污染源的确定

肖力文 U201911484

人工智能与自动化学院

污染源的确定

摘要

本文建立回归分析模型，确定了污染物随距离的扩散关系，并模拟了整个平面受污染源污染一段时间后该污染物的扩散情况，还根据污染源的个数、平面上的受污染状况确定了不同情况下污染源所在位置及其初始浓度。

针对问题一，由实验室得到的数据，建立回归分析模型，用两种解法分别拟合出污染物浓度关于扩散距离的衰减函数的两种可能形式。比较两种表达式各自与样本点间的 MSE 值，并做假设检验，以判定两种函数模型的合理性，从而确定出最合理的衰减函数表达式。之后，模拟出当污染源位于$[0,50]\times[0,80]$的平面中心处，即坐标为（25，40），且初始浓度为 10 的情况下，整个平面最终的污染情况，从而把握污染物浓度随扩散距离变化的基本规律。

针对问题二，根据附件中给出的某时刻在观测点所测得的污染物浓度，将污染源的横、纵坐标以及初始浓度视为参量，同样建立回归分析模型，用最小二乘法求解出这三个参数的问题。拟合出函数模型后，求出拟合模型的均方误差 MSE，并对结果做 t 检验、p 值检验等假设检验，以讨论结果精确度。

针对问题三，由附件中给出的数据，用问题二中所述的方法，将两个污染源的横、纵坐标以及初始浓度均视为参量，建立与问题二相似的回归分析模型，求出这六个参数的最小二乘估计。对所得结果做问题二所述的一系列检验，讨论结果的精确度。

综上所述，本文应用数理统计的知识，建立回归分析模型，通过所给数据拟合出污染物随距离扩散的关系，并求解出在不同情况下污染源的位置坐标以及初始浓度。随后用多种假设检验的方式详尽讨论结果的精确度。

关键词： 回归分析 最小二乘法 假设检验

1 问题的重述

1.1 背景资料和条件

为了确定在一个平面区域$[0,50]\times[0,80]$中某污染源的位置（在平面内），常常通过多个仪器在不同的位置进行测定，对测得的数据加以数学的方法进行计算，可以得到污染源的大致位置，并且还可以计算出污染物在初始位置的浓度值。

1.2 需要解决的问题

假设该污染物（假定是点污染源，中心向外扩散，每个方向的扩散没有差异性）在平面扩散的过程中，浓度会随着扩散距离进行衰减，其衰减函数大致是指数函数关系（关于距离），为了确定这种函数关系，可以在实验室（或相对理想的环境下）进行测定不同位置的浓度。实验数据详见表中所示：

距离	0	10	20	30	40	50	60	70	80	90	100
相对浓度	1	1.22	1.49	1.82	2.23	2.73	3.32	4.05	4.95	6.06	7.41

根据相关信息，建立数学模型，对以下问题进行求解

问题 1 根据实验室得到的数据，模拟得到该污染物的扩散情况。

问题 2 根据附件中给出的某时刻在观测点所测得的污染物浓度，试确定污染源的大致位置及初始浓度，并讨论结果的精确度做分析。

问题 3 如果存在两个污染源，试由附件中给出的数据给出污染源位置及初始浓度，结果精度如何？

2 问题的分析

确定污染源的位置以及初始浓度之前，要先利用实验室数据求出污染物浓度随扩散距离变化的关系式，再根据问题 2 与问题 3 具体的数据，建立回归分析模型去确定污染源的位置以及初始浓度。而求出污染物浓度随扩散距离变化的关系式的过程，也是建立回归模型，再用最小二乘法求解。

2.1 针对问题 1 的分析

目标是根据样本点拟合出最恰当的指数型函数模型，使得该函数在最小二乘准则下是最优的。求解回归模型有不同的方法：第一种是化非线性回归模型为线性回归模型，然后按线性回归求解；第二种是按最小二乘优化的思想，直接求出函数模型中参数的最小二乘估计。由于不同方法求解出的结果不同，因此有必要

对不同模型的合理性进行分析，并将最优的模型确定为染物浓度随扩散距离变化的关系式。

2.2 针对问题 2 的分析

由问题 1 中求解得到的关系式，在问题 2 的条件下可以得到一个含参的函数式，参数为污染源的横、纵坐标以及其初始浓度。再由附件中给出的样本点的数据，去拟合出最小二乘准则下最优的函数，从而解出所求的参数。由于题目要求分析结果的精确度，因此计算得到的模型与样本点的 MSE 值，并做 t 检验、p 值检验两种假设检验，从而有效分析结果的精确度。

2.3 针对问题 3 的分析

问题 3 与问题 2 类似，不同点在于增加了一个污染源，增加的污染源带来了更多的参数。除参数变多以外，没有其他变化。因此可以考虑与问题 2 完全相同的思路进行求解，分析结果精度的方式也同问题 2。

3 模型假设

（1）问题 1 的表格数据中，相对浓度 $=\frac{\text{标准浓度}}{\text{绝对浓度}}$，且问题 1 中的标准浓度为 c_s。

（2）当平面上有两个点污染源时，平面上任一点处的污染物浓度是两个点污染源对该点影响的简单加和，不会产生附加反应以增强或减弱污染。

4 符号与说明

符号	说明
c_s	问题 1 中的标准浓度
c_0	问题 2 中污染源的初始浓度
c_1	问题 3 中其中一个污染源（污染源 1）的初始浓度

c_2	问题 2 中另一个污染源（污染源 2）的初始浓度
(x,y)	任一点位置
(x_0,y_0)	问题 2 中污染源坐标
(x_1,y_1)	问题 3 中污染源 1 的位置
(x_2,y_2)	问题 3 中污染源 2 的位置
d	平面上任一点到单一污染源的距离
d_1	平面上任一点到问题 3 中污染源 1 的距离
d_2	平面上任一点到问题 3 中污染源 2 的距离

5 模型的建立与求解

5.1 问题 1 的模型建立与求解

5.1.1 模型的准备

根据实验数据，求出扩散距离与绝对浓度间的关系，如表 1：

表 1. 扩散距离与绝对浓度间的关系

距离	0	10	20	30	40	50	60	70	80	90	100
绝对浓度	c_s	0.820 c_s	0.671 c_s	0.549 c_s	0.448 c_s	0.366 c_s	0.301 c_s	0.246 c_s	0.202 c_s	0.165 c_s	0.145 c_s

5.1.2 模型的建立

由于已知衰减函数是关于距离的指数关系，可设函数表达式为 $c = a_1 e^{a_2 d}$。建

立回归分析模型，求解出回归系数a_1、a_2。

先用问题分析中所述的第一种方法求解:等号两边取对数得$\ln c = \ln a_1 + a_2 d$，将非线性回归模型转化为线性回归模型。则由 Matlab 程序求解线性模型的方法，利用 *polyfit* 函数解出衰减函数的表达式为

$$c = 0.9829 c_s e^{-0.0197 d} \qquad (1)$$

再用问题分析中所述第二种方法，直接用 lsqcurvefit 函数，求出a_1、a_2的最小二乘估计，得到的表达式为

$$c = 1.0008 c_s e^{-0.0202 d} \qquad (2)$$

5.1.3 模型的求解

绘制两种结果的拟合效果图，如图 1 所示。求出它们的 MSE 值，并将 t 检验与 p 值求出，如表 2。

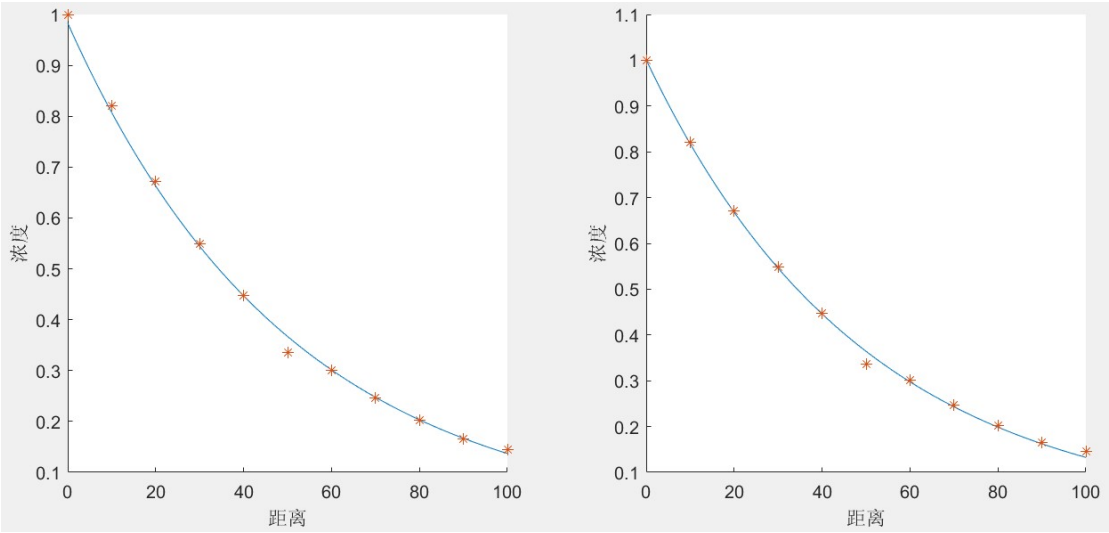

图 1. 函数$c = 0.9829 c_s e^{-0.0197 d}$（左）与函数$c = 1.0008 c_s e^{-0.0202 d}$（右）对样本点的拟合效果图

表 2. 两函数对样本点的 MSE 值、t 检验结果以及 p 值

	$c = 0.9829 c_s e^{-0.0197 d}$	$c = 1.0008 c_s e^{-0.0202 d}$
MSE	1.4363×10^{-4}	0.9290×10^{-4}
t 检验结果	1%的显著水平下显著，即有 99%的把握认为这个模型是正确的	1%的显著水平下显著，即有99%的把握认为这个模型是正确的

p 值	2.3163×10^{-7}	1.9951×10^{-7}

假设检验的结果中，表达式（2）的 p 值略微小于表达式（1），但二者的 p 值都已经足够小，几乎趋于 0，且只在10^{-7}数量级以后有差异。而表达式（2）的 MSE 值小于表达式（1），并且这种差异在10^{-5}数量级上就比较明显。因此以 MSE 值为主要参考因素，选择$c=1.0008c_s e^{-0.0202d}$为衰减函数的表达式更好。故最终确定衰减函数的表达式为$c=1.0008c_s e^{-0.0202d}$。

接下来根据表达式$c=1.0008c_s e^{-0.0202d}$，模拟出当污染源位于$[0,50]\times[0,80]$的平面中心处，即坐标为（25，40），且初始浓度为 10 的情况下，整个平面最终的污染情况。模拟结果如图 2。

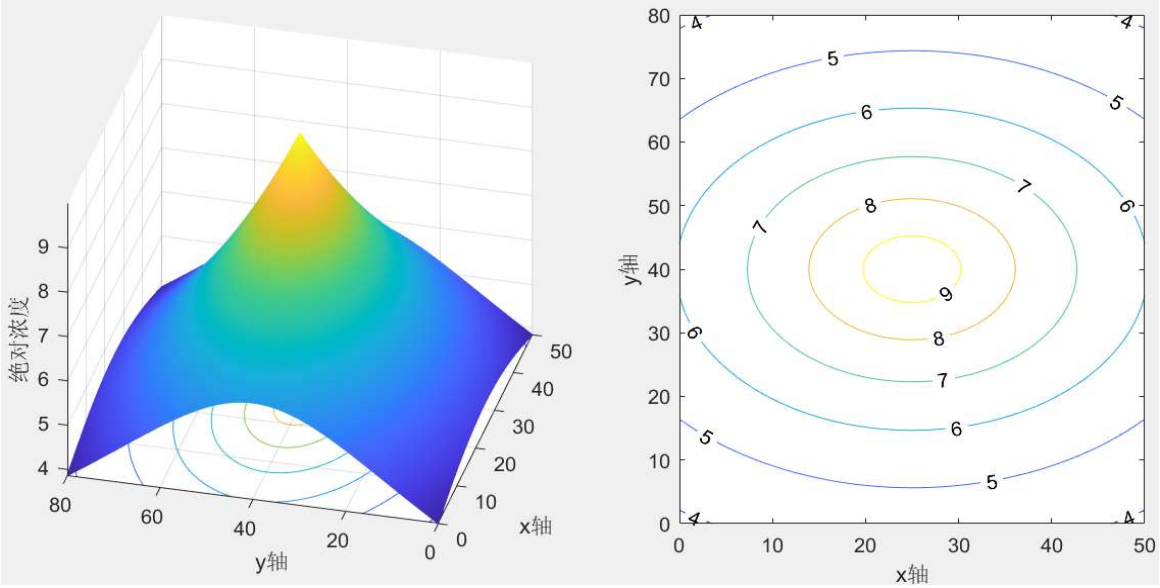

图 2. 污染物扩散情况模拟效果图

5.2 问题 2 的模型建立与求解

5.2.1 模型的分析

由问题 1 的求解知，该时刻各点的绝对浓度c与的距离d及标准浓度c_s有关系式（2）。而这里标准浓度就是污染源的初始浓度，距离是各点与污染源在坐标平面上的欧式距离，即：

$$c_s = c_0 \qquad (3)$$

$$d=\sqrt{(x-x_0)^2+\left(y-y_0\right)^2} \tag{4}$$

则各点的绝对浓度与其坐标的关系是：

$$c=1.0008c_0e^{-0.0202\sqrt{(x-x_0)^2+(y-y_0)^2}} \tag{5}$$

现在目的是根据已知表达式(5)以及平面上 10 个点的信息，求解出参数c_0、x_0、y_0，使十个数据点与函数（5）在最小二乘准则下拟合得最好。则问题 2 转化为，建立回归分析模型，用最小二乘法求解模型。

5.2.2 模型的建立

我们已知函数表达式与十个点的坐标与绝对浓度，建立了函数模型（5），即$c=1.0008c_0e^{-0.0202\sqrt{(x-x_0)^2+(y-y_0)^2}}$，之后通过 lsqcurvefit 函数用最小二乘法求出参量$\hat{c}_0$、$\hat{x}_0$、$\hat{y}_0$。

5.2.3 模型的求解

用 lsqcurvefit 求解出的$\hat{c}_0$、$\hat{x}_0$、$\hat{y}_0$值为

$$\begin{cases}\hat{c}_0=5.6139\\ \hat{x}_0=19.2981\\ \hat{y}_0=38.6938\end{cases},$$

将得到的$\hat{c}_0$、$\hat{x}_0$、$\hat{y}_0$代回式（5），得到

$$c=5.6184e^{-0.0202\sqrt{(x-19.2981)^2+(y-38.6938)^2}} \tag{6}$$

图 3 给出函数式（6）的图像以及浓度等高线示意图：

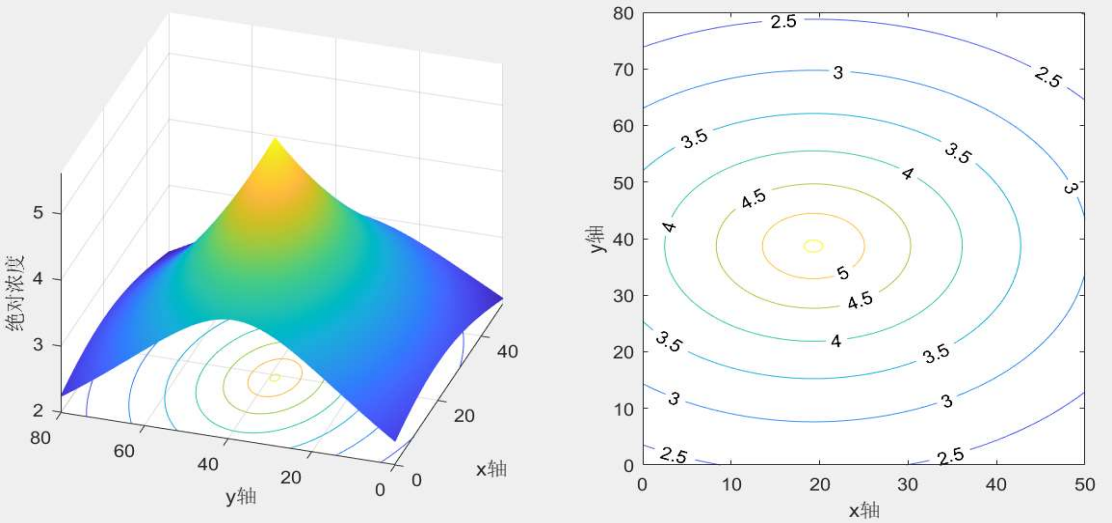

图 3. 函数表达式（6）的图像以及浓度等高线示意图

算出式（6）与十个数据点的 MSE 值，并对所得模型进行假设检验（t 检验的原假设为：假定数据测量的真实值与模型的预测值来自均值等于零且方差未知的正态分布，即认为二者相等），结果如表 3：

表 3. 函数（6）对样本点的 MSE 值、t 检验结果以及 p 值

MSE 值	0.0267
t 检验结果	0.01 的显著水平下不显著
p 值	0.9686

可以看到，该表达式求出的 MSE 值很小，说明函数预测值与真实值契合得很好；而 t 检验结果在α=0.01 的显著水平下接受原假设，则说明有 99%的把握认为数据测量的真实值与模型的预测值是相等；p 值为 0.9686 > α，说明如果拒绝“数据测量的真实值与模型的预测值相等”这一假设，犯第一类错误的概率为 96.86%，因此不应该拒绝该假设，而应接受原假设。综上，我们有充分的理由认为，该表达式的预测值与数据测量的真实值相等，即该表达式几乎符合真实情况。也就是说，求解得到的污染源的位置以及其初始浓度的精度很高。

5.3 问题 3 的模型建立与求解

5.3.1 模型的分析

由式（2）与假设知，任一点(x,y)在该平面内，其污染物浓度与两个污染源之间的关系是

$$c = 1.0008\left(c_1 e^{-0.0202\sqrt{(x-x_1)^2+(y-y_1)^2}} + c_2 e^{-0.0202\sqrt{(x-x_2)^2+(y-y_2)^2}} \right) \quad (7)$$

用问题 2 所述方法，求出c_1、x_1、y_1、c_2、x_2、y_2六个参量后，代入式（7），再求出 MSE 值，做假设检验，以讨论其精度。

5.3.2 模型的建立

以附件中浓度最大的两点的信息为拟合参数的初始值，用最小二乘法求出参数值，得到参数的结果为：

$$\begin{cases} \hat{c}_1 = 12.3581 \\ \hat{x}_1 = 40.9789 \\ \hat{y}_1 = 19.9756 \\ \hat{c}_2 = 8.3923 \\ \hat{x}_2 = 11.0830 \\ \hat{y}_2 = 70.0370 \end{cases},$$

代回（7）式得到

$$c = 1.0008\left(12.3581 e^{-0.0202\sqrt{(x-40.9789)^2+(y-19.9756)^2}} + 8.3923 e^{-0.0202\sqrt{(x-11.0830)^2+(y-70.0370)^2}} \right) \quad (8)$$

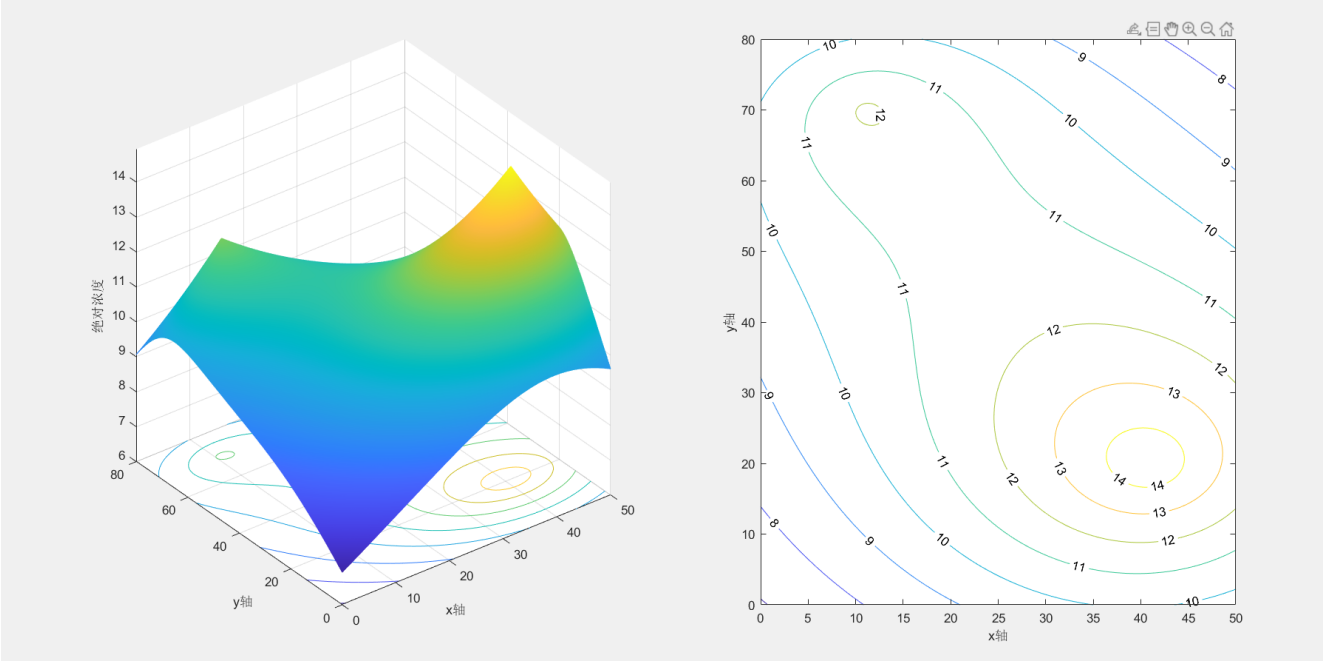

图 4. 函数表达式（8）的图像以及浓度等高线示意图

算出式（8）与十个数据点的 MSE 值，并对所得模型进行假设检验（t 检验的原假设是：假定数据测量的真实值与模型的预测值来自均值等于零且方差未知的正

态分布，即认为二者相等），结果如表 4：

表 4. 函数（8）对样本点的 MSE 值、t 检验结果以及 p 值

MSE 值	6.4104×10^{-5}
t 检验结果	0.01 的显著水平下不显著
p 值	0.9985

可以看到，该表达式求出的 MSE 几乎趋于 0，说明函数预测值与真实值契合得非常好；而 t 检验结果在α=0.01 的显著水平下接受原假设，则说明有 99%的把握认为数据测量的真实值与模型的预测值是相等；p 值为 0.9985 > α，说明如果拒绝“数据测量的真实值与模型的预测值相等”这一假设，犯第一类错误概率为 99.85%，因此不应拒绝该假设，而应接受原假设。综上，我们有充分的理由认为，该表达式的预测值与数据测量的真实值相等，即该表达式几乎符合真实情况。也就是说，求解得到的污染源的位置以及其初始浓度的精度很高。

6 模型的评价

模型基于回归分析理论与最小二乘法，根据现有数据确定了污染物随距离的扩散关系，模拟了整个平面受污染源污染一段时间后该污染物的扩散情况，还根据污染源的个数、平面上的受污染状况确定了不同情况下污染源所在位置及其初始浓度。通过比较 MSE 值、t 检验和 p 值检验的方式，分析了所有结果的精确性，证明了求解出的结果有极高的精度，说明模型的建立很合理，有效地达到了预期目的。

7 模型的改进

在结果精确度的检验方面，还可以使用更多的方法进行检验，以加强可信度。

附录

1 Matlab 代码索引

问题 1

m1_1.m

m1_2.m

问题 2

m2.m

问题 3

m3.m

2 Matlab 代码

m1_1.m

```
x1=[0 10 20 30 40 50 60 70 80 90 100];
y1=[1 0.820 0.671 0.549 0.448 0.336 0.301 0.246 0.202 0.165 0.145];
y1=log(y1);                                    %转化为线性回归的方式进行求解
p=polyfit(x1,y1,1);
a2=p(1);
a1=exp(p(2));
yhat=a1.*exp(a2.*x1);
mse=sum((exp(y1)-yhat).^2)./length(y1)         %求MSE值
[h p]=ttest(y1,yhat,'Alpha',0.05)              %t检验并求p值
hold on
subplot(1,2,1);
t1=0:100;
yhat=a1.*exp(a2.*t1);
hold on
```

```
plot(t1,yhat);
plot(x1,exp(y1),'*');
xlabel('x轴');
ylabel('y轴');

canshu0=[a1 a2];
f=@(canshu0,x1)canshu0(1).*exp(canshu0(2).*x1);  %直接在最小二乘准则下求解
canshu=lsqcurvefit(f,canshu0,x1,exp(y1));
yhat=canshu(1).*exp(canshu(2).*x1);

mse=sum((exp(y1)-yhat).^2)./length(y1)
[h p]=ttest(y1,yhat,'Alpha',0.01)
t1=0:100;
yhat=canshu(1).*exp(canshu(2).*t1);
subplot(1,2,2);
hold on
plot(t1,yhat);
plot(x1,exp(y1),'*');
xlabel('x轴');
ylabel('y轴');
```

m1_2.m

```
%模拟扩散情况
tx=linspace(0,50,1000);
ty=linspace(0,80,1000);
ty=ty';
tz=1.0008.*10.*exp(-0.0202.*sqrt( (tx-25).^2+(ty-40).^2 ));

hold on
subplot(1,2,1);
meshc(tx,ty,tz);
xlabel('x轴');
ylabel('y轴');
zlabel('绝对浓度');
%等高线图
subplot(1,2,2)
[C h]=contour(tx,ty,tz);
clabel(C,h);
xlabel('x轴');
ylabel('y轴');
hold off
```

m2.m

```
loc=xlsread('data1');
x2=loc(:,2);
y2=loc(:,3);
z2=loc(:,4);
xydata=[x2,y2];
canshu0=[4.5590 4.84 36.3];
f=@(canshu0,xydata)1.0008.*canshu0(1).*exp(-0.0202.*sqrt( (xydata(:,1)-
   canshu0(2)).^2+(xydata(:,2)-canshu0(3)).^2 ));
lb=zeros(1,3);
ub=[inf 50 80];
canshu=lsqcurvefit(f,canshu0,xydata,z2,lb,ub)
zhat=1.0008.*canshu(1).*exp(-0.0202.*sqrt( (x2-canshu(2)).^2+(y2-
      canshu(3)).^2 ));
mse=sum((zhat-z2).^2)./length(z2)          %mse
[h p]=ttest2(z2,zhat,'Alpha',0.01)

tx=linspace(0,50,1000);
ty=linspace(0,80,1000);
ty=ty';
tz=1.0008.*canshu(1).*exp(-0.0202.*sqrt( (tx-canshu(2)).^2+(ty-
canshu(3)).^2 ));

hold on
subplot(1,2,1);
meshc(tx,ty,tz);
xlabel('x轴');
ylabel('y轴');
zlabel('绝对浓度');

subplot(1,2,2)
[C h]=contour(tx,ty,tz);
clabel(C,h);
xlabel('x轴');
ylabel('y轴');
```

m3.m

```
loc=xlsread('data2');
x3=loc(:,2);
y3=loc(:,3);
```

```
z3=loc(:,4);
xydata=[x3,y3];
canshu0=[11.5348 21.14  29.51 11.9530 27.39 36.86];
f=@(canshu0,xydata)1.0008.*(canshu0(1).*exp(-0.0202.*sqrt( (xydata(:,1)-
    canshu0(2)).^2+(xydata(:,2)-canshu0(3)).^2 ))+canshu0(4).*exp(-
    0.0202.*sqrt( (xydata(:,1)-canshu0(5)).^2+(xydata(:,2)-
    canshu0(6)).^2 )));
lb=zeros(1,6);
ub=[inf 50 80 inf 50 80];
canshu=lsqcurvefit(f,canshu0,xydata,z3,lb,ub)
zhat=1.0008.*(canshu(1).*exp(-0.0202.*sqrt( (xydata(:,1)-
     canshu(2)).^2+(xydata(:,2)-canshu(3)).^2 ))+canshu(4).*exp(-
     0.0202.*sqrt( (xydata(:,1)-canshu(5)).^2+(xydata(:,2)-
     canshu(6)).^2 )));
mse=sum((zhat-z3).^2)./length(z3)          %mse
[h p]=ttest2(z3,zhat,'Alpha',0.01)
tx=linspace(0,50,1000);
ty=linspace(0,80,1000);
ty=ty';
tz=0.9829.*canshu(1).*exp(-0.0197.*sqrt( (tx-canshu(2)).^2+(ty-
canshu(3)).^2 ));
hold off
hold on
subplot(1,2,1);
meshc(tx,ty,tz);
xlabel('x轴');
ylabel('y轴');
zlabel('绝对浓度');
subplot(1,2,2)
[C h]=contour(tx,ty,tz);
clabel(C,h);
xlabel('x轴');
ylabel('y轴');
```