附录 1：

本科生毕业设计[论文]

（题目）：海洋表面二氧化碳分压数据的不确定度定量化算法

院　　系　物理学院

专业班级　应用物理学 1101

姓　　名　李翔

学　　号　U201117605

指导教师　戴民汉，陈长军

2015 年 6 月 8 日

学位论文原创性声明

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包括任何其他个人或集体已经发表或撰写的成果作品。本人完全意识到本声明的法律后果由本人承担。

作者签名：　　年　月　日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保障、使用学位论文的规定，同意学校保留并向有关学位论文管理部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权省级优秀学士论文评选机构将本学位论文的全部或部分内容编入有关数据进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于 1、保密口，在　　年解密后适用本授权书

2、不保密口　。

（请在以上相应方框内打“√”）

作者签名：　　年　月　日

导师签名：　　年　月　日

摘　要

鉴于海洋表面二氧化碳分压的数据采样后经过网格化处理来表征该区域的平均水平这一过程带来的巨大的不确定度，对于海气二氧化碳通量的计算的可靠性的影响是至关重要的。然而一般的不确定度的计算方法，即通过一组数据标准差的计算会混合多种不确定性的来源，甚至错误的估计了一个区域的不确定度的大小。针对目前存在的问题，本文设计了一个基于 Kriging 估值法的简单易于重复的定量算法，用于计算空间数据的不确定度，为不同的案例研究提供一个可靠的不确定度报告。该算法主要依赖于空间统计学对于空间数据相关性的量化从而更好的修正有效观测的测量次数，并通过 MATLAB 的运算平台设计一个运算软件，方便不同的案例研究中的数据更好更快捷的获取不确定度报告。

关键词：空间数据；不确定度分析；海表二氧化碳分压；Kriging 估计；

Abstract

Given that the uncertainty in the grideded sea surface pCO2 data is critical to assess the reliability of the CO2 flux estimated from measurements of air-sea pCO2 difference. Common uncertainty quantification by Standard Deviation will mix up the different sources of uncertainty. In this paper, combining the Kringing Estimation Method and spatial analysis, I provide a standard protocol to optimize three sources of uncertainty, especially spatial variability and its undersampling uncertainty. Using the remote sensing-derived and field-measured pCO2 data and spatial analysis, this method can offer an quantitatively uncertainty report. Considering its potential wide application, I wrote an automatic program in MATLAB platform to complete the uncertainty report.

Key Words： Spatial Data; Uncertainty Analysis; Sea Surface pCO_2 ; Kriging Estimation;

目 录

1. 绪论

1.1 背景介绍

闭合的碳收支和减少海气二氧化碳通量估计中的不确定度，因此更好的限制在海洋碳汇格局中空间时间的变异并最终更好的预测气候系统的变化，已成为众多国际合作的目标。其中，最有希望的一个方法用于估计海气二氧化碳的通量是基于测量海气的二氧化碳分压差在海表水中二氧化碳的分压差是最主要的决定因素，因为在大气中的二氧化碳分压相对的均匀和很好的定义。

在**[14]** 之后，自从 1972 年第一个将海表二氧化碳分压测量汇编，人们付出了巨大的努力去获得海表二氧化碳的数据，最主要的通过走航获取的全球数据。当估计海表二氧化碳的通量的时候，观察的数据被代表性的分入到网格化的格子中。因此本质上，三种误差来源有着不同的应用去计算网格化数据中的不确定度来源。第一个是，分析误差在二氧化碳分压决定和相联的环境参量用于获取二氧化碳分压。 第二，海表空间上的二氧化碳分压空间分布不均，因此在调查的海域会有不均匀的变异，特别是在近岸和靠近海洋边界和峰前。任何插值方法在一个特定的空间区域都会因此注定产生误差。最终，误差也可以来自于对一些没有实际测量数据或者再取样区域的外插值过程。因为样本站也许不均匀分布在整个调查区域，因此缺乏样本是很普通的现象。在**[8]**的气候学研究中，这些误差部分被仔细的检查，但他们用了一个标准差进行了平均二氧化碳分压差的估计。平均来说，整个全球海洋，基于所有的网格点的测量，平均月二氧化碳分压误差估计在 0.8uatm。除此之外，他们还评估了系统变差在水表二氧化碳分压由于再取样， 他们的插值方法基于二维传导对流输运方程。一个偏差的组成部分通过海表温度和气候的海表温度对比而估算。但是，对于区分三种不同的误差来源这个任务依然艰巨。甚至更困难去定量计算这些误差来源，在一个限定的空间和时间尺度内，其观察数据相关联。在一些其他的研究中，特别是在近岸的海洋区域，因为不足的数据覆盖范围，二氧化碳通量的数据通常被报道又一个方差代表不确定度和一个平均值来表示一个网格内的所有误差。这个简化的不确定度定量计算方式会混合不同的

不确定度来源，甚至会对数据的内插外插值带来超出样本点的困难，或者去优化设计走航观察方案来减少不确定度。这个研究从之前简化的二氧化碳分压 数据不确定度表示方式向前推进了一步，而且探究了每一种不确定度的贡献的估算的可行性因此可以帮助评估 CO2 通量的可信度。我们的目标就是探究三种不同来源的不确定度，在网格化的二氧化碳分压数据中，1，分析误差，产生于分析方法和数据约化过程。2.空间变异，代表了二氧化碳分压空间变异特性在一个给定的区域.3.来自于再取样的偏差，有空间变异和有效观察数共同决定。使用在中国东海网格化的二氧化碳分压测量数据，我们试着去阐释一种定量的方法去区分这三种不同的不确定度来源**[13]**。

1.2. 理论方法：空间数据分析

1.2.1 什么是空间数据分析？

空间性分析包含三个要素，1.制图建模。每个数据集被表示为一幅地图或是基于地图操作生成的新的地图。2.空间性分析的数学建模，通过模型输出对象之间的空间性相互作用形式，或空间性关系，或模型中对象的地理位置。3.对空间性数据进行适当的分析的统计技术的开发和应用，以及要充分利用数据中的空间性参考。

1.2.2 为什么要空间数据分析？

在数字全球化，空间数据迅猛增多的今天，有关空间性数据分析的理论，方法及其应用引起了人们的很大关注。分析空间性数据可以帮助人们更好处理科学和决策领域的诸多空间性问题。空间性数据有很诸多不同分类方式，如表面数据，对象数据。根据不同的数据种类，相应的分析方法也有不同的形式**[3]**。

1.2..3 空间数据分析的意义

在观测科学的许多领域，在数据库中记录单个事件发生的时间和地点很重要。记录事件发生的地点和时间意味着观测事件与其他数据库中的相应数据进行关联成为可能，从而在寻求科学解释方面起着重要作用。在地理空间性中的观测是非

独立的，地理位置的领巾的观测值是趋于相似的，比那些相隔较远的观测要更为相似。这是地理空间的基本特性，可以在解决如空间性插值问题的情形中被开发利用。另一方面，空间性相关这个性质不可避免的对经典统计理论的应用产生影响，因为相关性导致相邻的数据产生冗余，这会影响样本的信息容量。

2. 空间统计理论

2.1 空间相关性的量化

因为潜在的空间连续性和各种各样的复杂过程的局域相互作用，所以使得空间相关性成为地理空间数据的内禀属性。这个内在属性将会被在地理空间上获取的数据所继承。继承的形式将依赖于，比如空间分辨率或样本密度，或对象空间中的聚合尺度**[16]** 。

量化空间相关性的第一步就是定义任何点集或面对象之间存在的空间关**[3]** 。许多形式的空间性数据分析要求这个初始化步骤。完成这个步骤后，有几种方法可以度量空间的相关性。这些度量方法可以用用于整个地图，以取得空间相关性的单个平均测度，如果怀疑异质性，还可以定义地理子集。

2.1.1 二维连续空间中的数据

假设采用点采样。标准的空间性关系分析是根据距离或距离条带而定义。在数据矩阵中，$S(i)$ 为计算任何一对点 i 和 j 之间的内点距离 $d(i,j)$ 提供了充分的信息。设标记 $[(i,j)|d(i,j)=h)]$ 表示假如选择 j 作为条件，从 j 到 i 的距离是 h. 这个条件可能会被 $[(i,j)|d(i,j)=h\pm\Delta)]$ 放宽。以便选择的 j 落入 i 的距离带内。距离带的作用是重要的，它允许数据点位置存在不精确性，并保证有足够的点对数量进行可靠的统计计算。但依然可能在一些带中有很多点对，其他带中则很少的情

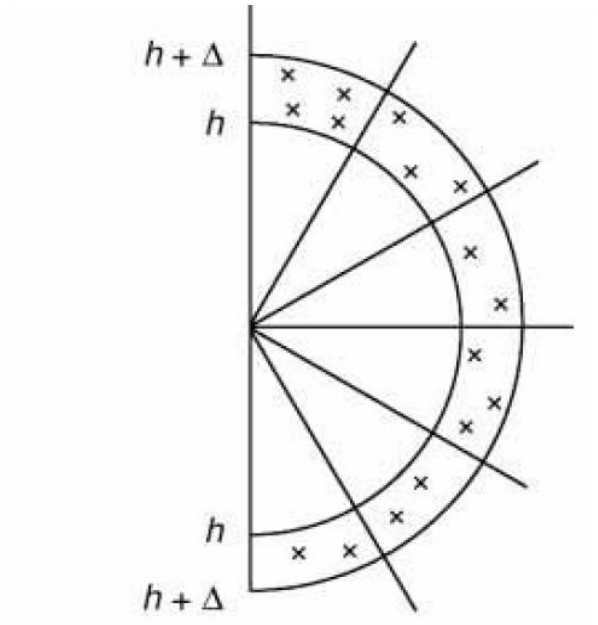


图 2-1 位于距离带内的数据点

况发生，因此造成估计精度的变化。应该考虑在不同带之间进行比较。

在数据点规则分布时，会存在一些带中可能没有点对，其他带中又有很多点对的情况。在数据充足的地方，配对也可以根据方向进行，使选择的 j 落入距离带内。如图[2-1].

接下来的讨论是基于：

$$[(i,j)|d(i,j)=h\pm\Delta)]$$

的简单事例，但是可以很容易推广到分带或分段的情况。对于任何距离 h，可以用二元散点图来图形化值得相似性。值得相似性可以由散射表明，这个散射是向右上倾斜的，被压缩在大约 45° 线附近。如果散射从这个对角线向外广泛传播，这就表示点对没有随着 h 的增加趋向产生相似性。

在测量的变量是序数，区间或比率的情况下，评估相似性的数字方法可以使用平方差 $(z(i)-z(j))^2$，如果 $z(i)$ 和 $z(j)$ 相似，则平方差小，否则平方差大。这个测度也可以通过叉积 $(z(i)-\bar{z})(z(j)-\bar{z})$ 来构造，其中 $\bar{z}$ 表示 $\{z(i)\}$ 的均值。如果 $z(i)$ 和 $z(j)$ 相似，这个量为正，否则为负。下面我们给出基于这两个量的空间相关性的数字描述方法。

对于任何给定的距离 h,这个量为：

$$\hat{\gamma}(h)=(1/2N(h))\underbrace{\sum_i\sum_j(z(i)-z(j))^2}_{[(i,j)|d(i,j)=h\pm\Delta)]} \qquad (2\text{-}1)$$

式中，$N(h)$ 表示按距离 h 相隔离的点对数，是在距离 h 处的半方差值，被距离 h 分隔的值越相似，它越小，如果值是不相似的，则它越大。因此，$\hat{\gamma}(h)$ 是随着 h 的增加而增加的。半方差函数是 $\{\hat{\gamma}(h),h\}$ 的图点，它提供了在不同距离上数据的相关性结构的图形描述。这个半方差计算了一半平均方差，也是空间自相关格雷检验的基础（Geary,1954）。图[2-2]是一个典型半方差的例子，在短距离上空间相关性强，然后随着 h 的增加，逐步减弱，直到超过一定的距离（变程），空间的相关性的变化（基台）接近 0。

另一个变量，对于任何一定的距离 h，有如下定义：

$$\hat{C}(h)=(1/N(h))\underbrace{\sum_i\sum_j(z(i)-\bar{z}(i))(z(j)-\bar{z}(j))^2}_{[(i,j)|d(i,j)=h\pm\Delta)]} \tag{2-2}$$

式中，$\bar{z}(i)$是指在第一个括号中所有的$\{z(i)\}$值得平均值，而$\bar{z}(j)$是第二个括号中所有$\{z(j)\}$的平均值。当然，将会有很多值对生成这两个平均值有贡献，但这两个平均值通常不会相等。为便于计算，式（2）可以写成：

$$\hat{C}(h)=[(1/N(h))\underbrace{\sum_i\sum_j(z(i)z(j))]}_{[(i,j)|d(i,j)=h\pm\Delta)]}-\bar{z}(i)\bar{z}(j) \tag{2-3}$$

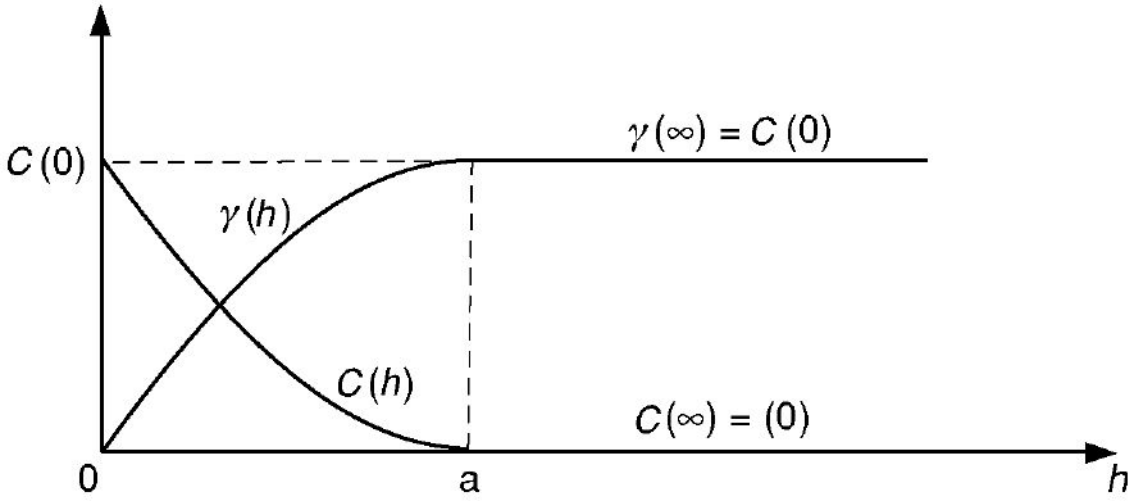


图 2-2

半方差函数$\hat{\gamma}(h)$和自协方差函数$\hat{C}(h)$的变化关系

$\hat{C}(h)$是在距离为 h 区域的自协方差（或空间协方差）的估计。它是叉积的平均值，并且当数值相似时，它的值大（叉积的值可能趋于正或负）；当数值不相似时，它接近 0（因为正负值将互相抵消）。

当 h = 0 时，根据式（2），$\hat{C}(0)$是$\{z(i)\}$的方差。如果$\hat{\sigma}(i)$和$\hat{\sigma}(j)$是式（2）中$\{z(i)\}$和$\{z(j)\}$对应的两个数据子集的标准差，那么：

$$\hat{R}(h)=\hat{C}(h)/\hat{\sigma}(i)\hat{\sigma}(j) \tag{2-4}$$

是在距离为 h 区域自相关（或者空间相关）的估计。自相关函数或者相关图的散点$\{\hat{R}(h),h\}$,有与自协方差函数一样的行为，但它是归一化后的函数。除了

$\hat{R}(h)=\hat{R}(-h)$ 边界影响，它可以被显示，并且与 $\hat{C}(h)=\hat{C}(-h)$ 类似。

2.2.2 离散空间中的数据

散点图使用的测量方法是基于均值平方差和均值叉积的，结合技术可以再次用于量化在离散空间中对空间对象进行测量的结合对之间的相关性。但是，在这种情况下，由 $\{S(i)\}$ 提供的数据信息则由其邻居来补充。邻居信息不仅由彼此相邻的对象对来定义，而且还需要量化这个相邻关系的紧密度。由 $\{S(i)\}$ 提供的信息可以用于定义邻居信息的过程，但也许需要使用其他数据和所做的假设（通常是不可测试的）。在离散空间这是需要的，在离散空间定义的许多过程类型不是单一的，或者是空间关系的自然定义。

用于定义邻居的指标包括：

直线距离：每一个点被连接到特定距离内的所有其他点。

最近邻居：每一个点被连接到与它的 k(=1，2,3，...)个最近邻居。（注意：如果点 A 是点 B 的 k 个邻居中最近的一个，这不能推导出点 B 也是点 A 最近邻居 k 个中的一个）。

Gabriel graphs：任何两个点 A 和 B 都可以进行连接，当且仅当所有其他点位于 A 和 B 的圆周，并处于相对立的点的圆周之外时。

Delaunay triangulation：在 Dirichlet 区域划分中，所有具有共享边界的点是相互关联的。Dirichlet 划分是基于一组点来创建的，保证了任何点 A 的周边地区包含所有比地图上任何其他点更靠近 A 的位置。图 2.3 表示了一组来已经构造的 Dirichlet 区域划分的点和在这个划分中他们是否共享一个公共边界为基础的点的结合情况。

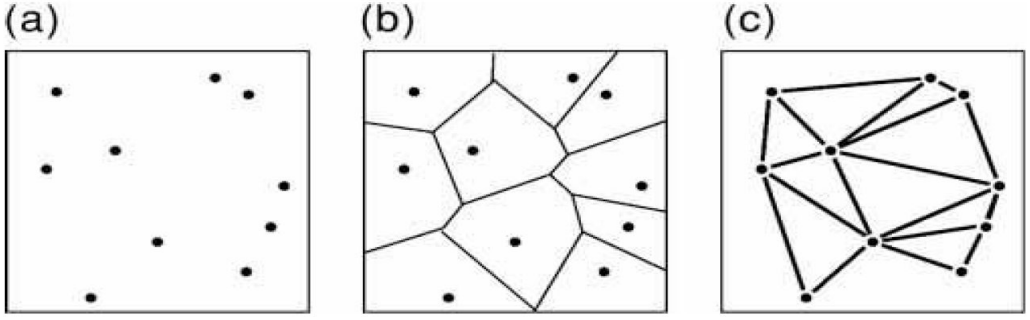


图 2-3

利用 Dirichlet 区域划分的邻居

像行政管理单元那样，在一组预定义的区域情况下，如果分隔研究区域，它们将共享公共边界，因此可以直接根据是否共享公共边界定义连接关系。如果该地区不是分割的研究区域，可由一个点（如地区或人口加权中心）定义它们的位置，然后运用上文所述的处理点对象的三种方法之一。

在离散空间中，在介绍用于描述空间相关性的式（2）之前，我们简要说明空间关系的矩阵方法描述。

空间性关系可以用二进制的连接形式或连接矩阵（C）去表示。如果有 n 个对象（点或面），定义一个与 n*n 个对象一样多的行和列的矩阵。每个面被唯一分配一个行和列。如果两个对象 i 和 j 是彼此连接的，那么；

$$c(i,j)=c(j,i)=1$$

其中，$c(i,j)$ 表示在矩阵 C 中的行 i 和列 j 上的实体项。否则，任何单元的值都为 0.任何一点或面 j 将被称为 i 的邻居，并用 N（i）表示。需要注意：一个对象不能与自身而连接，因此对于全部 i，$c(i,i)=0$，矩阵不会延着对角线对称。

如果 C 自乘，$C^2=C\times C$，则根据非零的单元实体可以识别所有的彼此可以以两步到达的面对象对。在结果矩阵中单元的值 $\{c^2(i,j)\}_{i,j}$ 确定了路径数。这个计数包括回溯路径（backtracking routes）。如果 i 与其他四个面对象相邻，那么 $c^2(i,i)=4$，因为在两步内从 i 回到自身的旅行路线将由四种方式。矩阵 C^3 替三阶的邻接区域提供了同样的信息，以此类推。这些关系，连同路径搜索的方法一起，包括多余回溯的路径，对编写有关区域规则的数据分析软件是重要的。但在分析像素数据时，这种矩阵是不必要的，因为矩阵是稀疏的，会浪费大量的存储资源和计算时间。

对于按照简单的出现为 1，不出现为 0 来描述对象之间的空间关系，没有特别规定。空间关系如果使用相互作用标准定义这些关系，可以使用更一般的权重矩阵 W 来定义。下面是不同类型权重矩阵的例子（如果他们被用于区域，距离可能按照参考区域中心来定义）：

（1）距离：$w(i,j)=d_{i,j}^{-\delta}$ 其中 $d_{i,j}$ 是 i 和 j 之间的距离，参数 δ ⩾0。

（2）距离的指数函数：$w(i,j)=\exp(d_{i,j}^{-\delta})$。

（3）公共边界：w(i,j)= $\left(l_{i,j}/l_i\right)^{\tau}$，其中$l_{i,j}$表示 i 和 j 之间公共边界的长度，$l_i$表示 i 的边界长度（除研究区域的边界外的任何线段）。参数$\tau \geq 0$。

（4）合并的边界和距离加权：$w(i,j)=d_{i,j}^{-\delta}\left(l_{i,j}/l_i\right)^{\tau}$。

（5）相互影响权重。输出权重：$w(i,j)=n(i,j)/n(i,\cdot)$。输入权重：$w(j,i)=n(j,i)/n(\cdot,i)$。$n(i,j)$是从 i 到 j 的空间交互影响；$n(i,\cdot)$是离开 i 的总交互影响；$n(\cdot,i)$是进入 i 的总交互影响。

符号W是标识一个普通的权重矩阵，因此它将有可能包括作为特殊情况的二进制连接矩阵 C 的可能性。注意，对于所有面对象 i 和 j，$w(i,j)\geq 0$，并且，通常$w(i,i)=0$。如果矩阵W按行进行标准化，那么行的总和等于 1，在行中的每个元素除以这个行的总和，记为W^*。于是：

$$w^*(i,j)=w(i,j)/\sum_{i=1}^{n} w(i,j\cdot)$$

Bavaud(1998)描述了一般权重矩阵的性质，并表明了它们如何影响空间系统的性能，如在样本总体区域内任何区域或地点的突出部分。一些学者已研究了其连接性与不同权重矩阵的特征值和特征向量的关系，并确定了它们是如何作用于空间结构特征的。矩阵 C 的主持征值提供了一组区域的连接性指标，同时相应的特征向量的单个元素表明了在整个结构内的每个地点的中心性。这项工作验证了一些基本的选择 C 或 W 的假设。存在邻居选择的进一步影响，这只有在选择表达空间变化的特殊模型时才变得明显。这些影响广泛适用于模型均值和方差的性质。

现在我们能够定义一组新变量，这些变量一般被作为空间平均变量，它们通过执行数据的空间操作，作为原始设置$Z_1, Z_2, \ldots Z_k$的函数而被获取。这些派生的变量一般标记为$WZ_1, \ldots, WZ_k$。前缀W简单地规定为在原始数据上的一些空间操作，它们是按照矩阵被获取的，例如：

$$WZ_1(i)=\sum_{j=1}^{n} w(i,j)z_1(j) \quad \text{j=1, . . . , n}$$

如果W是一个行标准化二进制连接矩阵，那么上式正好是在相邻区域的值的均

值。这种操作对表示 i 区域周围邻居条件是有用的。如果W是一个基于距离函数的行标准化加权矩阵，并且$w(i,j)\neq 0$，那么这个操作对一些形式的数据平滑是有用的。如果W是非标准化二进制连接矩阵，并且$w(i,j)=1$，那么上式是区域 i 和它的邻居的值的总和。这一操作广泛用于诸如此类的我一些聚类的检测方法中。

总之，用在离散空间中的空间分析的数据由于具有数据值的原始数据矩阵和空间对象的位置的标识符组成：

$$\{z_1(i),z_2(i),...,z_k(i)\big|s(i)\}_{i=1,...,n}$$

然而，除此之外至少需要有一个权重矩阵(W)用于获取空间关系。这些空间关系定义了每个 S(i)的所有邻居。成对的$\{s(i),N(s(i))\}$或者{i，，N｝的集合定义了一个图形，从这个矩阵中可以构造新变量$\{WZ_i\}$。

对应于式(1)和式(2)的一般表达式，可以用于离散空间量化在不同尺度上的空间相关性的表达式是：

$$\hat{\gamma}(C^1)=(1/2\big|N(C^1)\big|)\underbrace{\sum_i\sum_j c(i,j)(z(i)-z(j))^2}_{[(i,j)\big|d(i,j)=h\pm\Delta)]} \tag{2-5}$$

且

$$\hat{C}(C^1)=(1/\big|N(C^1)\big|)\underbrace{\sum_i\sum_j c(i,j)(z(i)-\bar{z}(i))(z(j)-\bar{z}(j))}_{[(i,j)\big|d(i,j)=h\pm\Delta)]} \tag{2-6}$$

这里，C^1用于表示连接矩阵C。$\big|NC^1\big|$表示计算中需要的对象的配对个数。在式(1)和式(2)中使用的求和项已被简化，定义在配对的数据值上的约束，是通过C矩阵的元素来使用的。邻居和滞后阶数可由连接矩阵C以及矩阵$C^2(C\times C)$, $C^3(C\times C\times C)$,...的幕定义的步数定义，适当地删除多余的步数。这种方法的半变差函数图和自相关函数图可由规则分布的空间对象获取。式(5)和式(6)的更一般的形式，可以使用W计算，而不是用C。这些在第 7 章中，当对序数级或更高级测量的数据进行广义自相关检验以及名义数据的广义结合计数检验中会遇到通常期望在邻近地点的值趋于相似。空间相关性表现为正相关，并且在

其数据集中体现。

通常期望在临近地点的值趋于相似。空间相关性表现为正相关，并且在数据集合中提到的空间正相关性或自相关性有共同点。对限定的连续空间情形(如任何两点之间的距离趋近零)，可视化任何其他形式的空间相关性都是困难的。在连续空间数据情形，点样本是分离的或像素有足够的大小，如果 z(i)是大(小)的，则 z(j)是小(大)的，这至少是可能的。这就是所谓的空间性为负相关，或称之为空间负自相关。对于连续空间，空间负自相关只发生在一定距离，如山顶和山谷之间。就离散空间来讲，竞争过程可能包括在邻居之间的空间负自相关，除了离散对象之外。例如，邻近的植物可能会为土壤养分竞争，这可能会产生植物大小的空间负自相关。

2.2.3 基于空间分析的拟合模型介绍

基于上文介绍的空间相关性分析的基础上，我们可以对所研究的数据进行相关性分析，通过自相关函数估计研究区域的块金值（小于实际取样尺度引起的变异．表示随机部分的空间异质性。），变程范围（具有相关性的范围），以及基台（空间相关性的变化范围）。因为步长的设置对变程的影响是显著的，需要调整到合适的值。为了保证其稳定性，对于图 2-1 所示的空间带的步长选取需要经行一定的调整与比较，本文中为了兼顾代码自动化和准确性，在实际操作中，假设研究区域的空间相关性变程小于该区域数据点相距最大距离的一半，即变程小于最远数据对距离的一半；同时这样设置的原因也出于对空间分析带内数据对的量的考虑，当数据对的距离为最远距离的一半时，对应的空间带内数据对的量接近峰值，随着距离的增大，数据对的量会急剧减少，其计算结果统计意义的有效性将急剧下降。然后将其空间分析范围 20 等份作为步长，实际情况可以根据数据的分布情况，可以进行变步长优化以保证不同带之间保持相对一致的数据对。

我们假设研究空间分布的数据是彼此关联，即具有空间相关性的。本文中，pCO_2 的数据表现为领近地点值趋于相似，则有空间正相关性。

基于 pCO_2 数据的上述特性，即有以下假设：

1. 两观测点点完全重叠时则相关性为 1。

2. 随着距离的增加，则空间相关性下降。

3. 当两观测点相遇超出了变程时，则相关性为 0，彼此没有关联。

4. 从统计平均上，空间的相关性只和距离相关，即各向同性。

基于以上假设，结合统计平均上，空间有各向同性，本文提出了如下的空间相关性拟合模型：

每个观测点具有辐射半径（变程），即对在其变程内各点有相关性。假设每个观测点等价，且具有统计平均意义上的相同变程。在本文中，因为考虑的为海表 pCO_2，相当于在分界面的行为，故可以将问题抽象为二维平面空间相关性分析问题。即，每个观测点的有效相关性覆盖为圆心为观测点坐标，半径为 1/2 变程的圆来表示。不同观测点之间的相关性可由两圆重叠的面积表示，即二倍圆缺面积。其面积计算表达式如下：

$$A_{overlaping} = \pi r^2 - d\sqrt{r^2 - d^2/4} - 2r^2 \arcsin d/2r \ (d <= 2r) \tag{2-7}$$

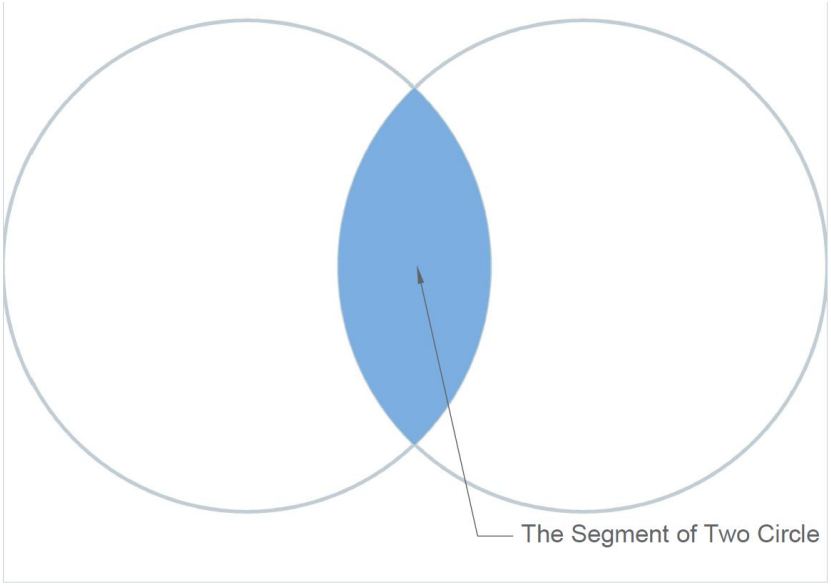


图 2-4

相关性的二维圆缺模型

式中，2r 即为变程值，d 为两观测点相距距离。上式推导见附录。下图为实际情况中，欲通过对实际数据的自相关函数计算后，通过对拟合函数搜索特殊值点来预估 r，继而推出变程的示意图。

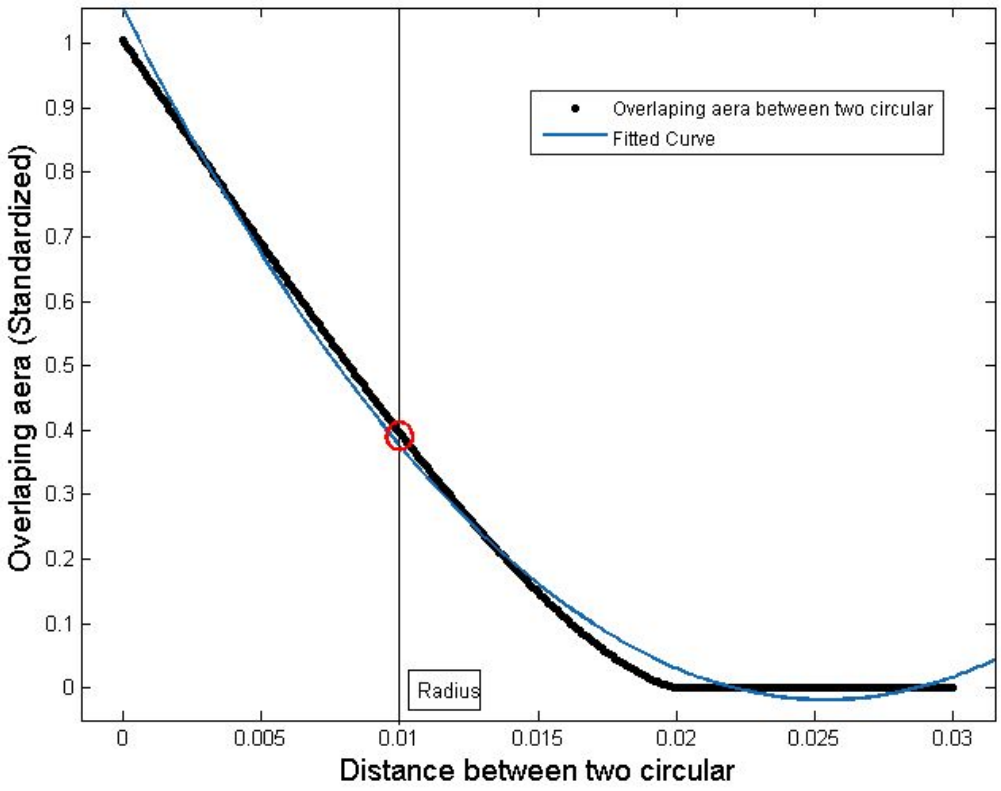


图 2-5

2.2.4 方法小结

本文描述了使用的空间数据分析框架，并汇集了一些模型和工作中出现的空间数据质量问题，如图 2-6。

任何空间数据的集合都是对自然中一个复杂现象的抽象表现。本节从其变量，空间对象和时间准确性以及精度，一致性和完整性等方面进行了概述，关于其数据质量的一般性质。主要目的是提醒分析员，不只是针对出现的问题，还要针对他们对一个研究区域是怎样产生不同的影响，或怎样根据时间或不同的研究区域进行比较。这是针对于用户进行其数据集合质量的评估，并建立合适的目标的。检查整个数据集或许是不可能的，因为这将大大增加数据获取的成本，但对数据的代表性样本应进行检查，以评价其质量。确定哪些误差是关键的和哪些误差不可能导致严重的后果也是必要的。

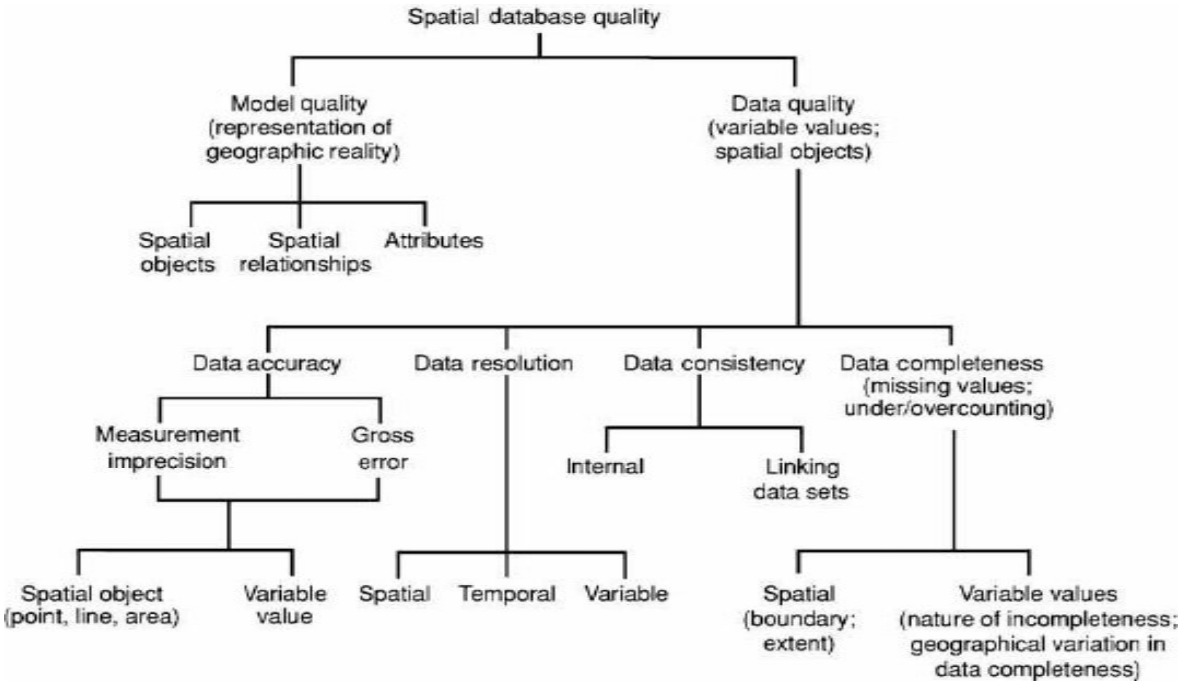


图 2-6

空间数据库方面的质量

3.基于空间分析的 Kriging 估计法

3.1 静态假设（Stationarity Assumption）

3.1.1 严格静态假设

不同的区域的不同的案例研究之间，数据的分布形式具有一定的特点**[1]** 。但是，从统计学上，我们可以做出如下假设，假设区域化的变量 Z(x)的任意分布函数不随着空间点 x 的位移 h 变化而改变。但一般情况这个假设的条件太强，实际上很多数据的分布随空间变化而变化。

3.1.2 二阶静态假设

当局域化变量 Z(x)满足如下条件，则可以称其为二阶静态的。

1.在研究区域内 Z(x)的数学期望都存在，且等于一个常数，即：

$$E[Z(x)]=m;$$

2.在研究区域内 Z(x)的协方差函数存在且静态（即只决定于基本步长 h,而与 x 无关)。

3.2 Kriging 插值法

3.2.1 背景介绍

Kriging 估值法，即空间区域估计或空间区域的插值方法，是空间统计分析的主要内容之一**[4]** 。该方法基于变异函数理论及空间结构分析，在一定区域内（有效范围）对局域化的变量的取值进行无偏最优估计的一种方法。这种估值方法最早是来自南非的矿山工程师 Kriging 和统计学家 Fisher 在 50 年代根据观测值不同的空间位置和其带来不同的相关程度，对每个观测值基于一定模型赋予一定的权重，进行滑动加权平均的处理，以此来估计特定位置点上的平均值的方法。

假设所研究的区域已经充分的进行了抽样，即可以当做从研究区域抽样的平均数的数学期望就代表该区域实际观测值的平均值。若该观测值服从正态分布，则

样品的平均观测值 x 对于所研究区域平均观测值 y 的回归直线方程，就应当代表 x 轴与 y 轴夹角的平分线 x=y，因为有 $E(x|y)=y$。但实际，数据集得到的是一条斜率 β （ β<1），且过（m,m）点的直线。这里 m 是 x 的均值，也是 y 的均值。则此回归直线方程为：

$$y-m=\beta(x-m)$$

然而实际情况，m 通常未知，只能用观测值的平均值去表示。于是，上式为：

$$y=\beta x+(1-\beta)m=\beta x+(1-\beta)*\frac{1}{n}\sum_{i=1}^{n}x_i$$

在某些特殊情况下，设 x 为 x(i)中的某个点（举例：x=x(1) ），则可得:

$$y=(\beta+\frac{(1-\beta)}{n})x_1+(1-\beta)*\frac{1}{n}\sum_{i=2}^{n}x_i$$

其中，令 $\beta+\frac{(1-\beta)}{n}=\lambda_1$，$\frac{(1-\beta)}{n}=\lambda_i$，i=2，3，...，n。综上：

$$y=\sum_{i=1}^{n}\lambda_i x_i \qquad (3\text{-}1)$$

其中 $\sum_{i=1}^{n}\lambda_i=1$。

3.2.2 Kriging 意义

变异函数模型的引入，其一个重要作用是用于空间上随机变量的估值或称作内插，即通过对某一化学生物或非这一类因子在空间上已充分抽样的数据来推测任意位置上未抽样点上的数值。其本质是使用区域化变量的最初数据和变异函数的模型结构等特点，对待估点的区域化变量的值进行了线性无偏最优估计的处理。从数学角度来说，该方法可以说是对空间结构数据的最优估计，线性，无偏内插估计（Best Linear Unbiased Estimation）方法。具体来说，它是根据待估值点有限范围内（变程）若干已观测的样点数据，结合了其分布形状，大小，和空间的相互位置关系等因素，进行了一种线性无偏最优估计。

Kriging 估值法和一般的估计不一样，它最大可能的利用了空间取样所提供的所有信息。在估计待测点数值时，它不仅分析了落在该样点的数据，而且还考虑

了变程范围内样点的数据；不仅分析了待估值点与邻近已知采样点的空间距离，而且还考虑了不同临近样点之间彼此的位置关系。再加上现有的观测值空间结构分布的特征，使这种估计方法比其他传统的方法更精确，符合实际，而且成功避免系统误差的出现，并同时给出估计误差和其精度。不过，如果其变异函数和空间相关分析的结果显示其区域化变量的空间相关性不明显，则该空间局部估值方法的并不适用。

3.2.3 Kriging 优点

1.无偏性

2.估计方差最优性：用 Kriging 模型计算的权重系数构成的估计值不仅在统计上是无偏的，而且其产生的估计方差最小，等于 Kriging 方差。

3.减弱了 Declustering Effect: 在该方法中，不会因为一些样点过于聚集在一个小区域内而增大其权重系数。

4.屏蔽效应：当块金值（Nugget Value）很小或几乎不存在时，已知样点的 Kriging 权重系数的大小决定于屏蔽效应。

5.权值可为负值：Kriging 权重系数从定义上说是可正可负的，这样可以保证其估值超出信息样品的最大最小值范围。

6.其拟合的变异函数模型，统计平均上，考虑了各向异性的情况。

3.2.4 Simple Kriging 推导

简单 Kriging 法是用于具有二阶静态且均值是已知的随机变量的一种估计方法。设变量 Z（x）满足二阶静态假设，则其具有如下性质：

1. $\mathrm{E}[Z(x)]=\mathrm{m};$

2. $Cov[Z(x+h),Z(x)]=C(h);$

3. $Var[Z(x)]=C(0)=\sigma^2$

由于是二阶静态变量，具有有限方差值，所以变量的空间变异分布结构可以用协方差函数来表示，也可以用其变异函数来表示，它们之间的关系为：

$$\gamma(h)=\sigma^2-C(h) \qquad (3\text{-}2)$$

简单 Kriging 法的估计公式为：

$$Z^*(x_0)=m+\sum_{i=1}^{n}\lambda_i[Z(x_i)-m] \qquad (3\text{-}3)$$

式中 $Z^*(x_0)$ 表示在 X0 位置的估计值，Z(x)代表在 Xi 位置的测量值，λ_i 代表估值分配给 $Z(x_i)$ 残差的权重，n 表示用于整个估计过程的观测值的个数，m 是均值。

根据 Kriging 法，应满足估计值的无偏性要求，即：

$$E[Z^*(x_0)-Z(x_0)]=0 \qquad (3\text{-}4)$$

将(3-3）带入到(3-4）得：

$$\begin{aligned}&E\{m+\sum_{i=1}^{n}\lambda_i[Z(x_i)-m]-Z(x_0)\}\\&=m+\sum_{i=1}^{n}\lambda_i\{E[Z(x_i)-m\}-E[Z(x_0)]\end{aligned} \qquad (3\text{-}5)$$

根据性质（1），上式化简为：

$$E[Z^*(x_0)-Z(x_0)]=m+\sum_{i=1}^{n}\lambda_i(m-m)-m=0 \qquad (3\text{-}6)$$

很显然，无论怎么取其权重，都可以满足无偏。

估计值误差的方差，见下式：

$$\begin{aligned}&Var[Z^*(x_0)-Z(x_0)]=E[\sum_{i=1}^{n}\lambda_i Z(x_i)-Z(x_0)]^2\\&=E\{[\sum_{i=1}^{n}\lambda_i Z(x_i)]^2-2\sum_{i=1}^{n}\lambda_i Z(x_i)Z(x_0)+[Z(x_0)]^2\}\\&=E\{[Z(x_0)]^2\}-2E[\sum_{i=1}^{n}\lambda_i Z(x_i)Z(x_0)]+E[\sum_{i=1}^{n}\lambda_i Z(x_i)]^2\\&=\sigma^2-2\sum_{i=1}^{n}\lambda_i C(x_0-x_i)+\sum_{j=1}^{n}\sum_{i=1}^{n}\lambda_i\lambda_i C(x_i-x_j)\\&=S\end{aligned} \qquad (3\text{-}7)$$

因为 Kriging 法要求估计值误差的方差最小，由此条件可以推出权重的计算 Kriging 线性方程。

$$\begin{aligned}&S=\sigma^2-2[\lambda_1 C(x_0-x_1)+\lambda_2 C(x_0-x_2)+\ldots+\lambda_n C(x_0-x_n)]\\&+\lambda_1[\lambda_1 C(x_1-x_1)+\lambda_2 C(x_1-x_2)+\ldots\lambda_n C(x_1-x_n)]\\&+\ldots\\&+\lambda_n[\lambda_1 C(x_n-x_1)+\lambda_2 C(x_n-x_2)+\ldots+\lambda_n C(x_n-x_n)]\end{aligned} \qquad (3\text{-}8)$$

为了得到方差最小值，即仿照最小二乘，求一阶偏导为 0,

$$\frac{\partial S}{\partial \lambda_1} = -2C(\mathrm{x}_0 - \mathrm{x}_1) + [\lambda_1 \mathrm{C}(\mathrm{x}_1 - \mathrm{x}_1) + \lambda_2 \mathrm{C}(\mathrm{x}_1 - \mathrm{x}_2) + ... \lambda_n \mathrm{C}(\mathrm{x}_1 - \mathrm{x}_n)] + ... = 0; \quad (3\text{-}9)$$

...

依次类推，可以列出一个线性方程组：

$$\sum_{i=1}^{n} \lambda_i C(x_i - x_j) = C(x_0 - x_j), j = 1,...,n \quad (3\text{-}10)$$

式中 $C(x_i - x_j)$，即待估点和观测点距离为 x_i 和 x_j 之间协方差值，求解（3-10）的方程组，既可以得出 n 个未知值的权重系数 $\lambda_1,...\lambda_n$。将这些权重系数带入（3-3），就可以计算估计值。

同时，也可以计算估计值的 Kriging 误差，从式（3-7）和（3-10）得：

$$\sigma_{SK}^2 = \sigma^2 - \sum_{\mathrm{i}=1}^{n} \lambda_{\mathrm{i}} \mathrm{C}(\mathrm{x}_0 - \mathrm{x}_{\mathrm{i}}) \quad (3\text{-}11)$$

3.2.5 Ordinary Kriging 推导

普通 Kriging 法，其定义要求变量满足二阶静态假设，比如变量需要满足固有假定条件，即对所有的 x 和 h，有如下统计性质：

1. $E[\mathrm{Z}(\mathrm{x}+\mathrm{h}) - \mathrm{Z}(\mathrm{x})] = 0;$

2. $Var[Z(x+h) - Z(x)] = 2\gamma(h)$

3. $\gamma(h) = \sigma^2 - C(h)$

推导过程与简单 Kriging 方法类似，估计式：

$$Z^*(x_0) = \sum_{i=1}^{n} \lambda_{\mathrm{i}} Z(x_i) \quad (3\text{-}12)$$

该公式估计误差的均值为：

$$\begin{aligned} \mathrm{E}[Z^*(\mathrm{x}_0) - \mathrm{Z}(\mathrm{x}_0)] &= \mathrm{E}\{\sum_{i=1}^{n} \lambda_{\mathrm{i}} Z(x_i) - Z(x_0)\} \\ &= \sum_{i=1}^{n} (\lambda_{\mathrm{i}} - 1) E[Z(x_0)] \end{aligned} \quad (3\text{-}13)$$

为了保证无偏估计，则有 $E[Z^*(\mathrm{x}) - \mathrm{Z}(\mathrm{x})] = 0$，即：

$$\sum_{i=1}^{n}\lambda_{\mathrm{i}}=1 \tag{3-14}$$

又：

$$\begin{aligned}\mathrm{S}&=\mathrm{Var}[Z^{*}(\mathrm{x}_0)-Z(\mathrm{x}_0)]\\&=\sigma^{2}-2\sum_{i=1}^{n}\lambda_{\mathrm{i}}\mathrm{C}(\mathrm{x}_0-\mathrm{x}_{\mathrm{i}})+\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_{\mathrm{i}}\lambda_{\mathrm{j}}\mathrm{C}(\mathrm{x}_{\mathrm{i}}-\mathrm{x}_{\mathrm{j}})\end{aligned} \tag{3-15}$$

式中$\gamma(h)$表示其变异函数，由上文引入$\sum_{i=1}^{n}\lambda_{\mathrm{i}}=1$的限制条件，由拉格朗日乘数法，引入μ：

$$\mathrm{S}=\sigma^{2}-2\sum_{i=1}^{n}\lambda_{\mathrm{i}}\mathrm{C}(\mathrm{x}_{\mathrm{i}}-\mathrm{x}_0)+\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_{\mathrm{i}}\lambda_{\mathrm{j}}\mathrm{C}(\mathrm{x}_{\mathrm{i}}-\mathrm{x}_{\mathrm{j}})+2\mu(\sum_{i=1}^{n}\lambda_{\mathrm{j}}-1) \tag{3-16}$$

同理，为使估计方差最小，由此得出计算权重系数的线性方程组，进而求解方程组：

$$\begin{vmatrix}C_{11} & \cdots & C_{1n} & 1\\ \vdots & & \vdots & \vdots\\ C_{n1} & \cdots & C_{nn} & 1\\ 1 & \cdots & 1 & 0\end{vmatrix}\begin{vmatrix}\lambda_1\\ \vdots\\ \lambda_n\\ \mu\end{vmatrix}=\begin{vmatrix}C_{01}\\ \vdots\\ C_{0n}\\ 1\end{vmatrix}$$

最后得出：

$$Z^{*}(x_0)=\sum_{i=1}^{n}\lambda_{\mathrm{i}}Z(x_i) \tag{3-17}$$

$$\sigma_{OK}^{2}=\sigma^{2}-\sum_{\mathrm{i}=1}^{n}\lambda_{\mathrm{i}}C(\mathrm{x}_0-\mathrm{x}_{\mathrm{i}})+\mu \tag{3-18}$$

（注意;按理论要求，对于分配给$x_1,x_2,\ldots,x_n$的权重，-1，$\lambda_1,\lambda_2,\ldots,\lambda_n$，应保证估计方差是非负的，即限制性正定条件。）

4. 海表二氧化碳分压数据的不确定度分析

4.1 海表二氧化碳分压数据库

本文中的数据库主要来源于两部分，①为 2009 年在中国东海（ECS）的一组走航测量的数据，时间间隔约为 2 个月，数据量约为 9000 个数据点。②数据主要来自于 LDEO 数据库，LDEO 来自于 the Carbon Dioxide Information and Analysis Center at the Oak Ridge National Laboratory, Oak Ridge, TN. 该数据库包含自 1957 年起截止 2015 年，大约 9 百万不在厄尔尼诺年内的走航二氧化碳分压的测量（该数据库只包含使用水汽平衡法测量出的数据）。

4.1.1 数据的来源（Data）

图 4-1 为 August 2009 年在中国东海为期一个月的实测二氧化碳分压数据，由数据的分布可以看出，其大致覆盖了整个空间，但是存在有几个区域有明显的数据不足的情况，所以，凭借该组数据对该空间的平均二氧化碳分压的评估存在较大的不确定度。借助于空间相关性的分析和 Kriging 估值法的结合，可以给出该区域的无偏最优估计和其最优方差的大小。在夏季，中国东海主要受四个水团影响，相对低温的北部黄海水；扬子江涌入水；和中间陆架区域的结合了一支陆架外黑潮流的南部台湾暖流。（高温高盐度）

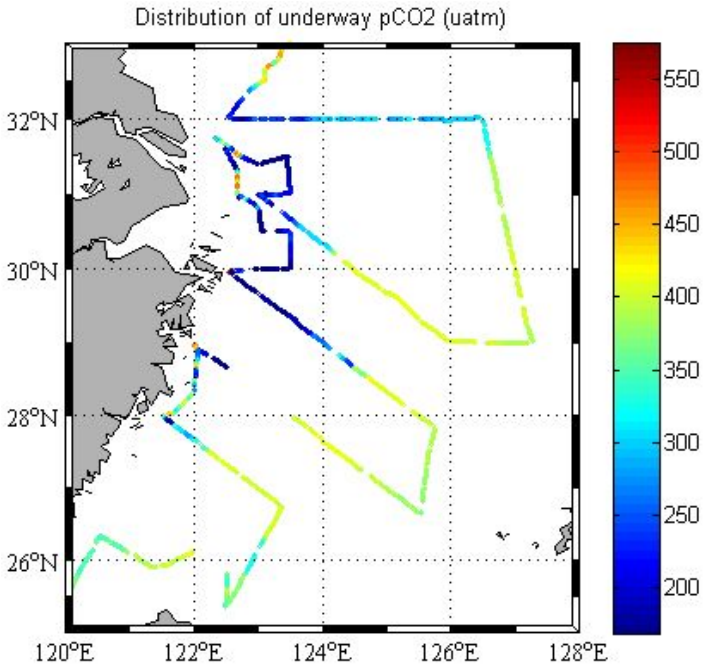


图 4-1

东海走航 pCO_2 数据（August 2009）

4.2 计算实现(Methodology)

4.2.1 空间相关分析，拟合函数估计

对于所研究的区域，即对应一组空间数据，如上图，第一步我们假设数据在分布上具有空间的相关性，彼此之间的联系与距离相关。通过第二章介绍的空间自相关分析方法，求出待研究数据的自相关函数变化，如下图。然后通过...的球状模型拟合函数对数据参数进行估计。这里为了使程序自动化，这里通过特殊值点，即自相关函数的 $1/2$ 点作为特征点进行搜索，从而对参数进行预估。本例中：通过第二章（3）式计算 C(h), (4)式计算 R(h)。然后由第三章介绍的方法，基台值 C=C(0)=7640.7 。通过对归一化的自相关函数进行多项式拟合，因为理想的球状模型中当 d=radius 时，R(radius) =0.3125*R（0），从拟合的归一化自相关函数中反推出半径 radius =1.09985，则有 a=2*radius=2.1997。

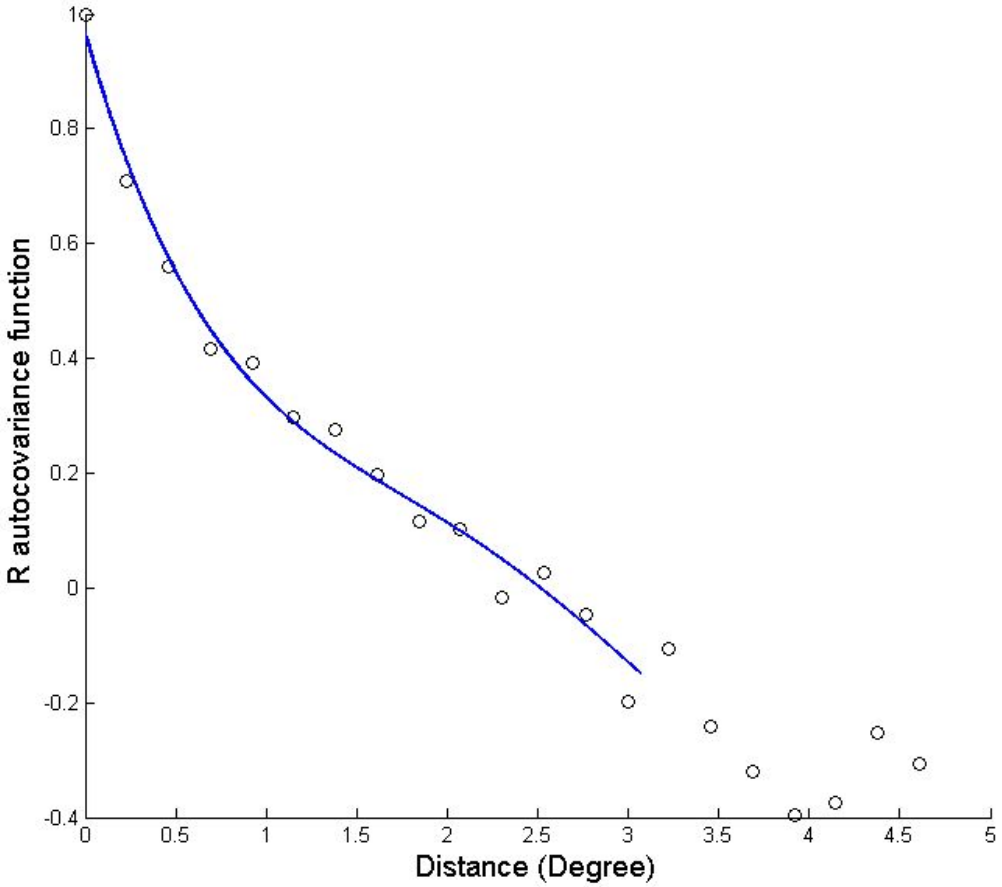


图 4-2

东海走航 pCO_2 数据归一化空间自相关函数分布

代入到球状模型拟合函数：$$C(h)=\begin{cases} C_0+C & h=0 \\ C[1-(\frac{3}{2}*\frac{h}{a}-\frac{1}{2}*(\frac{h}{a})^3] & 0<h\le a \\ 0 & h>a \end{cases} \quad (4\text{-}1)$$

或圆缺模型：

$$C(h)=\begin{cases} C_0+C & h=0 \\ \frac{C}{\pi r^2}[\pi r^2-d\sqrt{r^2-d^2/4}-2r^2\arcsin d/2r] & 0<h\le a \\ 0 & h>a \end{cases} \quad (4\text{-}2)$$

上式为第二章中，本文介绍的一个基于二维平面空间分析的圆缺模型拟合函数，实际应用中，有多种拟合模型函数，其中应用最广泛的为球状模型拟合函数，其原理和圆缺模型类似，使用的是球缺来表征相关性，适用于空间三维的情形。除此外，还有一些诸如指数模型，高斯模型，霍尔模型等拟合函数模型。本文中，分别测试了圆缺和球缺模型，然后通过对数据的空间分析估计的参数构造不同的模型，然后准备进行最后的空间估值和不确定度分析，具体实现如下。

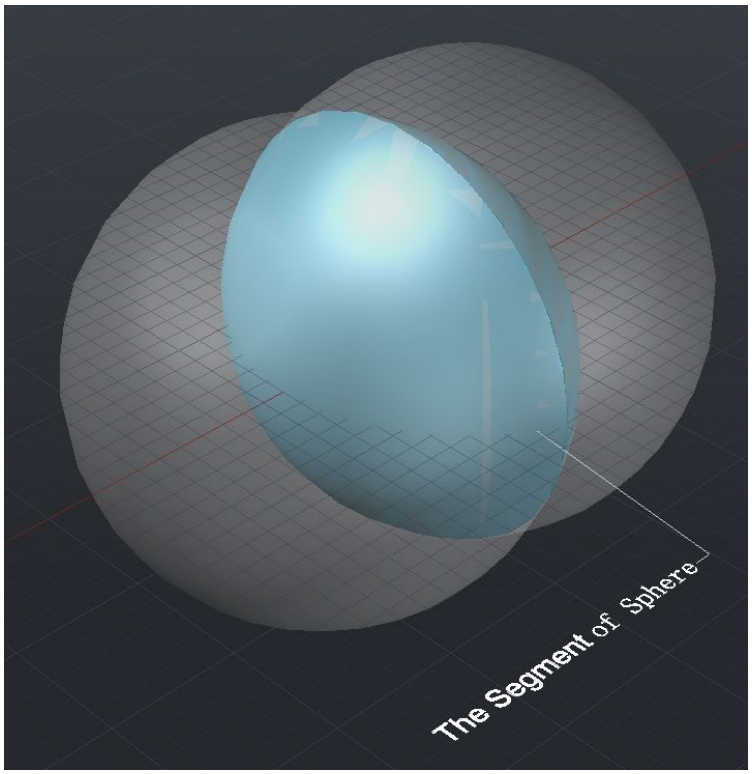


图 4-3 球状模型拟合函数示意图

4.2.2 Kriging 估值法具体算法

通过得出的拟合函数，对应于待估值点 x_0 计算下图矩阵中的各个元素，式中 $C(x_i-x_j)$，即距离 x_i 为和 x_j 之间的协方差值。

$$\begin{vmatrix} C_{11} & \cdots & C_{1n} & 1 \\ \vdots & & \vdots & \vdots \\ C_{n1} & \cdots & C_{nn} & 1 \\ 1 & \cdots & 1 & 0 \end{vmatrix} \begin{vmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu \end{vmatrix} = \begin{vmatrix} C_{01} \\ \vdots \\ C_{0n} \\ 1 \end{vmatrix} \tag{4-3}$$

然后解方程组得出权重矩阵。$\{\lambda_1, \lambda_2, \ldots \lambda_n, \mu\}$。

注意：为使矩阵有唯一解则上式矩阵$\left[C(x_i - x_j)\right]$必须满秩，但是由于实际数据中，会有一些观测点拥有一个相同的地理坐标，则对应协方差矩阵就有两行或多行相同元素，导致矩阵不具有唯一解，针对这个办法，本文提出一个解决办法，即在原有空间坐标的基础上，对其经纬度经行一段随机的微小位移调整，（调整范围小于实际坐标的最后一位有效数字的位数，但不能过于小，因为变量精度的原因，当微小调整过小时，近似于原点，矩阵依然不满秩。这样，因为生成的随机数是服从正态分布的，在统计平均下，其产生的误差的和会趋于 0 而不对最终结果产生影响。经过这样处理的空间坐标，对应的空间协方差矩阵有解。同时并不影响空间相关性的块金效应的计算，因为随机的小位移的偏离的范围远小于相关分析的步长，故影响很小。

即,普通 Kriging 线性估计量为:

$$Z_0^* = \sum_{i=1}^{n} \lambda_i Z(x_i) \tag{4-4}$$

对应该点的估值方差为：

$$\sigma_{OK}^2 = C(0) - \sum_{i=1}^{n} \lambda_i C(x_0, x_i) + \mu \tag{4-5}$$

4.3 计算结果与其不确定度报告

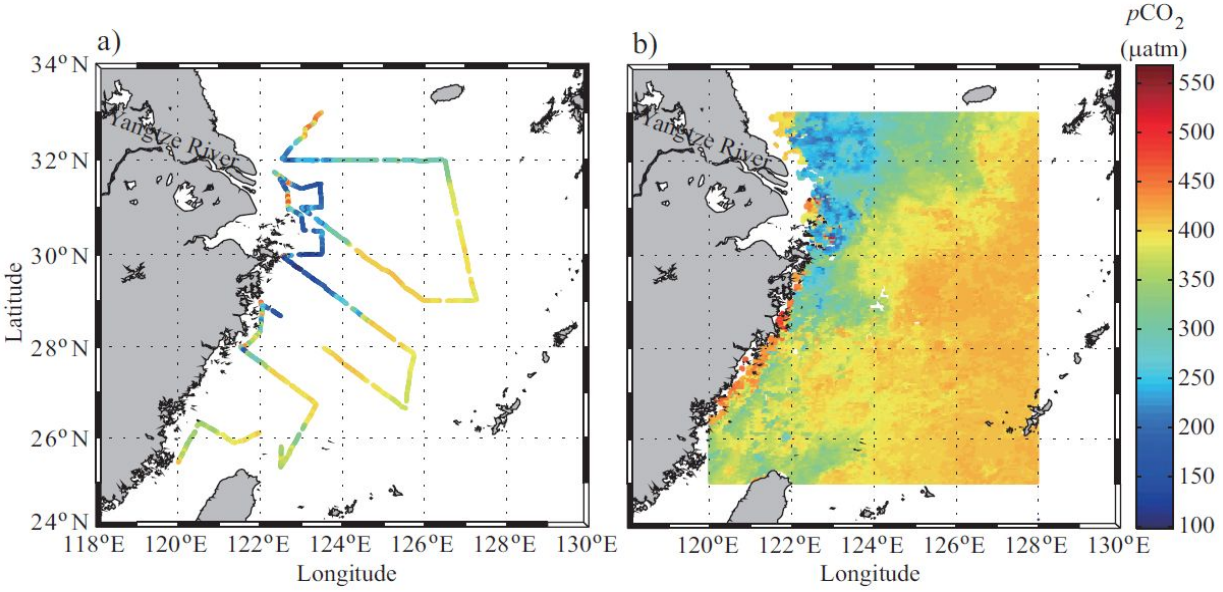


图 4-4

a)为图 4-1 走航数据(Wang, 2014)　　b)为同期对应的遥感数据(Bai, 2015)

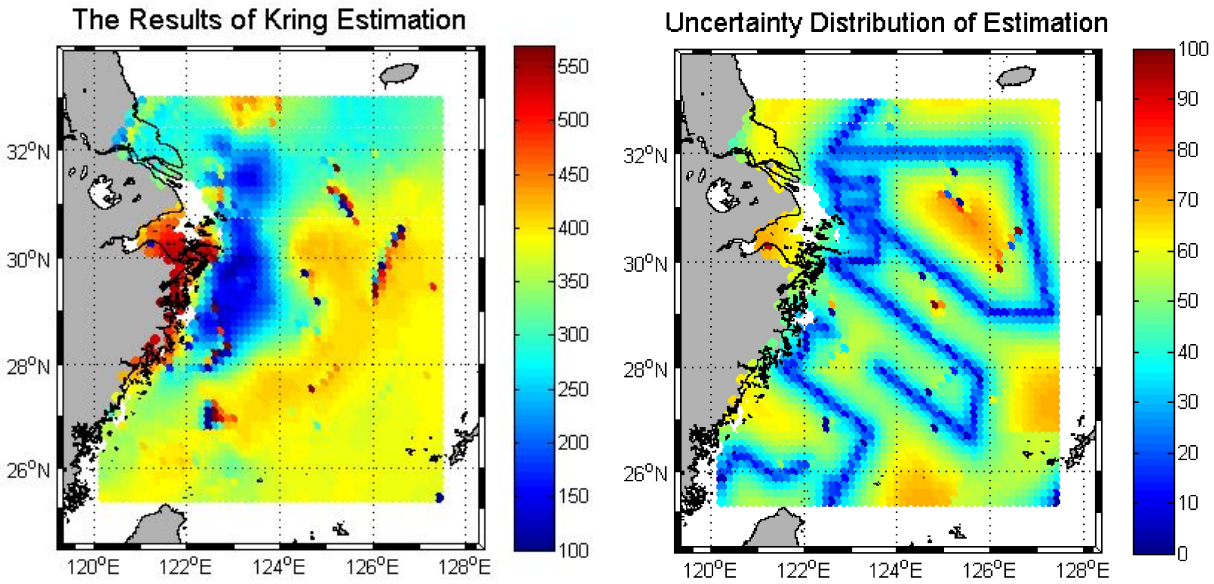


图 4-5

左图为由(August 2009 图 4-1)所示的数据经过 Kriging 估值法处理后得到的该研究区域 pCO_2 的估值空间分布。右图为对于 pCO_2 估值点的不确定度评估。

上述估值结果经过空间区域处理，只保留了海平面的数据，因为插值结果是不进行区域选择的；所以，陆地上的插值结果不具有实际的物理意义，故需要进行

筛选。具体的筛选方法为，将海岸线数据和空间部分的端点相连构成闭合多边形，从而可以通过 Inpolygon 函数进行处理。下图为**[13].**中的走航数据和当时通过遥感同时期测得的全空间的数据。可以作为对走航数据进行空间估值后的一个有效的比对评估。由图 4-4（a）和下方右图进行比较可以得知，基本上插值的结果与同期实测的遥感数据吻合的比较好，但依然存在一些极个别的异常点，多出现在数据分布严重不足的区域。为了保证该组数据的实际有效性，建议在后期实际应用中应予以剔除异常点。

同时如图 4-4（b）所示，可以看出，估值的误差分布极大的受走航的数据分布影响。即，空间数据采样所获取的有效样本量对于该空间的估值误差的影响是显著的。当估值点在实际采样点附近时，则误差很小，小于$\pm 10\,\mu atm$。而随着估值点距离采样点的的距离增加，即对于空间样本覆盖较差的区域，我们可以看到，其估值带来的不确定度的风险是显著的，恰好对应图中的偏红黄色区域。

1. 通过 Kriging 估值后，对网格化空间的pCO_2数据，一个网格内空间本身的变异由下式计算：

$$\sigma_{\mathrm{s}}^2 = \frac{1}{N}\sum_{i=1}^{N}(Z_i - \overline{Z}), \tag{4-6}$$

其中，N 是该网格内数据点个数，Z_i是单个点值，$\overline{Z}$ 是真值，由该网格内空间样本的平均值估计。

2.估值带来的误差，即样本数据不足（Undersampling）导致的不确定度由下式计算。

$$\sigma_u^2 = C(0) - \sum_{\mathrm{i}=1}^{n}\lambda_{\mathrm{i}} C(\mathrm{x}_0, \mathrm{x}_{\mathrm{i}}) + \mu\,, \tag{4-7}$$

即 Kriging 估值的估值方差。

3.采样本身对数据的初处理和仪器误差，即分析误差: E_m^2

因为采样数据的分析误差，空间本身的变异和估值方差彼此之间是独立的，故，针对研究区域的整体不确定度为**[11]**：

$$\sigma_T = \sqrt{E_m^2 + \sigma_s^2 + \sigma_u^2}\,, \tag{4-8}$$

由于实际情况中，一方面分析误差造成的不确定度远小于其他类不确定度，在本文中不予考虑，故研究区域的整体不确定度即为空间变异以及采样不足对空间

估计带来的不确定度的平方和。

4.4 实际应用小结

以上提出的基于空间相关性分析，构造拟合函数模型，结合 Kriging 估值法的分析手段，在实际应用中也存在一些问题，下面将附上该方法的有效性检测过程和相关局限性分析。走航数据分布分布和其空间自相关函数分布。

1.一维情形，航次：PYXS-March-2012 截取了一段数据。

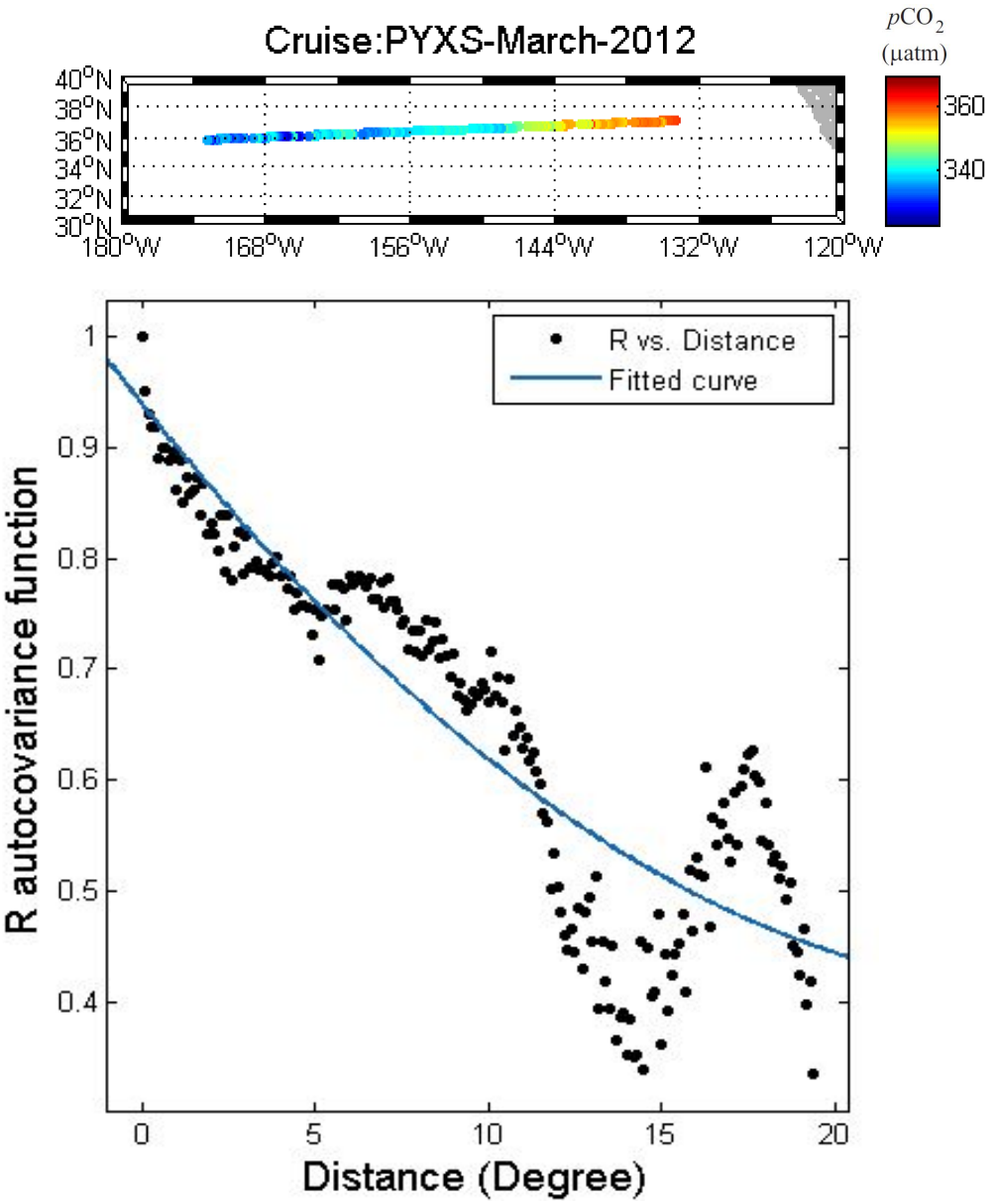
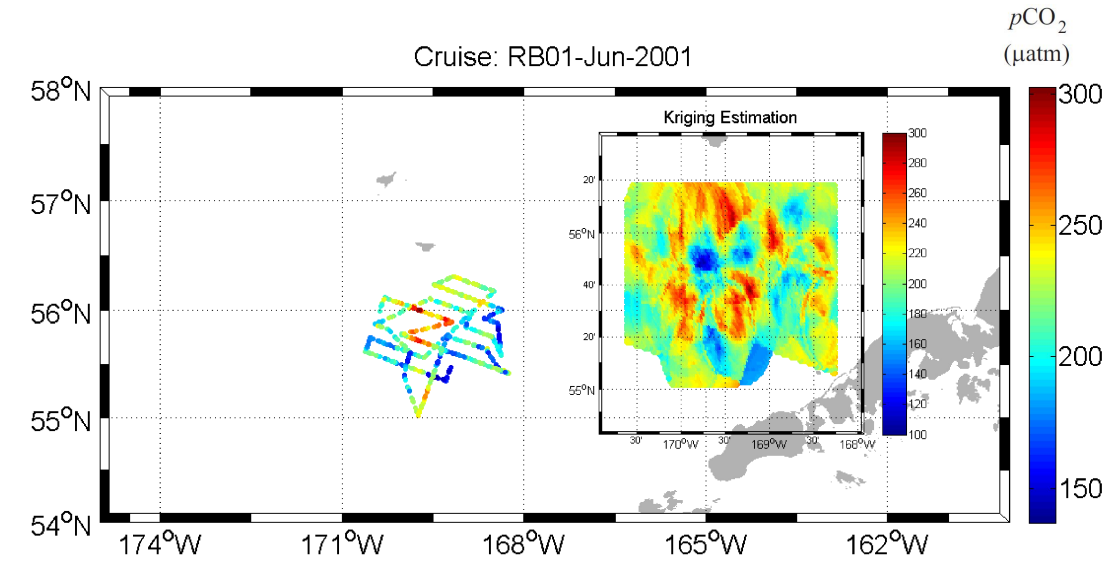


图 4-6

PYXS—March-2012 航次的分析，一维结构

2.二维情形，航次：RB01-Jun-2001



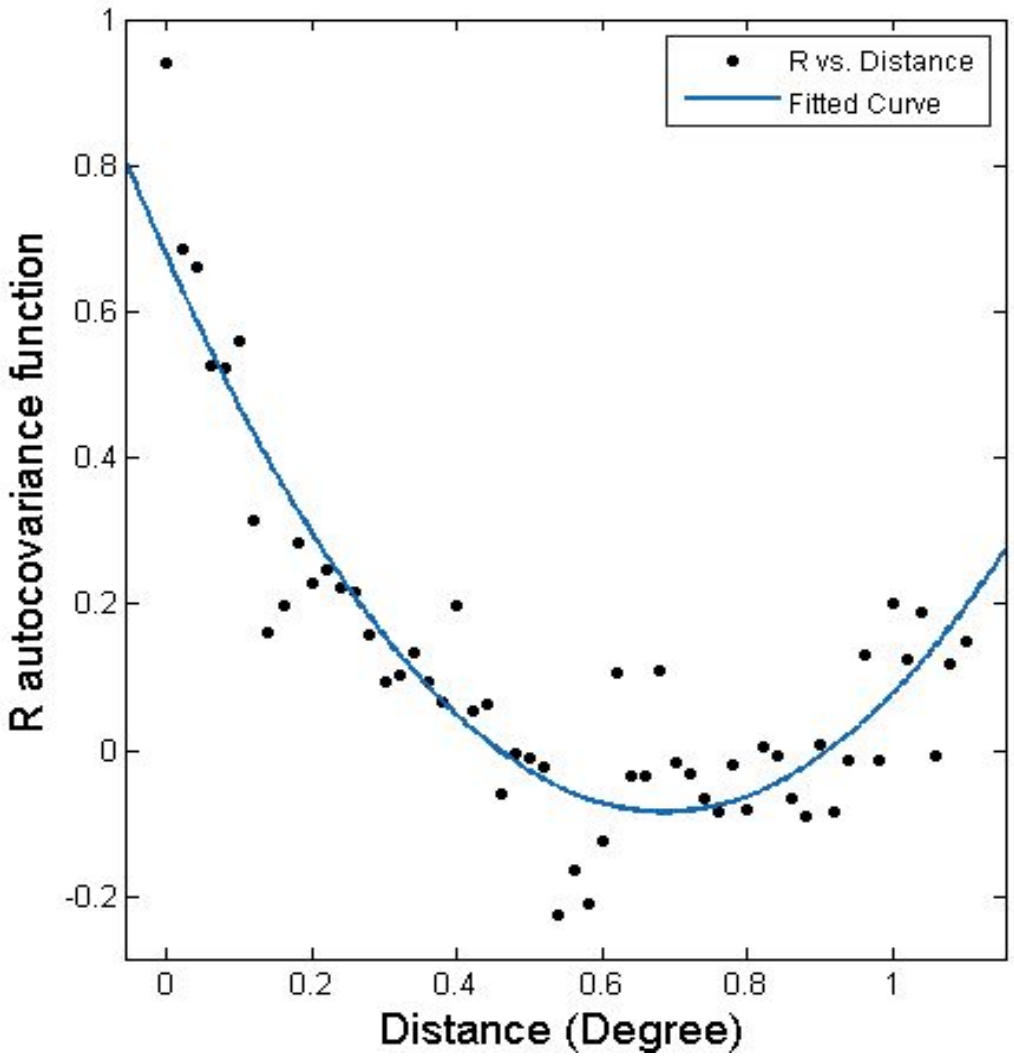


图 4-7

RB01—Jun-2001 航次的分析，二维结构

3. 二维特殊情况（存在空间异质性）

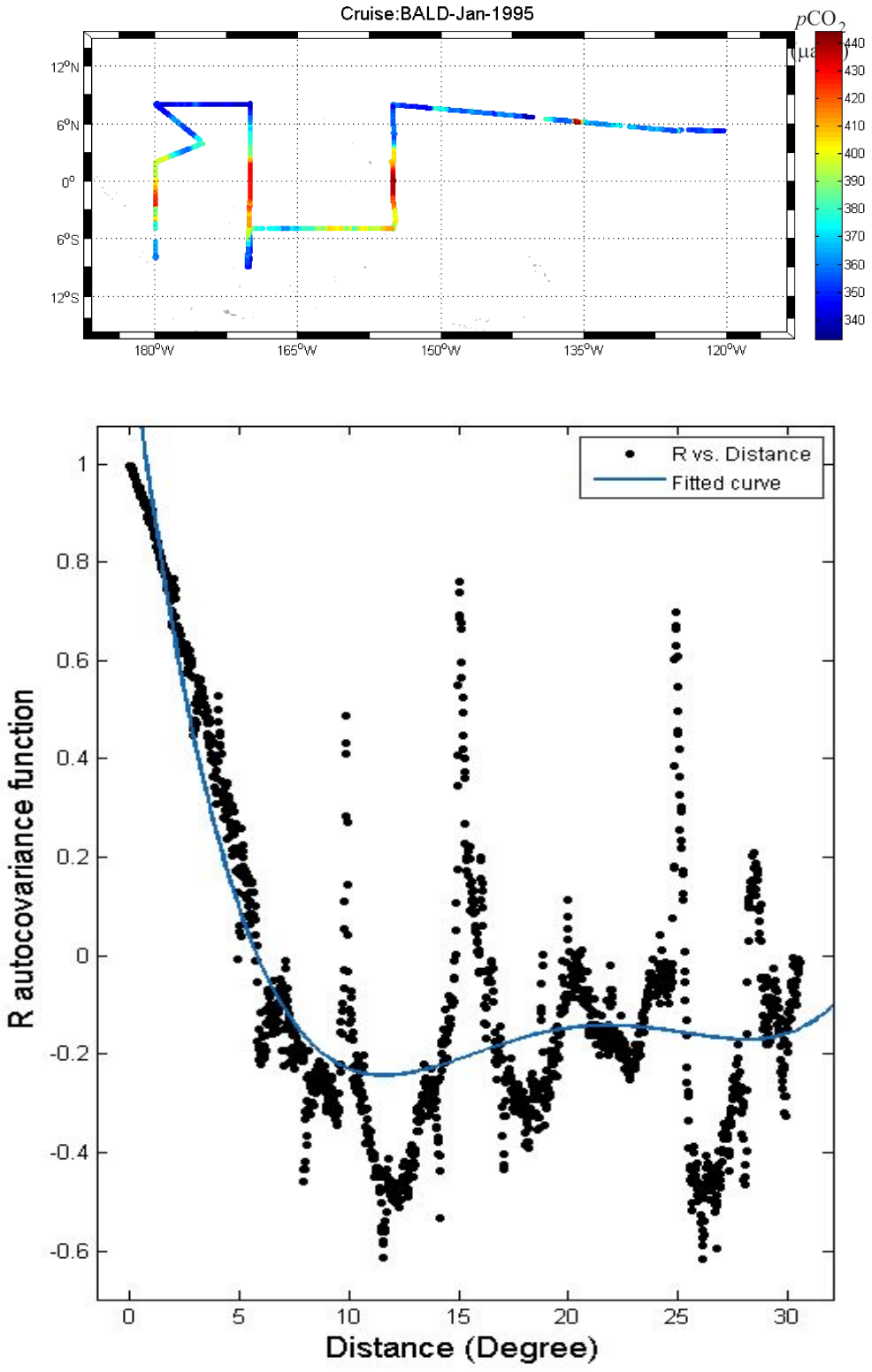


图 4-8

BALD—Jan-1995 航次的分析，具有空间异质性的半二维结构

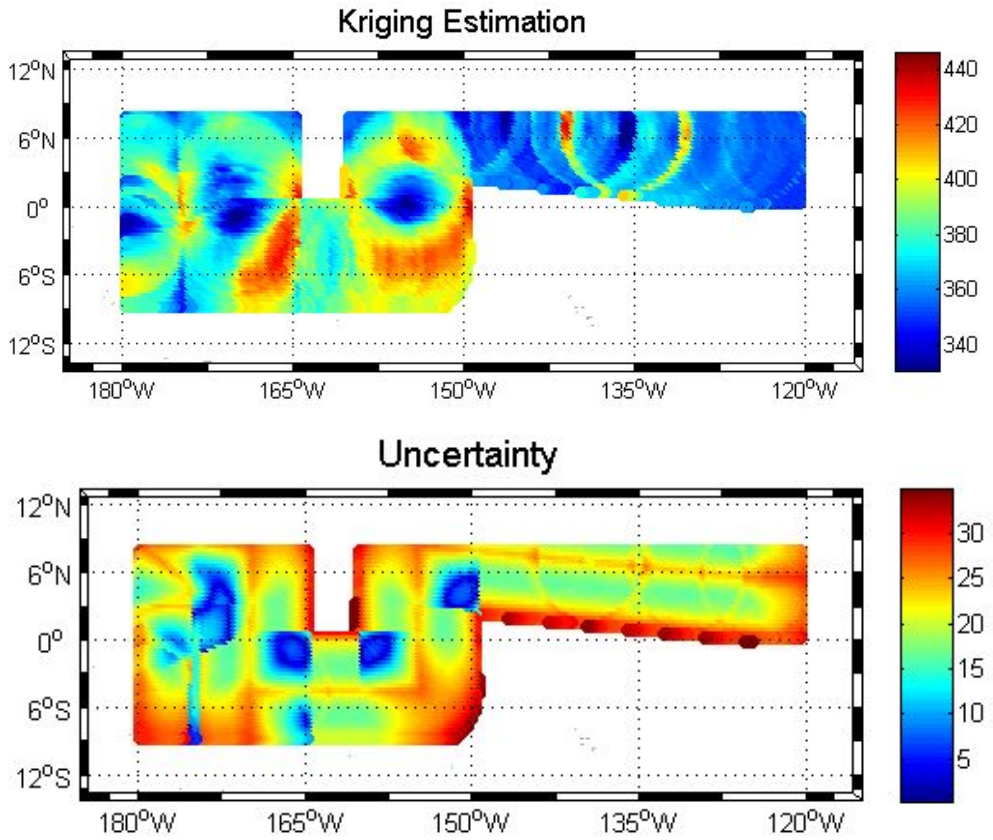


图 4-9

BALD—Jan-1995 航次的分析，具有空间异质性的半二维结构

以上展示了几个不同案例的空间相关性分析。大体上符合理论模型，在案例 3 中，出现了当距离为 10,16，25 度时又出现了一些极大相关值点。这是这个研究区域为太平洋的赤道附近，在太平洋的赤道附近其海表 pCO_2 分布并不是空间各向同性的，而呈条带状，即随纬度的变化影响显著，而经度对其影响较少。则在该区域存在空间异质性，对于该问题的目前尚没有很好的解决方法，建议解决方案就是细分研究区域，将异质性的空间继续划分为一维的条带，从而避免了异质性，保证其空间相关分析的稳定性。但，及时在实际应用中，对于这样研究区域具有明显的空间异质性的情况，该相关性分析依然可以很好的计算出最小的变程范围，如图，可知最小的变程约为 5°。即依然可以通过拟合函数进行后面的 Kriging 估值法。但是对潜含条带状的空间结构并不能很好的进行估值。

5. 总结

本文主要的内容即，基于空间统计学中的空间相关性分析，结合理想的拟合函数模型进行参数预估，由估计而成的参数结合 Kriging 估值法的分析手段，对所研究的区域进行 Kriging 估值，并给出其估值带来的不确定度报告。该方法较成功的解决了海表 pCO_2 数据的空间变异及其不确定度分析。其基于严格的数学统计模型上的推导证明，定量分析了研究区域全空间的空间变异及其不确定度分布，且该方法具有无偏最优估计等特点。另外结合了自动拟合和特殊点的参数搜索法的计算代码，使得针对不同的空间问题可以进行自动化处理较便捷的得出结果。

但，该方法也有一些局限性，在实际过程中，例如在诸多广泛使用的 GIS 系统中，Kriging 估值的参数需要进行人工的测试和调教，从而筛选优化出较好的参数进行随后的空间估值，但本文中，基本上采取了全自动的一步完成，但其代码和计算方法本身并不具有对于不同问题广泛的适应性，所以，可以预见本方法在一些特殊情况中会有较大的系统误差的引入。但，因为本文中的空间相关性分析，假设了空间数据在统计平均上是各向同性的。这种假设从统计平均上来讲是满足的，对于这类问题，本文中的方法具有很好的结果和应用。但对于部分采样严重不足不能反映空间二维结构的情形，则会遇到很多问题，如应用中展示的不同的案例分析。

同时，基于空间分析的空间相关性分析，也是对于空间数据的时间分析处理的检验手段。如：对一定空间研究区域，不同时间段采样获得的数据具有时间带来的变异性。而在如**[8]**中，采取了去季节化的处理手段，即选取观测较多的连续年，通过平均的月偏移来表示其时间变异，这种处理手段从统计整体平均上是具有一定有效性的，但是通过本文中的空间相关性分析可知，在较大数据量的统计平均来修正时间变异（去季节化），对于较小尺度的空间较精细的结构是具有逆向影响。如下例：

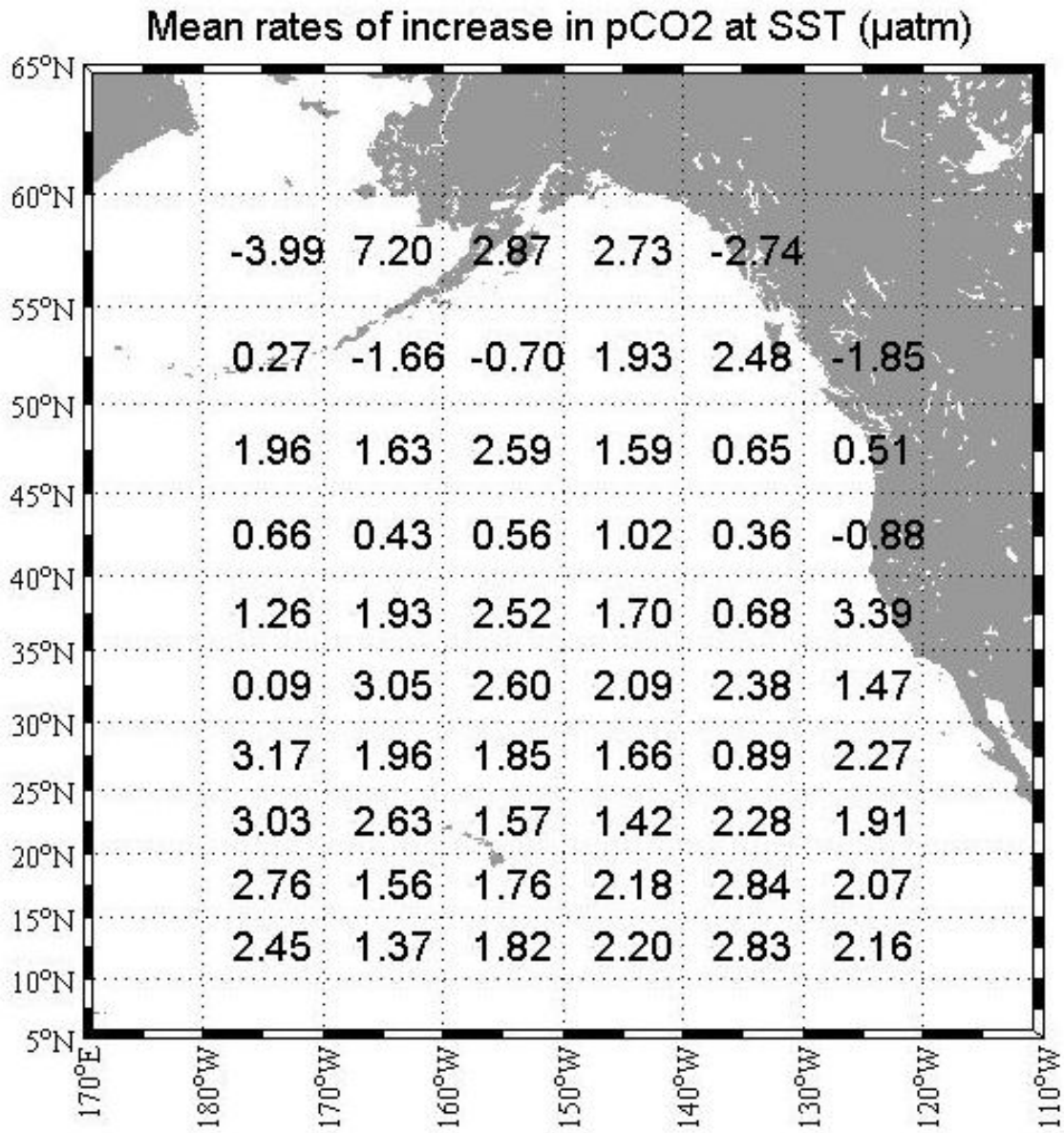


图 5-1

参见[Takahashi et al., 2002], Deseasonlization.

上图为参照**[8]**中的 Deseasonlization 法求出的不同区域（分辨率 10°×5°）的 pCO_2 平均年增长率，参考年选取为 2000 年。为了减少计算量，于是选取了数据较多的 North Pacific (35N-40N,130W-120W)区域，并结合了**[8]**中，改进后的 Deseasonlization 法，算出平均年增长率约为 4.035 uatm/yr,(145 个月份的数据)。改进后的结果与之前的方法算出结果（3.390 uatm/yr）接近。下一步，按照文献中的处理方法，对所有月份进行校正。

对校正后的数据进行空间相关性分析，发现普遍相关性降低。该结果意味着，

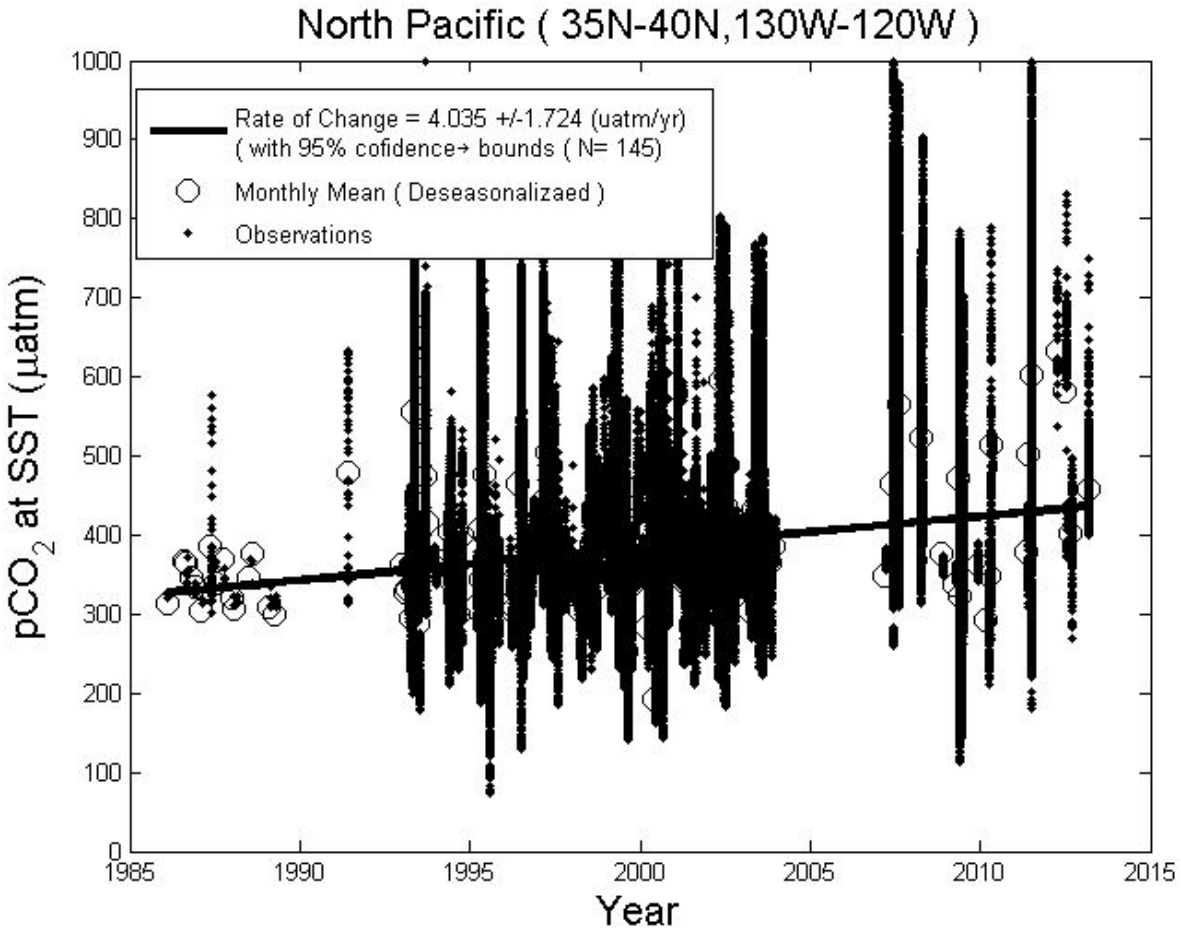


图 5-2

参见[Takahashi et al., 2002], Deseasonlization

这样的校正途径，从时间上整体平均来说可以反映整体的趋势，但对于空间结构，特别是pCO_2的空间结构的分布的影响是显著的。因此，基于以上的研究，本文提出这样的 Deseasonlization 方法还有待进一步的研究和优化，而空间相关性的分析也可以作为其一种有效的检测手段。

致谢

感谢在毕业设计期间相关的老师，同学和朋友的交流支持和鼓励。如，在这期间王桂芝老师，杨爽老师，果果师姐以及相关同学在开题报告准备时的倾听和建议。

谢谢在物理学院的四年中受到诸多老师的耐心指引和悉心照顾，如陈长军老师的实习指导和毕设的院内指导。

最后，特别感谢毕设期间从头至尾，从生活学习，毕业设计工作交流乃至一些非常难得个人思想上，来自戴老师的关心照顾和指引，您的关心和教诲是学生在厦门这数个月来的生活中最难忘的经历和收获。

参考文献

[1] Cressie, N., and N. H. Chan (1989), Spatial Modeling of Regional Variables, *Journal of the American Statistical Association*, *84*(406), 393-401.

[2] Cressie, N. (1992), STATISTICS FOR SPATIAL DATA, *Terra Nova*, *4*(5), 613–617.

[3] Haining, R. (2003), Spatial Data Analysis, *Spatial Data Analysis, by Robert Haining, pp. 452. ISBN 0521773199. Cambridge, UK: Cambridge University Press, June 2003.*, *1*(1), 452.

[4] Legendre, P., N. L. Oden, R. R. Sokal, A. Vaudor, and J. Kim (1990), Approximate analysis of variance of spatially autocorrelated regional data, *Journal of Classification*, *7*(1), 53-75.

[5] Legendre, P. (1993), Spatial autocorrelation: trouble or a new paradigm, *Ecology*, *74*(6), 1659-1673.

[6] Shen, S. S. P., H. Yin, and T. M. Smith (2007), An Estimate of the Sampling Error Variance of the Gridded GHCN Monthly Surface Air Temperature Data, *Journal of Climate*, *20*(10), 2321-2331.

[7] Shen, S. S. P., C. K. Lee, and J. Lawrimore (2012), Uncertainties, Trends, and Hottest and Coldest Years of U.S. Surface Air Temperature since 1895: An Update Based on the USHCN V2 TOB Data, *Journal of Climate*, *25*(12), 4185-4203.

[8] Takahashi, T., et al. (1997), Global air-sea flux of CO2: An estimate based on measurements of sea–air pCO2 difference, *Proceedings of the National Academy of Sciences*, *94*(16), 8292-8299.

[9] Takahashi, T., et al. (2002), Global sea–air CO2 flux based on climatological surface ocean pCO2, and seasonal biological and temperature effects, *Deep Sea Research Part II: Topical Studies in Oceanography*, *49*(9–10), 1601-1622.

[10] Takahashi, T., et al. (2009), Climatological mean and decadal change in surface ocean pCO 2 , and net sea–air CO 2 flux over the global oceans, *Deep Sea Research Part II: Topical Studies in Oceanography*, *56*(11), 554–577.

[11] Taylor, J. R. (1997), An introduction to Error Analysis, *University Science Books*, *56*(4), 471–475.

[12] Wanninkhof, R., G. Park, and T. Takahashi (2012), Global ocean carbon uptake: magnitude, variability and trends, *Biogeosciences Discussions*, *10*(8), 1983-2000.

[13] Wang, G., M. Dai, S. S. P. Shen, Y. Bai, and Y. Xu (2014), Quantifying uncertainty sources in the gridded data of sea surface CO2 partial pressure, *Journal of Geophysical Research: Oceans*, *119*(8), 5181–5189.

[14] Tans, P. P., I. Y. Fung, and T. Takahashi (1990), Observational contrains on the global atmospheric co2 budget, *Science*, *247*(4949), 1431-1438.

[15] Bai, Y., W.-J. Cai, X. He, W. Zhai, D. Pan, M. Dai, and P. Yu (2015), A mechanistic semi-analytical method for remotely sensing sea surface pCO2 in river-dominated coastal oceans: A case study from the East China Sea, *Journal of Geophysical Research: Oceans*, *120*(3), 2331-2349.

[16] Chou, Y. H. H. (1991). Map resolution and spatial autocorrelation, *Geographical Analysis,* 23, 228-46.

附录

1. 独立代码

```
%This code is independent and produced by Lux.
%Email:lixiang19930214@gmail.com
% The input file should be with a standard format. Like:
%latitede,longtitude,date,value
%Value A,Value B, string(normally represents date), Value C

clear all;clc;close all;% First part of the code: Insert Data.
name='RB01_Jun_2001';
filepath=strcat(name,'.txt');
[lat,lon,date,pco2]=textread(filepath,'%f%f%s%f','headerlines',0);

%Random test
% order_r=randperm(6512);
% lon1=lon(order_r(1:4000));
% lat1=lat(order_r(1:4000));
% pco21=pco2(order_r(1:4000));
% lon2=lon(order_r(4001:6512));
% lat2=lat(order_r(4001:6512));
% pco22=pco2(order_r(4001:6512));

%Randomly add up a small displacement for each data's coordinate. In order
%to make autocorrelation matrix has solutions.
num_station=length(pco2);
for i=1:num_station
       lat(i)=lat(i)+0.001*(rand-0.5);
       lon(i)=lon(i)+0.001*(rand-0.5);
end
%Calculate each data pair distance.
for i=1:num_station
    for j=1:num_station
       d(i,j)=((lon(j)-lon(i))^2+((lat(j)-lat(i))^2))^0.5;
    end
end
%Define a resolution for correlation calculation. Normally, each continuous
%data points have a average distance approaching to 0.003. In this case,
we set
%it as 0.02, in order to get a sufficient statistics population. For each
%single band. (But we could enlarge them, and it would make a difference
to our results.)
```

```
%%%
%There is another method to divide its sub-zone. Like 20 gaps.
num_gap=20;
h_resolution=max(max(d)/2)/num_gap;
h_lag=h_resolution;

%This part is to calculate the spatial correlation with respect to the
%distance. And R is the autocorrelation function which has been normalized.
sum_total_i_j=0;
for h=0:num_gap
    h_x(h+1)=h*h_resolution;
    if (h==0)
        [x_d y_d]=find(d==0);
    else
        [x_d y_d]=find(d>=h_resolution*h&d<h_resolution*(h+1));

    end
    num_d_inside(h+1)=length(x_d);
    sum_total_i_j=sum_total_i_j+length(x_d);
    sum_total=0;
    for i=1:length(x_d)
        sum_total=sum_total+pco2(x_d(i))*pco2(y_d(i));
    end
    Covarance_(h+1)=
sum_total/length(x_d)-mean(pco2(x_d))*mean(pco2(y_d));
    R(h+1)=
Covarance_(h+1)/((length(x_d)-1)/length(x_d)*var(pco2(x_d))*...
        (length(x_d)-1)/length(x_d)*var(pco2(y_d)))^0.5;
end

% Curve fit, and estimate the radius of circular.
[a]=polyfit(h_x,R,4);
h_r=[0:0.0001:max(max(d)/3)];
% The reason why I use max(d)/3 is that prevent the fitted curve have two
potiential answer.
R_fit=a(1)*h_r.^4+a(2)*h_r.^3+a(3)*h_r.^2+a(4)*h_r+a(5);
[b]=polyfit(h_x,Covarance_,4);
C_fit=b(1)*h_r.^4+b(2)*h_r.^3+b(3)*h_r.^2+b(4)*h_r+b(5);
% plot(h_r, R_fit,'r');
% hold on;
% scatter(h_x, R);
% title('Autocorrelation');

%This part try to find the coordinate x which R_fit equals to 0.3125*R(0)
to represent Radius.
```

```
while 1

[order]=find(R_fit<(0.3125+h_resolution)*(max(R_fit))&R_fit>(0.3125-h_
resolution)*(max(R_fit)));
    if (length(order)==0)
        h_resolution=2*h_resolution;
    elseif(length(order)>2)
        h_resolution=0.5*h_resolution;
    else
        break;
    end
end
if (length(order)==1)
    radius=h_r(order);
else
    radius=mean(h_r(order));
end

%After we obtaining the radius and we will know the range,nugget effect
%value and drill value.
%Clear the previous data, keep data to calculate range,nugget effect
%value and drill value.
clearvars -except radius R lat lon pco2 d num_station Covarance_ C_fit h_r
h_x
%C=C0+C   a represents range.
min_long=min(lon)-2*radius;max_long=max(lon)+2*radius;min_lat=min(lat)
-2*radius; max_lat=max(lat)+2*radius;
aera=(max_lat-min_lat)*(max_long-min_long);
amount=0;
C=Covarance_(1);
a=2*radius;
C0=C-C_fit(1);
% After we obtaining model parameters, we start to do the Kriging
% Estimations next.
x_c=min(lon);%start point
y_c=min(lat);
x_size=round((max(lon)-min(lon))/(radius/8)); %Kriging Interpolation
resolution equals to radius/8
y_size=round((max(lat)-min(lat))/(radius/8));

for m=1:x_size
    for n=1:y_size
        order_output=n+(m-1)*y_size;
        x_c(order_output)=min(lon)+(m-1/2)*(radius/8);
        y_c(order_output)=min(lat)+(n-1/2)*(radius/8);
```

```
        C_0=0;C_=0;d_0=0;d=0;
        %Point Kriging interpolation.
        for i=1:length(lat)%Calculation estimating point's distances among
other points.
            d_0(i)=
((lon(i)-x_c(order_output))^2+((lat(i)-y_c(order_output))^2))^0.5;
        end
        [d_temp]=find(d_0<=a);%Reserve the data points which are included
inside the range.
        display(length(d_temp));
        Kriging_points=length(d_temp);% A single estimation which are
determined by how many data points.

        %Construc the Autocorrelation Matrix. Ci,j
        if (Kriging_points>0) %If the estimating point have more than one
data in its range.
            d_0=d_0(d_temp);pco2_temp=pco2(d_temp);
            lon_temp=lon(d_temp);lat_temp=lat(d_temp);

            for i=1:Kriging_points

C_0(i,1)=Covarance_(1)*[1.5*(1-d_0(i)/a).^2-0.5*(1-d_0(i)/a).^3]; %Auto
correlation function
                %Spherical model
            end
                C_0(Kriging_points+1,1)=1;

            for i=1:Kriging_points
                for j=1:Kriging_points

d(i,j)=((lon_temp(j)-lon_temp(i))^2+((lat_temp(j)-lat_temp(i))^2))^0.5
;

C_(i,j)=Covarance_(1)*[(d(i,j)/a)-0.5*(1-d(i,j)/a).^3];
                end
            end

            for i=1:Kriging_points
                C_(i,Kriging_points+1)=1;
                C_(Kriging_points+1,i)=1;
            end

            %The detail of constructing the metrix will be show in P88.
            %Solve the equations.
            lamda=inv(C_)*C_0;
```

```
          pco2_e(order_output)=0;

uncertainty_e(order_output)=Covarance_(1)+lamda(Kriging_points+1);

          for i=1:Kriging_points

pco2_e(order_output)=lamda(i)*pco2_temp(i)+pco2_e(order_output);

uncertainty_e(order_output)=uncertainty_e(order_output)-lamda(i)*C_0(i)
;
          end

       else
%           pco2_e(order_output)=-999;
%           uncertainty_e(order_output)=-999;
       end

   end
end
%Draw an research region with data distribution.
order=find(pco2_e>0);
pco2_e=pco2_e(order);
uncertainty_e=abs(uncertainty_e(order));y_c=y_c(order);x_c=x_c(order);
uncertainty=sqrt(uncertainty_e);
m_proj('mercator','lat',[min_long,max_long],'long',[min_lat,max_lat]);
[x y]=m_ll2xy(x_c,y_c,'clip');
scatter(x,y,30,pco2_e,'filled');
m_gshhs_f('patch',[0.7 0.7 0.7],'edgecolor','none');
m_grid('box','fancy','tickdir','in');
colorbar;
title('Uncertainty','fontsize',14);

% test_num_Men=10000;
% %Using mento carlo to estimate the aera of observations.
% for i=1:test_num_Men;
%    x_test=rand*(max_long-min_long)+min_long;
%    y_test=rand*(max_lat-min_lat)+min_lat;
%    for j=1:num_station
%       d_test=(abs((lon(j)-x_test))^2+(abs(lat(j)-y_test))^2)^0.5;
%       if(d_test<=radius)
%          amount=amount+1;
%          break;
%       end
%    end
% end
```

```
%
% eff_obser=(amount/test_num_Men)*aera/(pi*radius^2);
% uncertainty_sample= var(pco2)/eff_obser;
```