颜色与物质浓度辨识

摘要

针对 B、G、R 与浓度之间的关系，本文在 Kubelka-Munk 理论的基础上，应用数学建模的方法对浓度与颜色的关系进行研究，通过对实验样本数据的选取与分析，建立了在相同物质浓度下 B、G、R 与物质浓度的曲线相关性分析模型和回归模型，并对该模型进行回归分析，分析结果表明，实验数据的误差绝大部分在 2%以内，说明该回归方程的拟合效果比较好，与 B、G、R 值相对应的所求解的物质浓度比实验数据的误差小，该研究精度高，拟合效果好，能够通过物质的颜色来判断物质浓度的需要，具有一定的实用价值。

关键字： 回归分析　相关性分析　浓度　B　G　R

一 问题重述

比色法是目前常用的一种检测物质浓度的方法，即把待测物质制备成溶液后滴在特定的白色试纸表面，等其充分反应以后获得一张有颜色的试纸，再把该颜色试纸与一个标准比色卡进行对比，就可以确定待测物质的浓度档位了。随着照相技术和颜色分辨率的提高，通过建立颜色读数和物质浓度的数量关系，即只要输入照片中的颜色读数就能够获得待测物质的浓度。附件　问题一中分别给出了 5 种物质在不同浓度下的颜色读数，讨论从这 5 组数据中能否确定颜色读数和物质浓度之间的关系，并给出一些准则来评价这 5 组数据的优劣

问题二通过对附件 Data2.xls 中的数据，建立颜色读数和物质浓度的数学模型，并给出模型的误差分析。

问题三探讨数据量和颜色维度对模型的影响。

二 问题假设

1.假设文件给出的数据和实际误差不是太大；

2.假设 R，G，B，S，H 和浓度的关系是相互独立；

3.假设物质的浓度没有达到饱和状态，没有析出晶体；

4.假设每种物质的纯净的，没有其他杂质，浓度的读数是完全正确。

三 符号说明

符号	说明

R	红色读数
G	绿色读数
B	绿色读数
H	色调
S	饱和度
ppm	物质浓度
p_s	相关系数

四 模型建立

4.1 问题一模型

4.1.1 模型建立

相关性分析[1]是指对两个或多个具备相关性的变量元素进行分析，从而衡量两个变量因素的相关密切程度。相关性的元素之间需要存在一定的联系或者概率才可以进行相关性分析。

关于浓度与颜色读数关系中，颜色的维数比较多，下面力图通过相关性分析， 发现每个维度与浓度的相关关系。在实际应用中可以选择不同类别中的评价指标。系数法、 Pearson 相关系数法和 Spearman 相关系数法是计算变量间相关性的常用方法。系数法用于计算定性变量之间的相关系数。Pearson 系数法衡量两变量间的线性相关性. 由于对数据分布敏感， 只有当数据服从或近似服从正态分布的时候，该方法才是可靠的， 否则可能产生错误的结论。Spearman 相关系数是一个非参数性质的秩统计参数，反映两变量之间联系的强弱程度，与数据分布形态无关. 本文选择 Spearman 相关系数法对图像融合质量评价指标进行相关性分析.设参与相关性分析的两个变量 X 和 Y 长度均为 Y_0。X 和 Y 均按降序排列后分别记为 X_{sorted} 和 Y_{sorted}。 X' 和 Y' 内分别记录 X 和 Y 中元素在 X_{sorted} 和 Y_{sorted} 中的位置， 并称其为秩次。记 $d_i = X'(i) - Y'(i)$，则 Spearman 相关系数 p_s 表示为 $p_S = 1 - \frac{\sum_{i=1}^{N} d_i^2}{N(N^2-1)}$。Spearman 相关系数的取值范围为 [-1，1]。当 $p_s(X,Y) = 1$

时，表示 X 与 Y 正相关，意味着 X 与 Y 的秩次完全相同；当 $p_s(X,Y)=-1$ 时，X 与 Y 负相关，意味着 X 与 Y 的秩次完全相反；当 $p_s(X,Y)=0$ 时，X 与 Y 不相关，意味着随着 X 的递增 (递减)，Y 没有增大和减小的趋势。

为了说明颜色读数与物质浓度之间的关系，可以先运用相关性分析，简略的了解在不同物质中，B、R、G 分别与相应物质浓度之间的大体关系。如下，表 1 是运用 SPSS[2]中的相关分析得出的组胺、溴酸钾、工业碱、硫酸铝钾、奶中尿素中 B、R、G 与物质浓度的相关系数。

表 1 不同物质中 B、R、G 与物质浓度的相关性

物质			B	G	R
组胺	浓度（ppm）	皮尔森(Pearson) 相关	-0.976**	-0.998**	-0.936*
		显著性 （双尾）	0.004	0.000	0.019
		N	5	5	5
溴酸钾	浓度（ppm）	皮尔森(Pearson) 相关	-0.956*	-0.872	-0.167
		显著性 （双尾）	0.011	0.054	0.788
		N	5	5	5
工业碱	浓度（ppm）	皮尔森(Pearson) 相关	-0.491	-0.664	-0.624
		显著性 （双尾）	0.264	0.104	0.134
		N	7	7	7
硫酸铝钾	浓度（ppm）	皮尔森(Pearson) 相关	0.623	-0.696	-0.648
		显著性 （双尾）	0.187	0.125	0.164
		N	6	6	6
奶中尿素	浓度（ppm）	皮尔森(Pearson) 相关	-0.960**	-0.345	-0.437
		显著性 （双尾）	0.002	0.503	0.386
		N	6	6	6

相关性在 0.01 层上显著（双尾）。**

相关性在 0.05 层上显著（双尾）。*

从表 1 可以看出，在组胺中，B 与物质浓度的 pearson 相关系数为-0.976，右上角标示“**”，显著性小于 0.01，说明在 0.01 的显著性水平上极显著，说明 B 与物质浓度呈显著负相关，即物质浓度的升高，B 随之降低。同理，可以看出 G、R 与组胺的物质浓度都呈显著负相关。在溴酸钾、奶中尿素中，除了 B 与物质的浓度呈显著负相关外，G、R 与物质浓度都呈负相关，而在工业碱、硫酸铝钾中，B、G、R 与物质的浓度呈现负相关的关系。综上所述，B、G、R 与物质浓度呈负相关的关系，即随着物质浓度的增加，颜色读数 B、G、R 随之降低。

为了更好的展现颜色读数与物质浓度之间具体的关系，可以通过回归分析[3]来拟合数据的模型，进而拟合出在不同物质中 B、G、R 分别与相应物质浓度的函数式。

回归分析是确定 2 种或 2 种以上变数间相互依赖的定量关系的一种统计分析方法。它既可以提供变量之间相关关系的数学表达式（经验公式），又可以利用概率统计的基础知识对该表达式进行分析，并判断所得到的数学表达式是否有效。然后利用所得的数学表达式，根据一个或几个变量值预测或控制另一个变量值，并可知道这种预测和控制所能达到的精确程度。

回归模型的一般形式，假设因变量 y 与自变量 $x_1, x_2, \cdots, x_p$ 有联系，那么可以认为因变量 y 由 2 部分组成，一部分由 $x_1, x_2, \cdots, x_p$ 决定，另一部分由众多未考虑的因素（随机因素）决定，则回归模型的一般形式为 $y = f(x_1, x_2, \cdots, x_p) + \varepsilon$ 式中，ε 为随机误差项；$y = f(x_1, x_2, \cdots, x_p)$ 为 y 对 $x_1, x_2, \cdots, x_p$ 均值回归函数。

曲线回归分析回归分析按照是否线性分为线性和非线性回归。变量间的关系虽不是线性关系，但可以通过自变量或因变量的函数变换转化为线性关系，然后应用线性回归方法求参数，并进行回归诊断，再经过适当的变换用原变量写出曲线回归方程进行预测与控制，此方法称为线性化方法。可化为线性回归方法的前提是根据实际样本数据找出适宜的曲线回归模型。本文针对颜色和浓度可以绘制散点图，当散点图显示某种曲线时，则构造相应的曲线回归方程。本实验基于 SPSS 软件进行，SPSS 软件给出了 11 种常见的可化为线性回归的曲线回归方程，并可直接选择曲线模型进行参数估计和回归诊断，其基本原理是线性化方法。

在进行回归分析之前，可以通过分别绘制 B、G、R 与物质浓度关系的散点图来大概了解 B、G、R 与物质浓度的大致关系图。如图 1 是溴酸钾中，B、G、

R 与物质浓度的散点图。

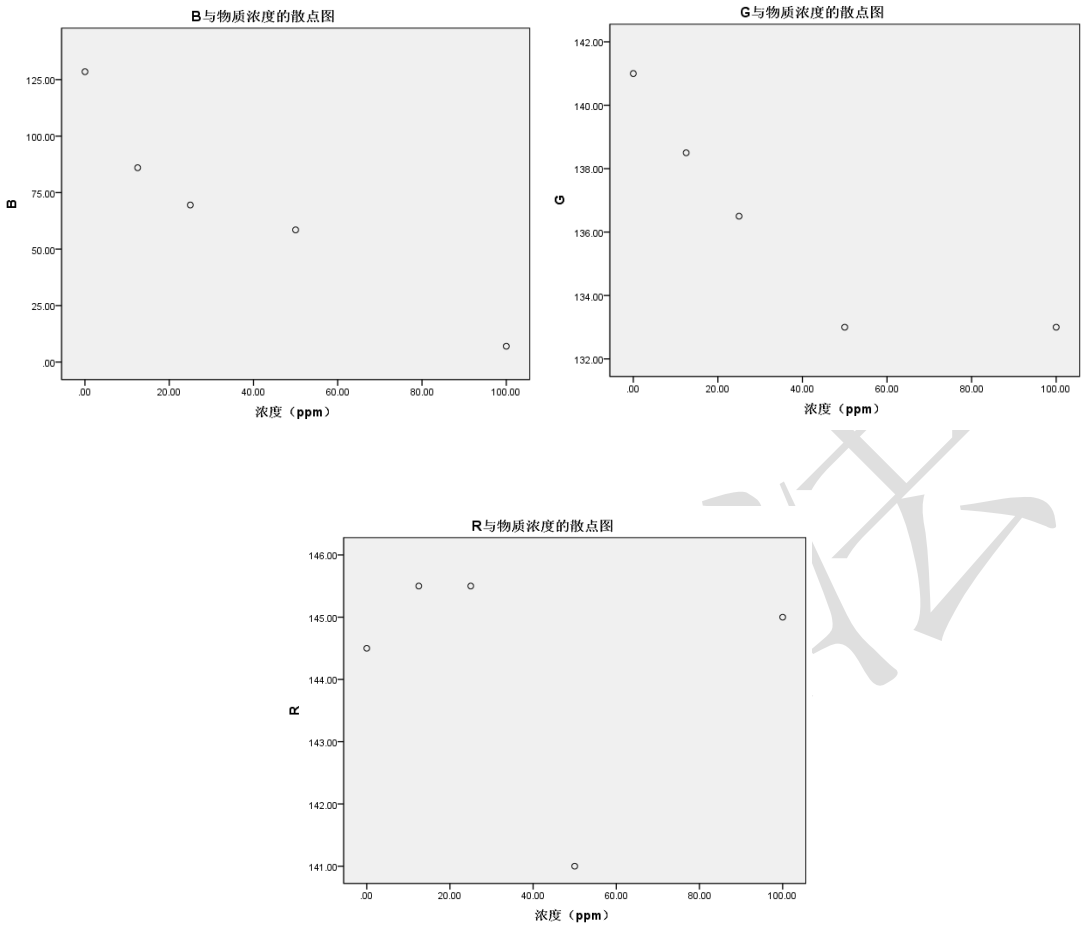


图 1　B、G、R 与物质浓度的散点图

由图 1 可知，在相同物质浓度的物质中，B、G、R 随着物质浓度的增加而降低，并且颜色读数 B、G、R 与物质浓度是某种曲线关系，从而可以运用曲线回归对模型进行求解。

对于物质浓度相同的物质，将颜色读数 B、G、R 与物质浓度 ppm 对应的数据在 SPSS 软件中进行曲线回归，由此可以得到在相同物质浓度的物质中，颜色读数 B、G、R 值与质量浓度 ppm 之间的关系模型为

$$\begin{cases} B = f_1(ppm) \\ G = f_2(ppm) \\ R = f_3(ppm) \end{cases}$$

以溴酸钾中 B 与物质浓度的关系为例分别对其作线性、二次、三次、复合、logistic 多项式拟合，得到如表 2 和图 2 的拟合结果。

表 2　B 与物质浓度拟合结果

模型摘要				
模型	R	R 平方	调整后 R 平方	标准偏差度错误
线性	0.956	0.915	0.886	14.879
二次	0.971	0.942	0.884	14.995
三次	0.999	0.999	0.995	3.094
复合	0.967	0.934	0.913	0.337
logistic	0.967	0.934	0.913	0.337

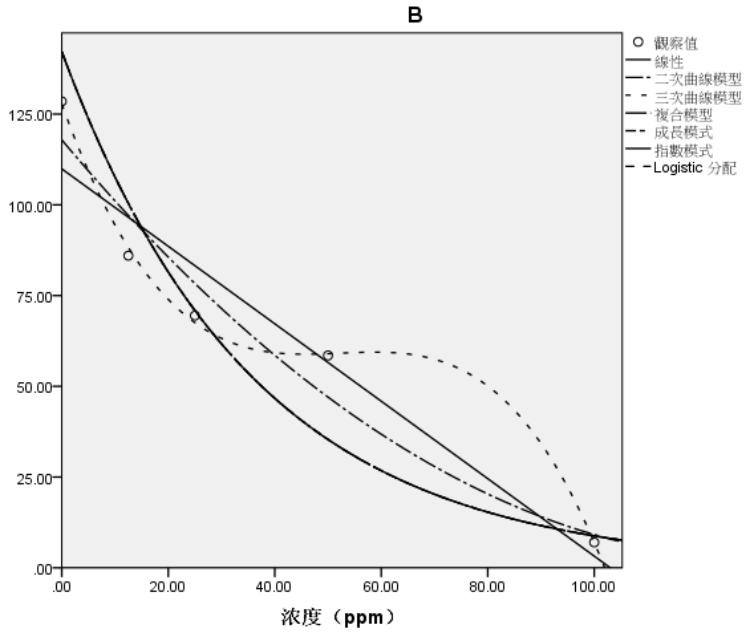


图 2　B 与物质浓度的拟合结果曲线

从表 2 以及图 2，可以看出各个模型的拟合优度，可以看出立方的 R 方最高。同时，从模型的显著性检验结果看，显著性小于 0.05，说明模型成立的统计学意义显著。从表 3 可以看出立方模型的回归系数检验都小于 0.5，说明立方模型的回归系数显著，可知最优模型立方模型的方程式为：B=127.768-4.075*ppm+0.079*ppm^2-0.001*ppm^3。

表 3 立方模型的参数检验

	非标准化系数		标准化系数	T	显著性
	B	标准错误	Beta		

浓度（ppm）	-4.075	0.374	-3.653	-10.889	0.058
浓度（ppm） ** 2	0.079	0.011	7.597	7.334	0.086
浓度（ppm） ** 3	-0.001	0.000	-4.988	-6.780	0.093
（常數）	127.768	3.006		42.500	0.015

同理，在溴酸钾中，

G=140.968-0.194*ppm+8.756E-6*ppm^3

R=144.389+0.222*ppm-0.009*ppm^2+7.206E-5*ppm^3

组胺

B=66.381+0.103*ppm-0.016*ppm^2+0.000122*ppm^3

G=109.453-0.535*ppm+0.003*ppm^2-1.931E-5*ppm^3

R=120.065-0.075*ppm+0.001*ppm^2-1.796E-5*ppm^3

工业碱

B=151.947-23.080*ppm+5.927*ppm^2-0.373*ppm^3

G=141.738+72.331*ppm+13.436*ppm^2+0.532*ppm^3

R=132.011-3.192*ppm+1.206*ppm^2-0.104*ppm^3

硫酸铝钾

B=116.482+45.325*ppm-15.444*ppm^2+1.576*ppm^3

G=123.475-10.174*ppm+1.468*ppm^2

R=100.257-79.706*ppm+32.703*ppm^2-3.842*ppm^3

但是，在奶中尿素中 B 与物质浓度的模型拟合能得到较好的拟合效果，即

B=120.026-0.014*ppm+3.464E-6*ppm^2

而 G、R 与物质浓度的模型拟合效果不太理想。如表 4，为 G、R 分别与物质浓度的模型拟合结果。

表 4　G、R 分别与物质浓度的模型拟合结果

		模型摘要			
颜色读数	模型	R	R 平方	调整后 R 平方	标准偏差度错误
G	线性	0.345	0.119	-0.101	2.141
	二次	0.683	0.467	0.112	1.923
	三次	0.698	0.488	-0.281	2.309
	复合	0.343	0.118	-0.103	0.016
	logistic	0.343	0.118	-0.103	0.016

	线性	0.437	0.191	-0.011	1.936
	二次	0.710	0.504	0.174	1.750
R	三次	0.712	0.507	-0.233	2.138
	复合	0.434	0.188	-0.015	0.014
	logistic	0.434	0.188	-0.015	0.014

由表 4 可知，即使是选择 R 方最大的拟合模型，拟合效果也是很不理想的。

评价准则：1）样本数量的多少；

2）选择合适的物质浓度梯度；

3）对于相同物质的相同物质浓度应尽可能多的重复测量；

4）对于测量的相同物质相同的物质浓度时，人为剔除与测量结果偏差较大的测量数据。

4.1.2 数据优劣评价

对于组胺、溴酸钾、工业碱三组数据集，样本数量相对较小，并且对相同物质的相同物质浓度进行重复测量时，测量次数较少，但是工业碱相对于组胺和溴酸钾的物质浓度梯度较合适；对于，奶中尿素的物质浓度梯度过大；而硫酸铝钾相对于其余四组数据是最好的，样本数量相对较多，梯度较合适，对于相同物质的相同物质浓度进行了多次重复测量。

4.2 问题二模型

4.2.1 模型建立

首先对数据集 Data2.xls 进行预处理，对于物质浓度相同时的重复测量结果人为剔除测量结果偏差较大的测量数据集，接着对于相同物质浓度的测量数据取平均值进而得到最终的数据集。见附件 Data.xls。

在进行模型拟合之前，先通过颜色读数 B、G、R 与物质浓度的散点图来观察他们之前的关系。如图 3 所示：

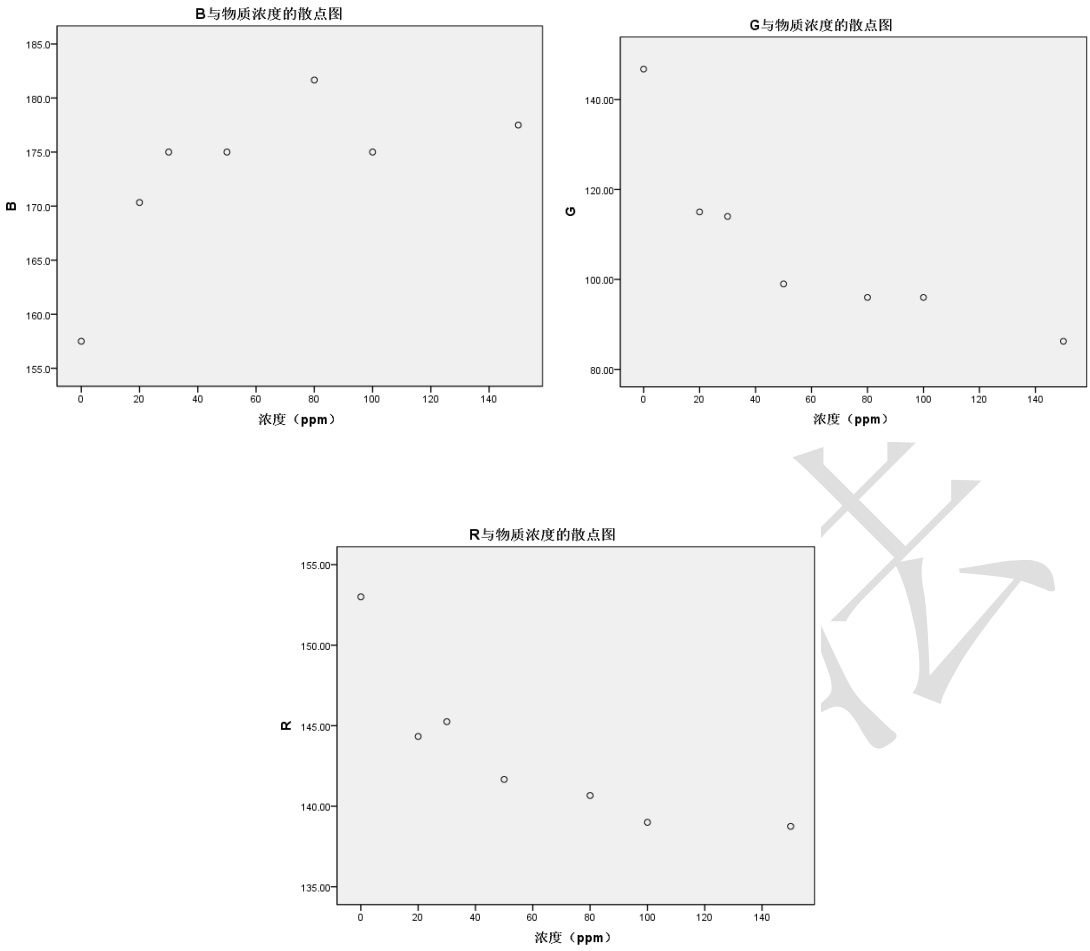


图 3　B、G、R 与物质浓度的散点图

由上图可知，B、G、R 分别与物质浓度具有一定的曲线关系，因而采用回归分析中的曲线回归进行模型的拟合。

以 B 与物质浓度的关系为例分别对其作线性、二次、三次、复合、logistic 等多项式拟合，得到如表 5 和图 4 的拟合结果。

表 5　B 与物质浓度拟合结果

模型摘要

模型	R	R 平方	调整后 R 平方	标准偏差度错误
线性	.671	.451	.341	6.240
二次	.893	.797	.696	4.240
三次	.964	.930	.860	2.880
复合	.669	.447	.337	.037

成长	.669	.447	.337	.037
指数	.669	.447	.337	.037
logistic	.669	.447	.337	.037

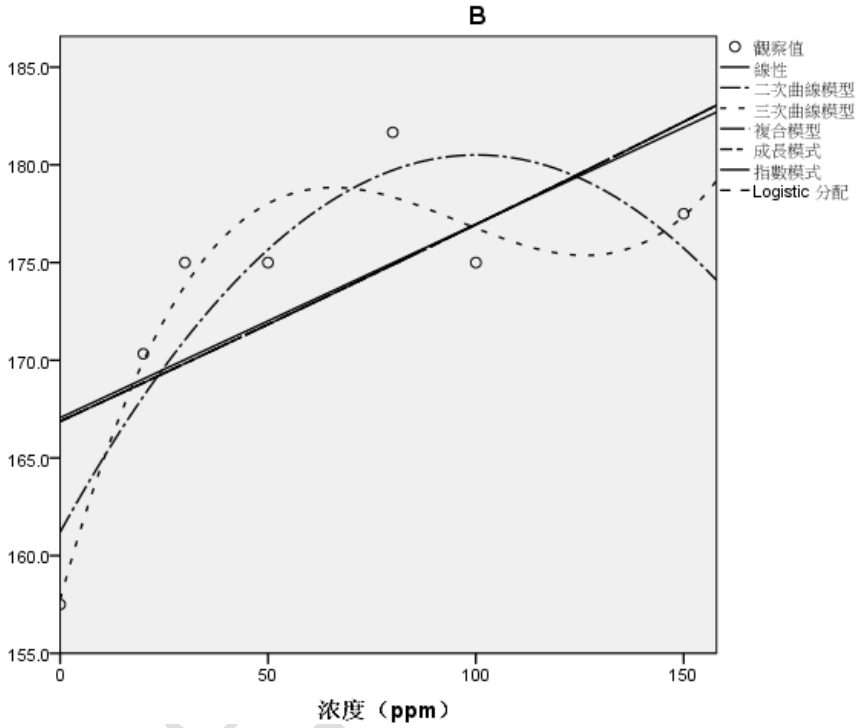


图 4　B 与物质浓度的拟合结果曲线

从表 4 以及图 4，可以看出各个模型的拟合优度，可以看出立方的 R 方最高。同时，从模型的显著性检验结果看，显著性小于 0.05，说明模型成立的统计学意义显著。从表可以看出立方模型的回归系数检验都小于 0.5，说明立方模型的回归系数显著，可知最优模型立方模型的方程式为：B=157.747+0.777*ppm-0.009*ppm^2+3.121E-5*ppm^3。

表 6 立方模型的参数检验

	非标准化系数		标准化系数	T	显著性
	B	标准错误	Beta		

浓度（ppm）	0.777	0.182	5.265	4.273	0.024
浓度（ppm） ** 2	-0.009	0.003	-9.474	-2.988	0.058
浓度（ppm） ** 3	3.121E-5	0.000	4.981	2.382	0.097
（常数）	157.747	2.743		57.502	0.000

同理，可得

$$G=145.909-1.689*ppm+0.019*ppm^2-6.657E-5*ppm^3$$

$$R=152.374-0.372*ppm+0.004*ppm^2-1.187E-5*ppm^3$$

4.2.2 误差分析

对于模型的误差分析，采用相对误差进行分析。相对误差指的是测量所造成的绝对误差与被测量〔约定〕真值之比。乘以 100%所得的数值，以百分数表示。一般来说，相对误差更能反映测量的可信程度。相对误差等于测量值减去真值的差的绝对值除以真值，再乘以百分之一百。

表 7 是通过 MATLAB（见附录 1）得到的相同物质浓度的颜色读数 B、G、R 值进行求解的，其是附件 Data.xls 中的数据和拟合数据的求解模型的误差分析结果：

表 7 相同物质浓度的 B、G、R 的模型误差分析

物质	实验值			拟合值			误差（%）		
浓度	R	G	B	R	G	B	R	G	B
0	153.00	146.75	157.5	152.56	145.59	157.66	0.2876	0.7905	0.1016
20	144.33	115.00	170.00	146.44	118.81	169.93	1.4619	3.3130	0.2173
30	145 .00	114 .00	175.00	144.39	109.83	173.83	0.5921	3.6579	0.6686
50	141.67	99.00	175.00	141.97	98.72	178.14	0.2118	0.2828	1.7943
80	140.67	96.00	181.70	141.37	93.28	178.63	0.4976	2.8333	1.6896
100	139.00	96.00	175.00	142.14	92.52	177.20	2.2590	3.6250	1.2571
150	138.75	86.25	177.50	143.82	77.67	178.42	3.6541	9.9478	0.5183

从表 7 可以看出，模型的误差绝大部分在 2%以内，从而得出回归模型的拟合效果还是不错的，可以得出因变量染色度数与自变量物质浓度之间的关系。

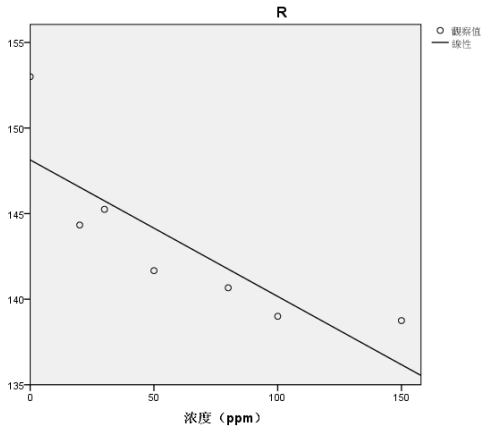
4.3 问题三模型

由于模型采用的是相关分析和回归分析，回归分析是建立在数据的基础上面

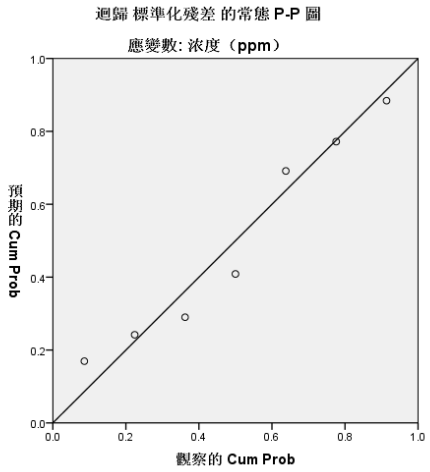
进行模拟曲线，所以数据量越多，模拟的效果越显著，越能分析模型的好坏，使模型更加符合实际情况。

对于维度，本文将维度分为一，二，三维度，通过拟合分析出维度越高，越能准确反应浓度与颜色的函数关系。

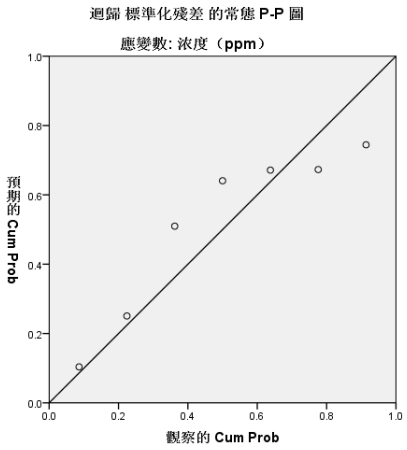
图 5 是在数据维度为一、二、三维度时拟合的结果图。



一维



二维



三维

图 5 模型拟合结果图

通过图形能反应出随着维度的升高，浓度与颜色关系越接近具体函数。

下表是 B、G、R 与物质浓度模型拟合的结果。

表 8 模型摘要 [b]

模型	R	R 平方	调整后 R 平方	标准偏差斜度错误
1	0.838[a]	0.701	0.642	31.185
2	0.855[a]	0.730	0.596	33.131

3	0.904[a]	0.817	0.634	31.536

a. 预测值：（常数），R、G、B

b. 因变量： 浓度（ppm）

模型 1：R 与 ppm 拟合的回归模型

模型 2：R,G 与 ppm 拟合的回归模型

模型 3：R,G,B 与 ppm 拟合的回归模型

在拟合的过程中，随着拟合的进行不仅数据量有所增长，而且数据的维度也在逐渐的增加，根据表 8 中的模型拟合结果可以看出，数据量增多与维度增加模型的拟合效果也有所增加。

五 评价模型

5.1 优点

1.进行回归分析时，先通过散点图进行大体观测，然后通过 spss 构建一次线性模型，二次线性模型，三次线性模型，对数模型，复合模型，logistics 模型，最终进行准确比较各种模型和实际数据直接的误差，最终得出最优的模型—三次模型，提高了模型的准确率，减少误差，使模型更符实际情况。

2.在问题二中，将每个数据与数据的众数进行比较，将数据与众数差别大的进行剔除，减少个别数据对整体的影响，然后取各组数据的平均值，进行回归分析，使模型更加准确。最后通过误差分析判断模型的合理性，误差相对比较小，说明模型符合实际问题，合理的解决了问题。

3.模型使用相关性分析和曲线回归模型，运用相对比较简单的模型解决实际问题。

5.2 缺点

1.模型在假设的基础上进行建模，没有考虑一些问题，例如人为测量出错，物质浓度没有达到饱和，没有晶体析出的等问题，与实际情况有误差。

参考文献

[1]张小利,李雄飞,李军. 融合图像质量评价指标的相关性分析及性能评估[J]. 自动化学报,2014,40(02):306-315. (2013-12-20)[2017-09-16]. http://kns.cnki.net/kcms/detail/11.2109.TP.20131220.0537.054.html

[2]李昕,张明明,李敏. SPSS 22.0 统计分析从入门到精通[M]. 电子工业出版社,2015.

[3]武洋洋,张秉森,李敏. 曲线回归分析在织物染色计算机配色中的应用研究[J]. 青岛大学学报(工程技术版),2013,28(01):22-26. [2017-09-16]. DOI：10.13306/j.1006-9798.2013.01.011.

附录 1

1.求解误差

```
a=xlsread('Data.xls');
ppm=a(:,1);
n=length(ppm);
D=[];
for i=1:n

  b=157.656+0.781*ppm(i)-0.009*ppm(i)^2+3.144*10^(-5)*ppm(i)^3;
  g=145.592-1.673*ppm(i)+0.018*ppm(i)^2-6.577*10^(-5)*ppm(i)^3;
  r=152.555-0.381*ppm(i)+0.004*ppm(i)^2-1.232*10^(-5)*ppm(i)^3;


    D=[D;r,g,b];

end
D;
D=roundn(D,-2);
d1=[];
for j=2:4
    d1=[d1,a(:,j)];
end
d1;

D1=D-d1;
[a1,b1]=size(D);
for i=1:a1
    for j=1:b1
        D2(i,j)=D1(i,j)/d1(i,j)*100;

    end
end
abs(D2)
```