。注意到两者的相关系数非常的小，该结果的准确性尚待商榷。为此使用低通滤波的方法作为对比。

### 5.3.2 低通滤波法

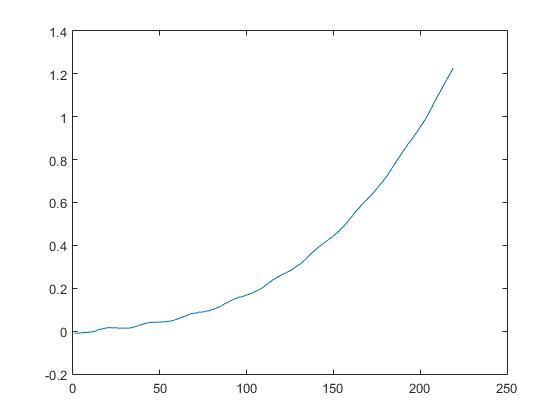
低通滤波是一种过滤方式，规则为低频信号能正常通过，而超过设定临界值的高频信号则被阻隔、减弱。但是阻隔、减弱的幅度则会依据不同的频率以及不同的滤波程序（目的）而改变。

低通滤波可以简单的认为：设定一个频率点，当信号频率高于这个频率时不能通过，在数字信号中，这个频率点也就是[截止频率](https://baike.baidu.com/item/%E6%88%AA%E6%AD%A2%E9%A2%91%E7%8E%87" \t "https://baike.baidu.com/item/%E4%BD%8E%E9%80%9A%E6%BB%A4%E6%B3%A2/_blank)，当频域高于这个截止频率时，则全部赋值为0。因为在这一处理过程中，让低频信号全部通过，所以称为低通滤波。

首先我们规定一天为最小的单位，那么原序列的采样频率fs=1hz，对原序列做快速傅立叶变换(FFT)

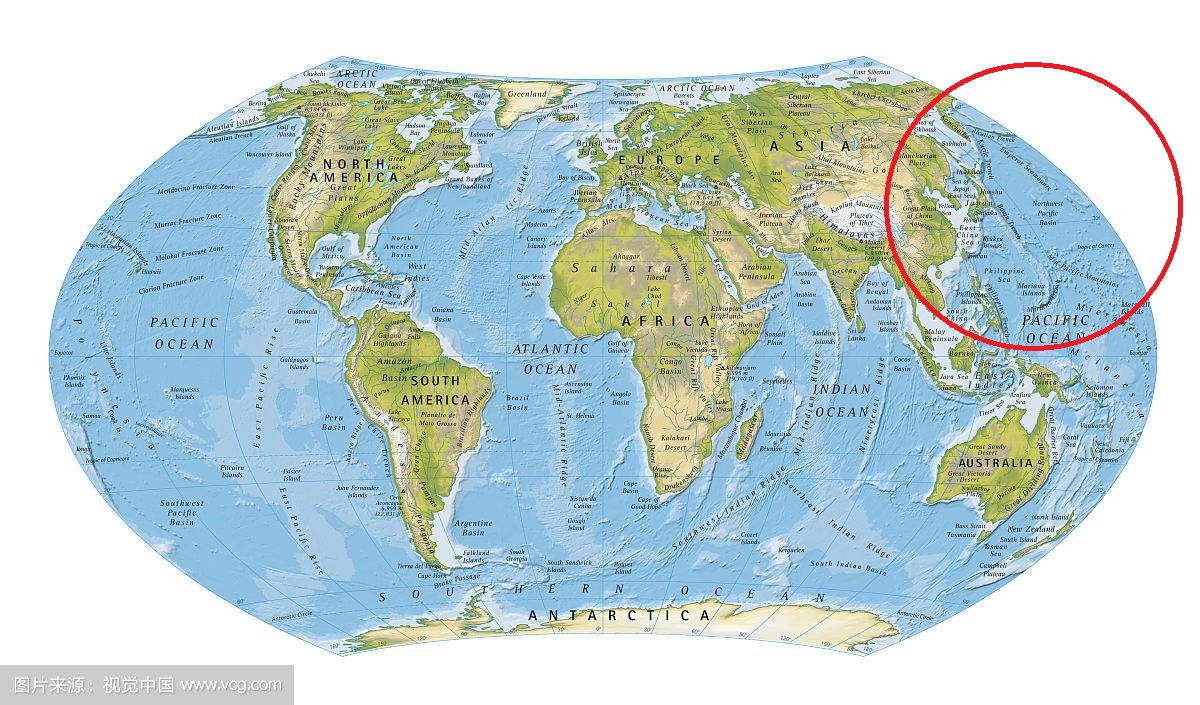
如图所示，通过计算得到其主要周期分量的周期为365，730，1095，1460，对应频率为1/365hz，1/730hz，1/1095hz，1/1460hz，通常低通数字滤波器的通带频率fp不宜过小，针对现有情况无法设计适合的低通滤波器。

考虑降低采样频率以满足设计滤波器的硬性要求，将采样频率设置为fs=1/15hz，经过FIR低通滤波器后，得到残留的近似直流分量为



根据上图，经过低通滤波后的序列呈上升趋势，表明该海区温度总体为上升趋势。

# 6.海温预测



通过东海、南海、台湾海峡、杭州湾海峡预测全球海洋变化趋势？

# 参考文献

# 附录