颜色读数辨识物质浓度

摘要

本文为了精准确定待测物质的浓度档位，试确立颜色读数和物质浓度的数量关系模型。针对问题一：颜色读数和物质浓度之间的关系，根据所给数据，将各种物质的实验结果绘制成色卡，直接观察颜色。发现颜色的变化与浓度的改变有关联。随后处理数据并用 EXCEL 绘出颜色读数与浓度的折线图，从图可观察出其颜色读数与浓度是有相关性。经过相关性分析发现有些物质 RGB 有很强的自相关性，因此我们引入灰度来代替原数据中的 RGB。得出组胺与溴酸钾两种物质的浓度与灰度有相关性，其余三种没有相关性。将组胺与溴酸钾的浓度与灰度进行一元线性回归，结果如下：组胺：浓度=-3.038*灰度+327.8 ；溴酸钾：浓度=-5.298*灰度+732.481；工业碱的数据中浓度为 0 到 7 的数据变化极差为 3，所以去除了浓度为 0 的数据组重新进行相关性分析，结果显示工业碱浓度与所有数据相关。将工业碱浓度与灰度导入 SPSS 进行一元线性回归，结果如下：浓度=-0.036*灰度+12.931（灰度<140）经过分析，硫酸铝钾的颜色读数与浓度只在是否存在该物质时存在差异，将浓度设置为存在或不存在，导入 SPSS 与灰度进行相关性分析，显示两者有相关性。奶中尿素的数据经过分析只有 B 与浓度有相关性，将浓度与 B 导入 SPSS 进行一元线性回归，结果如下：浓度=-112.475*B+13571.908（B<140）建立“三准则”分别判断实验数据的准确性。一、RGB2HS 吻合度，利用 RGB 与 HS 关系检验每条数据准确性；二、同物同浓下变异系数，检查同物质同浓度下数据是否稳定；三、同物异浓离散度分析，检查同物质不同浓度下颜色读数是否存在差异。

针对问题二：首先对附件二中的数据进行检验，绘制色卡进行观察，发现 HS 的数值有很大误差，同种浓度下 RGB 变化极小。经过计算发现 HS 的数值相反。更正之后取同种浓度下数据的平均值经计算出平均值的灰度。绘制折线图，发现颜色读数与浓度间存在较小关系。将数据导入 SPSS 进行相关性分析，确定颜色读数与浓度间存在相关性。之后对 RGB 进行相关性分析，发现有很强的自相关性。因此我们取灰度与浓度的数据导入 MATLAB 进行一元线性回归，得到结果为：浓度=-3.612*（0.2989*R+0.587*G+0.114*B）+515.3，并发现模型误差较大，后建立多元 Logit 模型和指数模型浓度 $= 1.653 * 10^7 * e^{(-0.1032*灰度)}$进行修正，发现误差得到极大改善可初步用于实际。

针对问题三：分别说明数据量和颜色维度对模型的影响。数据量分析取数量性差较大的工业碱和硫酸铝钾进行分析。工业碱的数值较少，所以数据的准确性很难检验，异常值不易发现，影响模型准确性。硫酸铝钾的数据较多，易检验准确性，但异常值出现几率大，处理不当会对模型会有很大影响。5 种颜色维度由于数据错误不全，单位不全，对模型有很大影响。随着颜色维度数据的增多，模型会更加稳定准确。

关键词： 比色法 RGB 与 HSV 变异系数 相关性分析 灰度

一、问题重述

比色法是常用的物质浓度检测法，由于不同人对颜色的敏感度不同，使得结果精度受到很大影响。随着照相技术的提高，比色法的使用日趋精准。要求通过照片中的颜色读数，建立与物质浓度间的数量关系，获得待测物质更准确的浓度。

1.1 问题一

（1）通过附件 1 所给出的 5 组数据确定各物质颜色读数与物质浓度的关系。

（2）给出评价标准并评价已知数据的精准程度。

1.2 问题二

（1）通过附件 2 所给出的数据，建立颜色读数与浓度间的模型

（2）对（1）中建立的模型进行误差分析

1.3 问题三

（1）分析数据量对模型的影响

（2）分析颜色维度对模型的影响

二、问题假设

（1） 假设 R、G、B、H、S 的读数未受主观因素影响

（2） 不考虑物质杂质对浓度（ppm）的影响

三、符号说明

符号	含义
V	颜色亮度
S	饱和度
H	色调
R	红色颜色值
G	绿色颜色值
B	蓝色颜色值
HD	颜色灰度
R'	归一化红色颜色值
G'	归一化绿色颜色值
B'	归一化蓝色颜色值
S'	通过归一化后的颜色值带入模型中计算出的饱和度（0-1）
S''	通过归一化后的颜色值带入模型中计算出的饱和度（0-255）
H'	通过 R、G、B 值计算出的亮度
$E1$	原数据饱和度与计算数据饱和度之差
$E2$	原数据亮度与计算数据亮度之差
SD	相同物质的 R、G、B、H、S 的标准差
N	数据的数量
Xi	每一分类的具体数值
$\overline{X}$	每一分类的平均数
$C.v$	变异系数
Me	总体的平均数
$RMSE$	均方根误差

四、问题分析

首先我们需要查找文献学习颜色 RBG 与 HSV 等专业术语的知识，为后续分析做准备。

4.1 问题一分析

问题需判断每种物质每次实验的颜色读数与物质浓度之间的关系，由于颜色读数由 RGB 值构成，所以我们队首先将各组 RGB 值录入 MATLAB，用数据绘制出色卡。将各组色卡进行比较，可直观的观察到实验中每组颜色与浓度之间的关系。之后将附件 1 的各组数据进行分组处理分析，绘制关于颜色与浓度之间的折线图，最后再将数据导入 SPSS 进行相关性分析，找出颜色与浓度关系，并建立数学模型，给出实验数据优劣的评价准则，并评价五组数据优劣。

自行建立准则，对已知数据进行评估。考虑该附件没有明确指明各数值的单位以及转换方式，所以针对此问题我们建立了关于 RGB 与 HS 的数学模型，通过模型的计算求解可得出一个准确的计算数值，然后将原数组与计算数组进行对比观察误差。

4.2 问题二分析

4.2.1

问题需要根据附件二中的数据，建立颜色读数与物质浓度间的模型。先对数据进行检验。绘制色卡进行观察。取同种浓度下颜色读数的平均值，绘制折线图并导入 SPSS 进行相关性分析。计算出平均颜色读数的灰度，做灰度与浓度间的折线图和相关性分析。之后导入 MATLAB 中进行拟合，得到模型。

4.2.2

问题需要将回归出的模型进行误差分析，将其导入 MATLAB 中进行绘图分析。

4.3 问题三分析

4.3.1

问题要求讨论数据量对模型的影响，取附件 1 中数据量差距最大的工业碱和硫酸铝钾分别进行分析。

4.3.2

问题要求讨论颜色维度对模型的影响，根据附件中颜色体系的数目多少及完整性对模型的影响进行分析。

五、模型建立与求解

根据参考文献[1]基于三刺激理论，我们的眼睛通过光对视网膜的锥状细胞中的三种视色素的刺激来感受颜色。这三种色素分别对波长为 630nm（R）、530nm（G）和 450nm（B）的光最敏感。通过对光源中的强度进行比较，我们感受到光的颜色。这种视觉理论是使用三种颜色基色：红、绿和蓝在视频监视器上显示彩色的基础，称为 RGB 颜色模型。除由了一组基色的表示方法，HSV 模型使用对用户更直观的颜色描述方法。为了给出一组颜色描述，用户需选择一种光谱色并加入一定量白色和黑色来获得不同的明暗、色泽和色调。这个模型的颜色参数的色彩（H）、色彩饱和度（S）和明度值（V）。HSV 模型的三维表示从 RGB 立方体演变而来于 RGB 和 HSV 是可以相互转换的，所以我们分析问题时主要用 RGB 分析物质浓度用 HS 进行检验，在文献中 RBG 和 HSV 转化的公式如下：

$$V = \max(R,\ G,\ B)$$

$$S = \begin{cases} V - \dfrac{\min(R,\ G,\ B)}{V} & \text{if } V \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

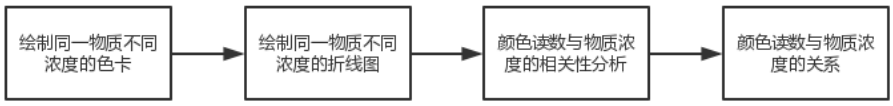
$$H = \begin{cases} \dfrac{60(G-B)}{V - \min(R,\ G,\ B)} & if\ V = R \\ 120 + \dfrac{60(B-R)}{(V - \min(R,G,B))} & if\ V = G \\ 240 + \dfrac{60(R-G)}{V - \min(R,G,B)} & if\ V = B \end{cases}$$

$V \in [0, ;\ S \in [0,1],\ H \in [0,360]$

在后续分析中，我们发现文献中的这个公式的取值范围和题目所给不同，甚至 H 和 S 的维度交换，在下面分析建模中我们将详细指出。

5.1 问题一模型的建立及求解

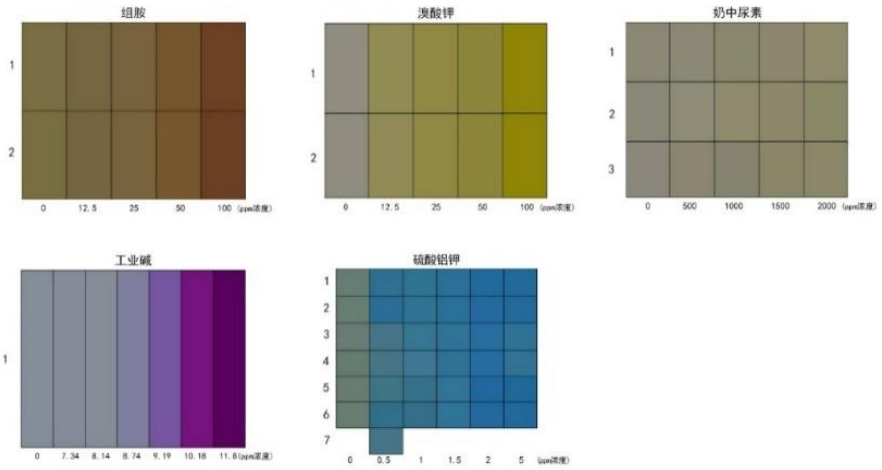
5.1.1 模型建立思路



5.1.2 问题一的分析求解

5.1.2.1 画色卡观察物质浓度间关系

在题目附件 1 中提供了 5 种物质的在不同浓度下的颜色读数 R、G、B、S、H。将每种物质每次测量的 RGB 数值转换为颜色后绘制色卡，清楚直观的看出颜色读数与物质浓度之间关系。因数据众多，为了方便制图，我们选择用 MATLAB 软件绘制（详细程序见附录）。在 MATLAB 中有一套自己的颜色数值标准，RGB 色彩模式的强度值为 0~255，而在 MATLAB 中 RGB 的强度值为 0~1，所以在程序(见附录)中要把数值转换。由此我们得出了相对应的色卡，再用 ADOBE PHOTOSHOP 进行整合以便对比。五种物质在不同浓度下的各次读数整理后如下图：



根据上图可得知组胺、溴酸钾的颜色物质浓度关系明显，工业碱在浓度 8ppm 以后颜色和浓度关系明显，而硫酸铝钾在 0 浓度和其他的浓度的差异比较明显，

而0.5至5ppm的颜色差异并不明显，而奶中尿素从肉眼上看颜色差异并不明显。

5.1.2.2 绘制浓度与颜色读数关系的折线图

将同一种物质不同浓度的数据进行分组，其中以组胺数据为例，整理后如下表（1），其他数据做相同处理见附件（1）。

组胺	浓度	B	G	R	H	S
组一	0	68	110	121	23	111
	12.5	66	102	118	20	112
	25	62	99	120	19	122
	50	46	87	117	16	155
	100	37	66	110	12	169
组二	0	65	110	120	24	115
	12.5	64	101	118	20	115
	25	60	99	120	19	126
	50	46	87	118	16	153
	100	35	64	109	11	172

表 1

之后将每个小组的数据分别绘制成折线图，可以观察颜色读数与浓度间的关系。组一如图（1），组二如图（2）。

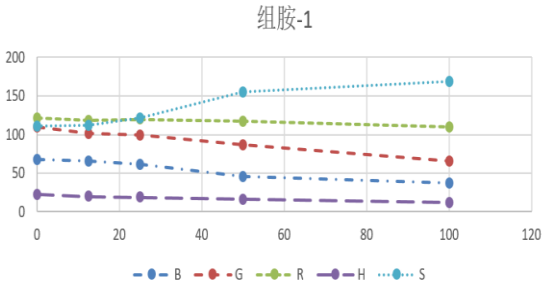


图 1

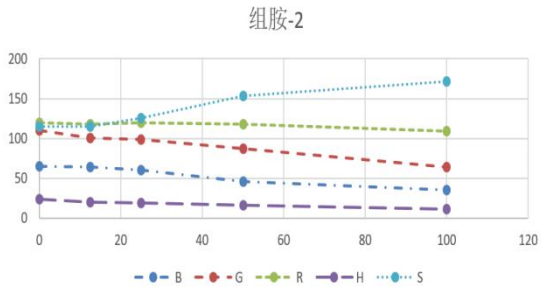
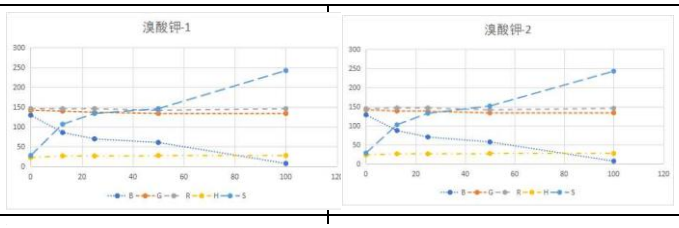
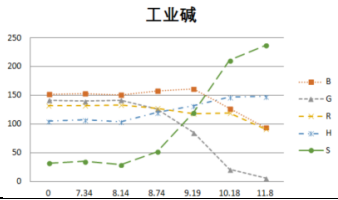
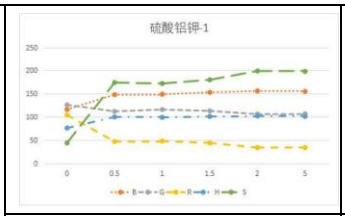
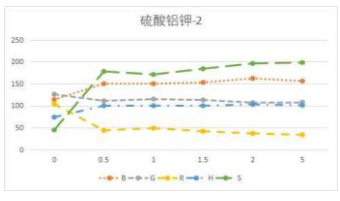
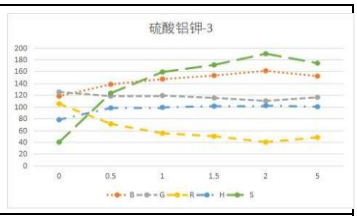
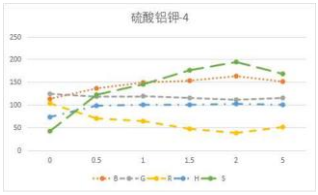
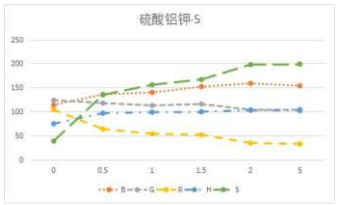
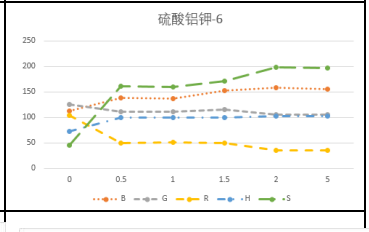
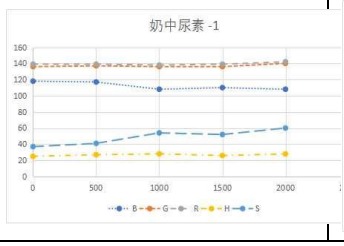
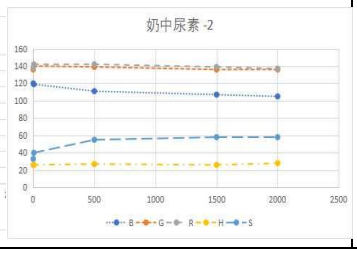
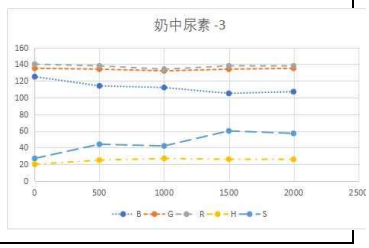


图 2

由图（1）和图（2），可观察出颜色读数在随着物质浓度而变化。可以初步认为组胺物质浓度与颜色读数间存在某种关系。同理，可以得到其他颜色读数与物质浓度的折线图如下表。我们发现组胺、溴酸钾在有两个颜色读数变化趋势明显，工业碱的在溶液 ppm 较高的时候读数变化明显，硫酸铝钾在 ppm 较低的时候变化明显，而奶中尿素在差异为 0-2000 的 ppm 中颜色读数差异较小。

溴酸钾		
工业碱		

硫酸铝钾			
			
奶中尿素			

将整理后数据导入 SPSS 中，运用皮尔逊分析法对数据进行相关性分析，可得出皮尔逊相关系数和显著性如下表

物质	实验组数	类型	B	G	R	H	S
组胺	1	皮尔逊相关性	-0.970	-0.998	-0.951	-0.983	0.955
		显著性	0.006	0.000	0.013	0.003	0.011
	2	皮尔逊相关性	-0.981	-0.997	-0.914	-0.978	0.974
		显著性	0.003	0.000	0.030	0.004	0.005
溴酸钾	1	皮尔逊相关性	-0.952	-0.862	-0.177	0.686	0.948
		显著性	0.012	0.060	0.776	0.201	0.014
	2	皮尔逊相关性	-0.960	-0.875	-0.152	0.722	0.957
		显著性	0.009	0.052	0.807	0.169	0.011
奶中尿素	1	皮尔逊相关性	-0.868	0.639	0.626	0.606	0.946
		显著性	0.057	0.246	0.259	0.278	0.015
	2	皮尔逊相关性	-0.949	-0.578	-0.585	0.646	0.840
		显著性	0.014	0.308	0.300	0.239	0.075
	3	皮尔逊相关性	-0.909	0.000	-0.289	0.741	0.910
		显著性	0.033	1.000	0.638	0.152	0.032
硫酸铝钾	1	皮尔逊相关性	0.579	-0.694	-0.611	0.495	0.586
		显著性	0.229	0.126	0.198	0.318	0.222
	2	皮尔逊相关性	0.552	-0.642	-0.595	0.487	0.575
		显著性	0.257	0.170	0.212	0.327	0.233
	3	皮尔逊相关性	0.586	-0.480	-0.615	0.497	0.608
		显著性	0.221	0.335	0.194	0.316	0.200
	4	皮尔逊相关性	0.550	-0.571	-0.589	0.486	0.589
		显著性	0.259	0.236	0.218	0.328	0.219
	5	皮尔逊相关性	0.676	-0.789	0.742	0.572	0.711
		显著性	0.141	0.062	0.091	0.235	0.113

	6	皮尔逊相关性	0.702	-0.681	-0.639	0.535	0.644
		显著性	0.120	0.136	0.172	0.274	0.167
工业碱	1	皮尔逊相关性	-0.491	-0.664	-0.624	0.708	0.658
		显著性	0.264	0.104	0.134	0.075	0.108

由上表可以确定组胺的颜色读数中的 R、G、B、H、S 都与浓度有很强的相关；溴酸钾的颜色读数中的 B、S 与浓度有较高的相关性；奶中尿素的颜色读数 R、S 与浓度有相关性，硫酸铝钾的颜色读数与浓度没有相关性。工业碱的皮尔逊系数和显著性表明没有相关性，而直观的从图可观察出其颜色读数与浓度是有相关性的，再者上述 RGB 颜色读数本身也具有自相关性，因此我们下面进行详细分析。

5.1.2.3 灰度分析

虽然我们发现有些物质颜色读书 RGB 与浓度有相关关系，当颜色读数 RGB 具有自相关性时，不能直接用 RGB 三个变量直接所为自变量与浓度进行回归分析，在图形学中，灰度具有很高的识别度，在医学中 CT 等更是由于识别度高等原因一直使用灰度照片。因此，我们这里选用利用 RGB 灰度这个值来代替 RGB 读数。根据文献[2]我们得到方法计算灰度，公式如下$HD = 0.2989R + 0.587G + 0.114B$。各物质平均浓度、灰度及其相关性如下：

物质	浓度	灰度	相关性	p 值	物质	浓度	灰度	相关性	p 值
组胺	0	108.17	-0.997	0	硫酸铝钾	0	117.63	-0.678	0.139
	12.5	102.26				0.5	101.00		
	25	100.94				1	100.36		
	50	91.43				1.5	98.88		
	100	74.99				2	92.04		
溴酸钾	0	140.61	-0.947	0.015		5	93.44		
	12.5	134.59			奶中尿素	0	270.1721	-0.204	0.742
	25	131.54				500	268.2063		
	50	126.88				1000	263.6968		
	100	122.21				1500	265.7953		
工业碱	0	140.14	-0.662	0.105		2000	269.627		
	7.34	139.08							
	8.14	140.32							
	8.74	129.93							
	9.19	103.52							
	10.18	62.37							
	11.8	41.44							

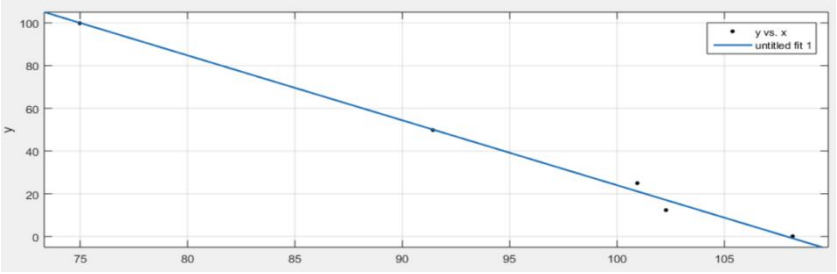
即：组胺和溴酸钾的浓度都与灰度有显著相关关系。

5.1.3 各物质颜色度数与浓度关系

5.1.3.1 组胺颜色读数与浓度关系

我们看到这两组颜色读数 RGB 差异很小，最大极差绝对值为 3，因此这里选用他们颜色读数的平均值来讨论，由于这组数据 RGB 自相关性较大两两相关性分别为：0.968431204、0.87053481、0.94458497：因此选用灰度作为自变量，对浓度进行一元线性回归得到：

y=-3.038x+327.8，该方程 p 值小于 0.05，说明该方程显著，回归系数 p 值也小于 0.05，说明回归系数显著。

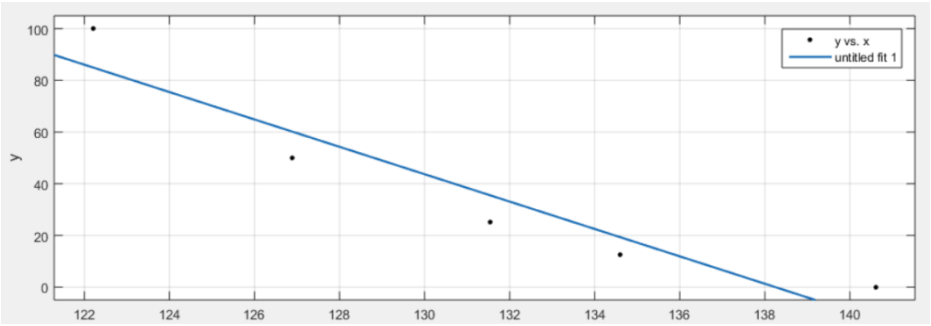


即：浓度=-3.038(0.2989R + 0.587G + 0.114B)+327.8，浓度与灰度负相关，与 RGB 读数都负相关。

5.1.3.2 溴酸钾颜色读数与浓度关系

我们看到这两组颜色读数 RGB 差异很小，最大极差绝对值为 2，因此这里选用他们颜色度数的平均值来讨论，在上述分析中我们发现这组从肉眼上看，其实随着浓度的变化颜色变化比较明显，但是从颜色读数上来说只有 B 与浓度有显著相关性，可能是因为其他颜色维度变化不大，由于这组数据 RGB 自相关性有一对较大分别为 0.897990569，0.07348246，0.45569145，因此选用灰度作为自变量，对浓度进行一元线性回归得到：

y=-5.298x+732.481，该方程 R=0.947，p 值小于 0.05，说明该方程显著，回归系数 p 值也小于 0.05，说明回归系数显著。



即：浓度=-5.298(0.2989R + 0.587G + 0.114B)+ 732.481 浓度与灰度负相关，与 RGB 读数都负相关。

5.1.3.3 工业碱与浓度关系

经过观察图像发现工业碱的颜色读数与浓度存在一定关系。但通过皮尔逊分析法和显著性的检验，显示两者没有相关性。我们继续分析，浓度为 0 与浓度 7 的色彩读数变化较小极差仅为 3，7 以上颜色读数变化明显。所以除去浓度为 0 的一组数据，重新进行计算。我们看到这两组颜色读数 RGB 差异很小，最大极差绝对值为 2，因此这里选用他们颜色读数的平均值来讨论，绘制折线图如图（3）

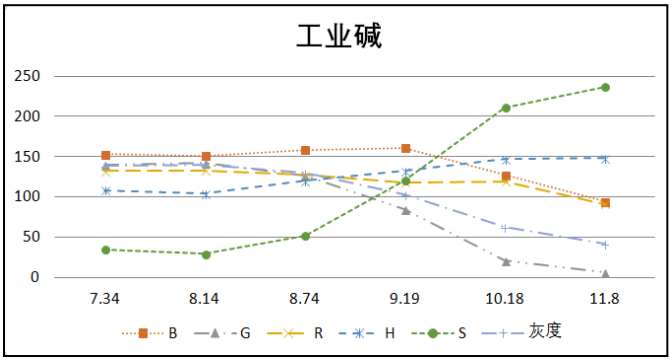


图 3

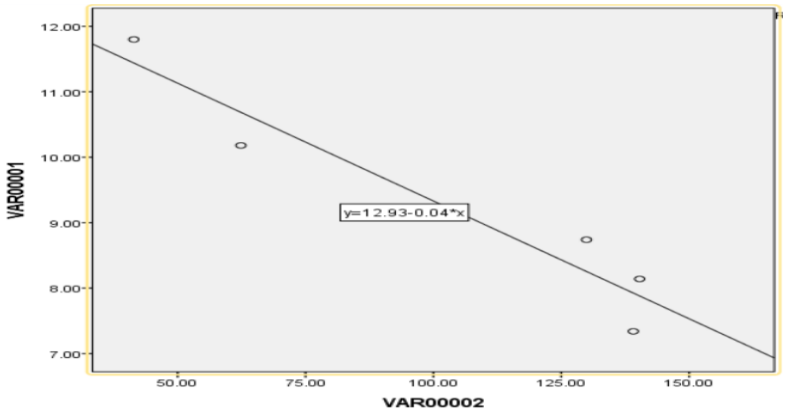
将数据导入 SPSS 进行相关性分析如表（2）

物质	类型	B	G	R	H	S	灰度
工业碱	皮尔逊相关性	-0.865	-0.941	-0.941	0.913	0.940	-0.958
	显著性	.026	0.005	0.005	0.011	0.005	0.003

表 2

根据图中的皮尔逊相关性与显著性可以确定，工业碱的颜色读数与浓度间有相关性。计算浓度与灰度相关性为 0.958 及 RGB 读数自相关分别为 0.833724，0.86190527，0.874326497，因此选用灰度作为自变量，对浓度进行一元线性回归得到：

y=-0.036x+12.931（x<140），该方程 R=0.95，p 值小于 0.05，说明该方程显著，回归系数 p 值也小于 0.05，说明回归系数显著。



即：浓度=-0.036 $(0.2989R + 0.587G + 0.114B)$+ 12.931$(0.2989R + 0.587G + 0.114B < 140)$ 浓度与灰度负相关，与 RGB 读数都负相关。

5.1.3.4 硫酸铝钾与浓度关系

硫酸铝钾这组数据比较特殊，每一种的读数差异比较大，但是实际颜色不算大，因此我们也只能选用平均值作为讨论数据，观察图像发现工业碱的颜色读数与浓度存在一定关系。但通过皮尔逊分析法和显著性的检验，显示浓度与各组读书及其平均值（参见附录）都没有显著相关性。我们继续分析，硫酸铝钾的色彩读数随浓度变化小，只在有浓度和浓度为零间有明显差异。将浓度为 0 的颜色读数与有浓度的颜色读数进行相关性分析，结果如下表：

物质	实验组数	类型	B	G	R	H	S
硫酸铝钾	平均	皮尔逊相关性	0.874	-0.745	-0.905	0.986	0.919
		显著性	0.000	0.001	0.002	0.003	0.004

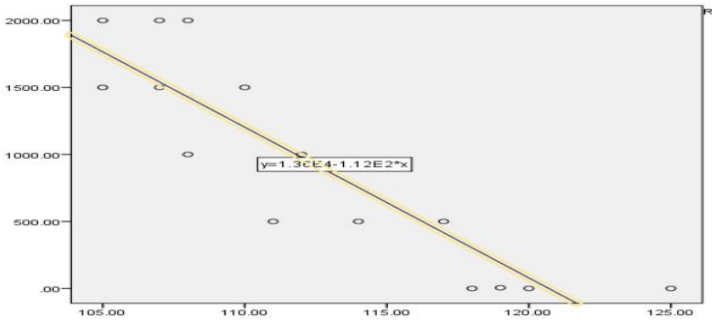
根据所得的皮尔逊相关系数和显著性，可以确定硫酸铝钾的颜色读数可以确定溶液中是否含有该物质,RGB 自相关性如下：0.920829968，-0.81369431，-0.94515598，因此可用灰度直接分辨溶液中该物质浓度是否大于 0.5ppm，即：灰度<100 时，溶液中物质大于 0.5ppm。

5.1.3.5 奶中尿素与浓度关系

奶中尿素从肉眼上看颜色差异并不明显，而观察折线图奶中尿素在差异为 0-2000 的 ppm 中颜色读数差异较小。但是在相关性分析中我们发现浓度与 RGB 中 B 具有显著相关性而与灰度的相关性也较差。根据文献[3]这可能是因为肉眼在该颜色的辨别能力比较差导致，因此以颜色读数 B 为自变量进行一元线性回归，得到：

y=-112.475x+13571.908（x<140），该方程 R=0.892，p 值小于 0.05，说明该方程显著，回归系数 p 值也小于 0.05，说明回归系数显著。

即：浓度=-112.475 B+13571.908，浓度与 B 读数负相关。

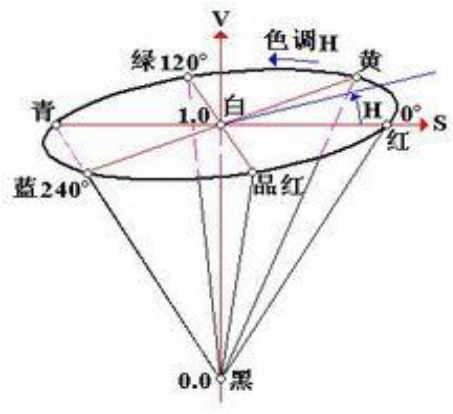


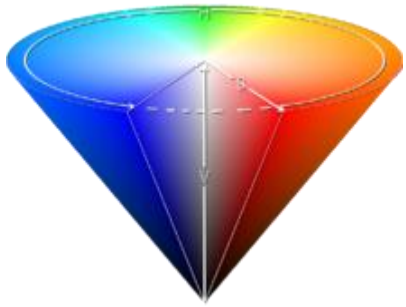
5.1.4 建立数据评价准则

5.1.4.1 评价准则 1——RGB2HS 吻合度（检验每一条数据的准确性）

RGB 与 HS 分别属于两种不同的颜色判定的体系，通过 RGB 与 HSV 模型之间的转换的公式，将每组数据中 R、G、B 的值变换到对应的 S、H、V 的值，并与附件 1 中 S、H 原始数据相比较求出误差。并定义吻合度为：饱和度吻合度 $=\left(1-\frac{\sum^{|(S-S'')|}_{S}}{n}\right)*100\%$，色调吻合度 $=\left(1-\frac{\sum^{|(S-S'')|}_{S}}{n}\right)*100\%$，即 1—误差绝对值除以原数据的和。吻合度越接近 1 表示该条数据正确性越高。通过模型公式解得 V、S’数值：

我们通过查阅文献[1]得知 HSV 颜色模型是根据颜色的直观特性创建的一种颜色空间,。这个模型中颜色的参数分别是：色调（H），饱和度（S），明度（V），其中 H 用角度度量，可视为一个圆形，取值范围为 0°～360°，从红色开始按逆时针方向计算，红色为 0°，绿色为 120°,蓝色为 240°。以 data2 的二氧化硫数据为例，我们计算得出的二氧化硫颜色读数 H 分布在 270°～287°，而附件 2 中读数 H 分布在 135°～141°显然不是[0,360)的取值范围而是[0,180)，因此我们认为原始数据中 H 的取值范围为 0°～180，如图：





因此计算得出的 H 数值乘以$\frac{1}{2}$再与原始数据比较。再者由于原始数据和计算数据的定义域不同，导致无法进行直观比较，发现原数据与计算数值如下关系：$S''=S'*255$。 因此可得如下计算表(计算过程见附件[2])：

	S吻合度	H吻合度
组胺	99.11%	96.18%
溴酸钾	98.94%	98.06%
工业碱	98.25%	93.49%
硫酸铝钾	99.17%	91.64%
奶中尿素	98.27%	97.91%

结论：五种物质实验数据记录都比较准确，但是更加准确的是组胺、溴酸钾和奶中尿素。

5.1.4.2 评价准则 2—同物同浓下变异系数（检验相同物质 ppm 下数据稳定度）

对实验数据准确性的评估，将同种物质，在相同浓度下的 R、G、B 数值应该相对稳定，利用数据变异系数 $C.v=\frac{SD}{Me}*100\%=\frac{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2}}{Me}*100\%$，建立准则 2：以组胺为例，将同种浓度数据整理，之后把数据代入模型中求出标准差 SD 变异系数 C.v 得下表：

变异系数	浓度ppm	B	G	R	H	S
组胺	0	3.19%	0.00%	0.59%	3.01%	2.50%
	100	2.18%	0.70%	0.00%	0.00%	1.87%
	50	2.32%	0.00%	0.00%	0.00%	2.28%
	25	0.00%	0.00%	0.60%	0.00%	0.92%
	12.5	3.93%	2.18%	0.65%	6.15%	1.24%
溴酸钾	0	0.55%	0.00%	0.49%	3.14%	2.57%
	12.5	1.64%	0.51%	0.49%	0.00%	2.72%
	25	1.02%	0.52%	0.49%	0.00%	0.53%
	50	3.63%	0.00%	0.00%	0.00%	2.87%
	100	0.00%	0.00%	0.00%	0.00%	0.29%
硫酸铝钾	0	1.71%	0.79%	0.61%	2.89%	6.09%
	0.5	4.31%	3.01%	19.26%	1.15%	16.22%
	1	3.76%	2.78%	10.88%	0.52%	6.26%
	1.5	0.27%	1.19%	7.98%	0.51%	3.64%
	2	1.65%	2.60%	6.19%	0.53%	1.72%
	5	1.36%	4.51%	20.64%	0.89%	7.45%
奶中尿素	0	4.75%	0.65%	1.20%	4.30%	7.85%
	500	2.63%	1.84%	1.49%	4.38%	15.80%
	1000	2.57%	2.11%	2.08%	2.57%	17.68%
	1500	2.34%	0.85%	0.42%	0.00%	7.35%
	2000	1.43%	1.93%	1.90%	4.22%	2.62%

通过查阅文献[4]得知，当变异系数大于 15%时视为数据有较大差异。当同种物质相同浓度下，颜色应为极其相近，对应的变异系数应该即为接近并且小于 15%。所以通过比较变异系数的大小，结合色卡颜色的变化并得出结论：组胺在相同浓度下，变异系数较小并小于 15%，所以该组实验数据准确性高。同理：溴酸钾在相同浓度下，变异系数较小，该组实验数据较为准确。奶中尿素相同浓度下，RGBS 变异系数较小，该组实验数据较为准确，而 H 变异系数大，因此可以多读书 H 存在差异性。硫酸铝钾相同浓度下，G、B、H 的变异系数较小数据较为准确，但 R 与 S 的在 5ppm 出现三组大于 15%的变系数，所以判定 R 与 S 不算十分精确。由于实验数据有限，工业碱仅有一组数据，无法进行计算，存在偶然性，故不适用于此方法。

5.1.4.3 评价准则 3——同物不同浓下变异系数离散度（同物质不同浓度颜色读书的区分度）

在对同一种物质不同浓度的比色法实验，我们希望每组读书差异相对较大，因此我们也利用变异系数 $C.v=\frac{SD}{Me}*100\%=\frac{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2}}{Me}*100\%$ 来建立准则 3 的指标。以组胺为例，将不同浓度的数据升序排列，得到表（1）后，把 R、G、B、H、S 值分别代入公式中求的最终 C.v 值 如下表：

		B	G	R	H	S
组胺	变异系数	24.27%	18.85%	3.79%	25.08%	19.20%
溴酸钾	变异系数	63.09%	2.56%	1.31%	7.23%	59.19%
工业碱	变异系数	16.85%	62.41%	12.24%	15.51%	87.12%
硫酸铝钾	变异系数	11.14%	5.46%	43.83%	11.01%	37.10%
奶中尿素	变异系数	5.27%	0.88%	1.03%	5.87%	21.38%

即：结合色卡颜色的变化并得出结论：组胺在相同浓度下，变异系数较大并大于 15%，所以该组实验数据使用比色卡方，颜色差异大，效果好。溴酸钾在不同浓度下使用 B、S 分析较为有效，工业碱在不同浓度下，比色卡颜色差异也比较大，硫酸铝钾在不同浓度下只有 R 与 S 差异行多较大，而奶中尿素在不同浓度下，比色卡颜色差异只有饱和度差异较大。

5.2 问题二模型的建立

5.2.1 数据评价

首先对数据进行检测，发现在同一浓度下的颜色读数极差不超过 4。绘制二氧化硫的色块如图（4）：

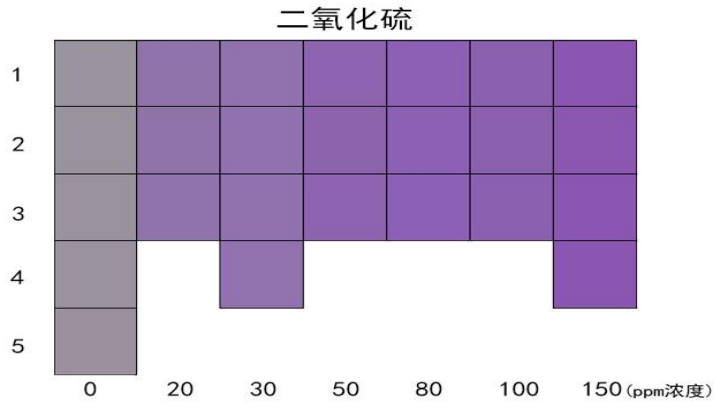


图 4：二氧化硫颜色卡

原题中附件 2 中提供了二氧化硫在不同物质浓度下的颜色读数，我们通过观察发现，H、S 的数据可能有误。我们利用公式将 R、G、B 数值转换为 H、S 数值，发现计算得出的 S 数值近似于附件 2 原始数据中的 H 数值，而计算得出的 H 数值约为原始数据中 S 数值的二倍。H、S 两组数据恰好相反。将数据代入公式中算得 S 吻合度与 H 吻合度分别为 98.83%和 99.62%，应数据比较精确；将 R、G、B、S、H 代入准则二模型中计算标准差 SD 与变异系数 C.v 得下表

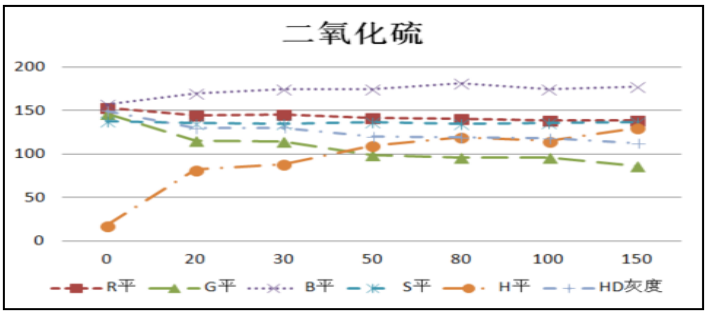
浓度ppm	B	G	R
0	0.29%	0.78%	0.35%
20	0.40%	0.00%	0.90%
30	0.34%	0.00%	0.47%
50	0.41%	0.00%	0.57%
80	0.41%	0.00%	0.32%
100	0.00%	0.00%	0.57%
150	0.36%	0.58%	0.33%

结论：相同浓度下 B、G、R、H 对应的变异系数 C.v 皆小于 15%，所以数据较为准确；但 S 中有一组数据大于 15%，相比较下 S 不算太准确； 将不同浓度下 R、G、B、S、H 取平均值后，与其标准值进行运算，最终得出变异系数如下表

	B	G	R	H	S
二氧化硫	3.51%	18.60%	4.46%	0.86%	40.07%

结论：不同浓度下 G、S 的变异系数较大并大于 15%，但 B、R、H 相对于偏小，所以 G、S 更加有效。

由上图可以直观的观察到同种浓度下色差极小，因此我们取在同种浓度下的颜色读数的平均值进行计算。计算出平均色彩读数的灰度并绘制出浓度 RGB 折线图，如图（5）



图中皮尔逊相关系数和显著性，可以确定二氧化硫的颜色读数与浓度间存在相关性。将 RGB 导入 SPSS 进行相关性分析，结果如表（8）

		R	G	B
R	皮尔逊相关性		0.988	-0.895
	显著性		0.000	0.007
G	皮尔逊相关性	0.988		-0.916
	显著性	0.000		0.004
B	皮尔逊相关性	-0.895	-0.916	
	显著性	0.007	0.004	

图 8

发现浓度与 RGB 强自相关，通过上图可以初步确定二氧化硫的颜色读数与浓度间存在相关性，将数据导入 SPSS 进行相关性分析，因此计算灰度并计算相关性如表（7）

物质	实验组数	类型	B	G	R	HD
二氧化硫	1	皮尔逊相关性	0.696	-0.867	-0.844	0.968
		显著性	0.000	0.000	0.000	0.000

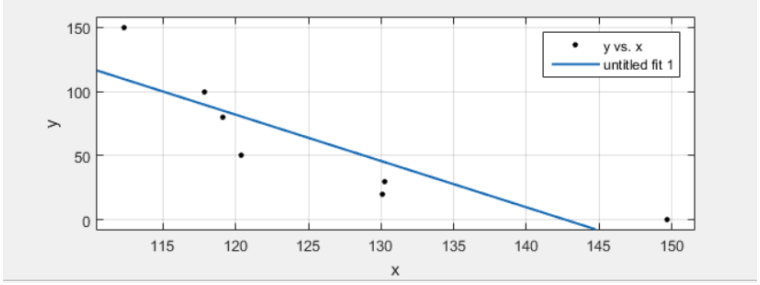
表 7：二氧化硫浓度相关性

5.2.1 预测模型一：

将浓度与灰度导入 MATLAB 进行一元线性回归，得到模型为：f(x) = -3.612*x

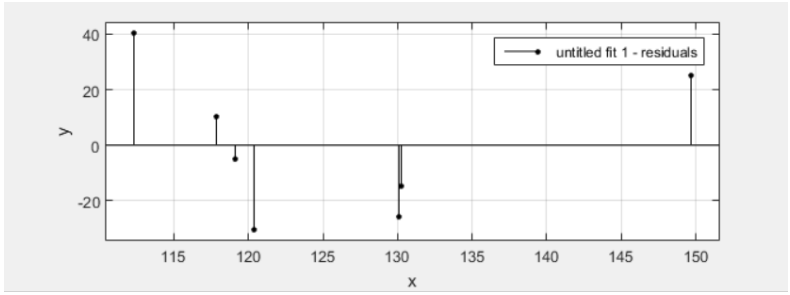
图 6

+ 515.3，　R 方值为 0.742 说明方程显著即：　浓度=-3.612*（0.2989*R+0.587*G+0.114*B）+515.3 图像如图



得到模型显著性为 0.013 说明模型显著。系数显著性为 0.013 说明系数显著。之后对一元线性回归的模型进行误差检验，结果如图（7）

图 7



由图可知误差最大值小于 3 倍 RMSE 值，所以模型误差较小。这个模型中自变量数据每组差异较小。但是这是模型中 y 的取值为 0-150，我们发现真的用这个模型预测浓度时两段的差异不能接受，如下表，因此建立模型二

浓度	绝对误差（平均值）		浓度	绝对误差（平均值）
0	25.31483056		80	5.084553
20	25.5079892		100	10.37095
30	14.7169913		150	40.45808
50	30.389314			

5.2.1 预测模型二：

我们将 0-150 的浓度看做有序分类变量，以灰度为自变量，建立一元多分类 logistics 回归，模型为：预测变量为 x 的基线-类别 logit 模型为：

$$\ln(\frac{\pi_j}{\pi_J}) = \alpha_j + \beta_j x, j = 1, \cdots, J-1$$

不管哪个类别作为基线，对于同一对类别都会有相同的参数估计；即基线类别的选择是任意的，这个模型虽然分类准确性极高，但是没有办法预测除 0，20,30,50,100,150 外其他浓度，因此并不能真正应用。

浓度	灰度	预测分类	预测概率		浓度	灰度	预测分类	预测概率
0	150.51	0	1		50	120.51	50	0.95
0	149.92	0	1		50	120.09	50	0.95
0	149.45	0	1		50	120.62	50	0.95
0	149.45	0	1		80	119.13	80	0.95
0	149.04	0	1		80	119.24	80	0.95

20	129.93	20	0.97		80	118.95	80	0.95
20	129.81	20	0.97		100	117.85	100	0.97
20	130.45	20	0.97		100	117.74	100	0.97
30	130.09	30	0.97		100	117.96	100	0.97
30	130.32	30	0.97		150	112.32	150	1
30	130.21	30	0.97		150	112.79	150	1
30	130.51	30	0.97		150	111.91	150	1
					150	112.32	150	1

5.2.1 预测模型二：指数模型

再观察散点图，发现灰度与浓度成指数关系，因此使用指数模型进行回归：得到模型：$y = 1.653 * 10^7 * e^{(-0.1032*x)}$,R 方为 0.966，拟合程度明显高于线性模型，即得模型为：浓度 $= 1.653 * 10^7 * e^{(-0.1032*灰度)}$

$= 1.653 * 10^7 * e^{-0.1032(0.04175*0.2989R+0.587G+0.114B)}$ 。

计算绝对误差如下：误差得到很大改善，可初步应用实际

浓度	线性误差	指数误差		浓度	线性误差	指数误差
0	25.31483056	3.240967985		80	5.084553	4.139326
20	25.5079892	4.495050236		100	10.37095	13.62429
30	14.7169913	6.059244417		150	40.45808	4.884433
50	30.389314	16.34987651				

5.3 问题三

5.3.1 数据量对模型的影响

因为讨论数据量对模型的影响，所以选区附件一中数据量最少的工业碱和数据量最多的硫酸铝钾分别进行分析。

工业碱的数据共有 7 个且浓度各不相同，无法进行同浓度下的数据检验。也无法证明数据的可靠性，如果出现异常值很难发现和处理。较少的数据进行回归会得到比较理想的模型，但也降低了模型的可靠性和预测范围。

硫酸铝钾的数据共有 35 个，有充足的数据进行数据间的交叉验证，提高了回归模型的可靠性和预测范围。但同时数据量的增加也增大了异常值出现的可能，单个数值的异常可能会对整体的分析造成很大的影响。

所以数据量过多或过少都会对模型造成不同程度的影响。如何检验数据的准确性和处理数据中异常值，对数据回归的模型会有很大的影响。

5.3.2 颜色维度对模型的影响

颜色维度也就是不同的数值，如 R、G、B、S、H 就为 5 种颜色维度。其中三组数据可组成一种颜色体系。如 RGB 和 SH 即为两种不同的颜色体系。不同体系间可对颜色维度的数据新型相互验证。附件中给出了 5 种颜色维度，两种不同的颜色体系。其中 SH 体系的数据不全，给之后两组体系数据的相互检验带来了很大的不便，使得数据的准确性降低，回归出的模型可靠性也得不到保证。

在分析 R、G、B、S、H 时,附件中未给出各颜色维度的单位。所以在分析的过程中，数据是录入错误还是单位不一致很难区分，给之后的模型回归造成了很大的困难。所以颜色维度的数据量给出的越多，单位越全，回归出的模型准确性越能得到保证。

六、模型的评价与改进

6.1 模型的评价

模型优点，根据题中所给的颜色读数绘制出色卡，可清楚直观地看出颜色与浓度之间是否存在关系，为接下来的讨论奠定了基础。绘制颜色与浓度关系的折线图，方便了我们观察两者间的关系。利用皮尔逊分析法对数据进行相关性分析，可得出皮尔逊相关系数和显著性，判断各物质的颜色读数于浓度是否具有相关性。当颜色读数 RGB 具有自相关性时，不能直接用 RGB 三个变量直接所为自变量与浓度进行回归分析，因此我们将 RGB 数值转为识别度很高的灰度进行分析。通过 H、S 吻合度检验，判断实验数据的准确性，并发现题目数据中的错误。

6.2 模型的改进

因题目中部分物质的数据不足，无法通过本模型对颜色读数与物质浓度间的关系进行准确地分析，得到的结果可能有较大误差。我们从线性模型到指数模型的误差确实得到了改善，但是在 50ppm 的误差依然有 10%，这实际还是不够好的，应该在更精确的仪器获得的数据基础上继续改善。

七、参考文献

书籍：

[1][美]赫恩(Hearn,D.)，巴克(Baker,M.P.)，卡里瑟斯(Carithers,W.R.)著；蔡士杰，杨若瑜译.《计算机图形学：第 4 版》.电子工业出版社,2014

[4] 罗良清、魏和清.《统计学》.中国财政经济出版社,2011

网站资源：

[2]百度文库.从 RGB 色转为灰度色算法.

https://wenku.baidu.com/view/0da4374549649b6648d747b2.html?from=search, 2015.9.11

[3]知乎.Wang J.普通人肉眼可以看出 RGB 在多大程度上的差别.https://www.zhihu.com/question/33688135?sort=created.2015.7.31

八、附录

MATLAB 绘制色快程序：

```
num = xlsread('d:\data1.xls','sheet1','B2:G6');color1 = num;
color1(:,2:4) = color1(:,2:4)/255;
s = size(color1);
gshu = s(1,1);
kuan=2;
%rectangle('Position',[1 2 kuan*gshu 5])
for i=1:gshu
    rectangle('Position',[1+(i-1)*kuan 2 kuan 5],'facecolor',[color1(i,4),color1(i,3),color1(i,2)])
end
```