The Association Between Urbanization and Water Chemistry Parameters

G11 AP Statistics

Cheng (Paul) Ai

Abstract

Due to the rapid development of human society, urbanization become a significant and an indispensable activity for the development in human civilization. Nevertheless, this action actually causes the deterioration of environment, which become a barrier to human. Thus, monitoring environment and properly control the step of urbanization become an important task.

This passage focusses on finding a relation between some water chemistry parameters, something that can indicate the quality of water, and the degree of urbanization by collecting data from a place in British from 1998-2005. The analysis of data shows mainly positive relationship between the urbanization degree and concentration of chemicals in nearby river system.

By using the relation found in the observational study, this passage continually to build a scoring system to evaluate the degree of urbanization in a certain area by establishing a linear system. This system tells people a predicting degree of urbanization, and whether this action is excess, thus helps the monitoring and protection of the environment.

Key Words: Water Chemistry Parameters; Urbanization Degree; Linear System

Content

1. Introduction

Since the end of two Industrial Revolutions, the development of human technology reaches to a peak in around 20 years. This development produce lots of convenience for human activities on excavate and use of natural resources, and also brings a completely new life for human in that time. However, the negative by-product of Industrial Revolution soon comes to people's mind, which is the defection on environment. The deterioration on the environment thus soon become a very serious problem in development of technology, and the progress of urbanization or agriculturalization and some other humanize progresses.

Long in the end of last century, American and Canada firstly revised environmental protection law for restrain the further deterioration of natural environment brought by industrialization and urbanization, and investigated some supervising methods to quantify the pollution by some certain index, in order to solve the problems brought by human activities.

This passage aims to analyze a certain case of urbanization affecting on nearby river system by using statistical evidence and simple regression model, and thus popularize the result to the normal river system in the world, therefore introduce a way to protect the environment.

2. Background Information

Urban Stream Syndrome stands for the "illness" of near-water environment due to the pollution of urbanization or agriculturalizing or some other natural effect of land use which defect the quality of the river system, such as eutrophication, or direct pollution to the river [1]. These defections will influence the whole water system which spread to the nearby urban city or residential area. Therefore, supervising quality of water system becomes an important task for every country. For a particularly water system, in order to quantify its quality, we choose some measurable quantity and create a certain mechanism to evaluate the quality of water, and give a final response of how those land use patterns may influence the river system, in order to make the best use of land without causing lots of serious environmental issue.

In the recent study, scientists have found that the effects of urbanization on river system based on system structure, peak flows or some other visible quantities. To illustrate, research done around Taihu Region, China indicated the relation between water surface area and population growth [2], another research done around San Antonio River Basin in Texas showed the relation between peaks flow and urbanization degree [3] … These researches firstly introduce the analysis of urbanization into account of river system defection in macroscopic view. However, there is lack of analysis from micro-view of water system, which called water chemistry parameters. These parameters include chemical properties such as metal or ion concentration, which create a complete evaluation system or index to help government supervise water quality [4].

3. Experimental Design

This passage mainly focuses on determine relation between urbanization degree, using environmental usage to predict, and water quality, which indicate by water chemistry parameters. Therefore, I create an observational study, from collecting data to data analysis and result summary.

3.1 Data Collection

Data was adapted from the UK Government River Water Quality Monitoring 1990 to 2018[5]. In this database, it provides a large amount of raw data collected from a certain river system in England, specific shows in the area below in Figure 1.

3.2 Variable Analysis

I choose land use patterns as urbanization index, thus set it as explanatory variable, and chemistry parameters, such as gases dissolved, ion concentration, as response variable. However, other aspects, natural environment, particularly states as primary basin, and time difference will be the confounding variable, which will be discuss specifically in method and procedure.

Figure 1. Data Covered Area

4. Procedure & Method

4.1 Random Sampling Methods

As stated above, since there are two confounding variables in the raw data, which is the "Primary Basin" and time difference. Therefore, here I firstly conduct a stratify sampling method in order to eliminate the effect from confounding variable, which is the difference between each different natural environment, since it will affect concentration of certain substance in the river, which will influence the determination of water quality based on water chemistry parameters.

Based on the raw data, I assign 56 stratums, from 1-56, which determined by how many "Primary Basin" has been supervised. After that, I filtered them by how many data in it in order to satisfy the condition of scope of research, the result is showed in Table 1. Since the scope of inference need to be larger enough for popularizing the result, and some invalid data in each stratum, I only choose strata that has larger than 1000 data, which is enough for analysis even regard with invalid data due to some unexpected reasons.

Table 1. Numbers of Data in Each Stratum

	A	B	C	D	E	F	G	H
1	Primary Basin	Numbers of data		Primary Basin	Numbers of data		Primary Basin	Numbers of data
2	Arney	3364		Fane	587		Newry	3142
3	Ballinamallard	2612		Faughan	3663		Owenkillew	5474
4	Ballinderry	5430		Finn	377		Quoile	2415
5	DRUMNAGRESHIAL TRIB AT DRUMNAGRESHIAL	984		Finn (Fermanagh)	2237		Roe	4046
6	Belfast Lough (North)	2181		Flurry	587		Roogagh	360
7	Blackwater	9875		Foyle (with Deele)	1148		SE Down Streams	2276
8	Blackwater (Ards)	6		Glenavy	1241		Sillees	2450
9	Blackwater (Ards) (Enler/Blackwater)	398		Kesh	2400		Six Mile Water	3893
10	Bradoge	801		Kilkeel & Mourne Streams	4130		Strule	1480
11	Burn Dennet	1500		Lagan	7655		Swanlinbar	765
12	Burnfoot	136		Lough Foyle (East)	10		Termon	528
13	Burnfoot (L Foyle West)	758		Lough Foyle (South)	1347		Upper Bann	8241
14	Bush	2104		Lough Neagh & Peripherals	2516		Upper Erne	3182
15	Camowen	2133		Lough Neagh Peripherals	137		Woodford	510
16	Castletown	949		Lower Bann	6883			
17	Colebrooke	4427		Lower Erne	3433			
18	Comber (Enler/Blackwater)	626		LEGNAGOOLY BURN AT LEGNAGOOLY BRIDGE	6286			
19	County	294		Mourne	1107			
20	Crumlin	1082		Mourne/Strule	84			
21	Derg	3167		Moyola	3126			
22	Drumragh	3607		N Down & Ards Peninsula	1038			
23	Fairy Water	2041		NE Coast	4202			

Therefore, there are 39 stratums that satisfy the condition. After assign and choose suitable stratums, a cluster sampling method will be again to use in order to eliminate the effect of time. I choose timeline from 1998 to 2005, each year as a cluster, example cluster, year of 1998 is showed below in table 2, and analysis data in it.

Table 2. Example Cluster: Year 1998

	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Site_Sta	Station_	RWB_ID	Primary	X	Y	Date	Time	NO3_N_MG	NO2_N_M	NH4_N_MG	PH	P_SOL_MG
8	Closed	KILFENNAI	<10km2	Foyle (witl	246079	418422	1998/07/1	10:35:00	0.32	0.02	0.04	7.7	0.04
9	Closed	KILFENNAI	<10km2	Foyle (witl	246079	418422	1998/08/1	10:35:00	0.695	0.02	0.09	7.23	0.06
10	Closed	KILFENNAI	<10km2	Foyle (witl	246079	418422	1998/09/1	09:30:00	1.73	0.03	0.08	7.39	0.08
11	Closed	KILFENNAI	<10km2	Foyle (witl	246079	418422	1998/10/1	10:55:00	2.3	0.02	0.07	7.65	0.09
12	Closed	KILFENNAI	<10km2	Foyle (witl	246079	418422	1998/11/1	11:10:00	3.4	0.03	0.16	7.46	0.05
13	Closed	KILFENNAI	<10km2	Foyle (witl	246079	418422	1998/12/C	11:10:00	3.1	0.02	0.07	7.69	0.04
97	Closed	KILFENNAI	<10km2	Foyle (witl	245401	416700	1998/07/1	10:45:00	1.74	0.07	0.51	7.79	0.21
98	Closed	KILFENNAI	<10km2	Foyle (witl	245401	416700	1998/08/1	11:00:00	0.47	0.02	0.09	7.29	0.09
99	Closed	KILFENNAI	<10km2	Foyle (witl	245401	416700	1998/09/1	08:55:00	0.98	0.09	0.18	7.19	0.76
100	Closed	KILFENNAI	<10km2	Foyle (witl	245401	416700	1998/10/1	11:15:00	1	0.05	0.92	8.12	0.75
101	Closed	KILFENNAI	<10km2	Foyle (witl	245401	416700	1998/11/1	11:30:00	1.3	0.02	0.18	7.29	0.16
102	Closed	KILFENNAI	<10km2	Foyle (witl	245401	416700	1998/12/C	11:30:00	1.5	0.06	0.55	7.6	0.58
202	Closed	PENNYBUI	<10km2	Foyle (witl	243938	418730	1998/01/1	16:40:00	2.9	0.05	0.95	7.11	0.05
203	Closed	PENNYBUI	<10km2	Foyle (witl	243938	418730	1998/02/1	16:45:00	1.99	0.06	1.03	7.22	0.06
204	Closed	PENNYBUI	<10km2	Foyle (witl	243938	418730	1998/03/1	11:55:00	1.89	0.06	1.09	7.27	0.07
205	Closed	PENNYBUI	<10km2	Foyle (witl	243938	418730	1998/04/1	13:30:00	1.45	0.03	1.29	7.38	0.07
206	Closed	PENNYBUI	<10km2	Foyle (witl	243938	418730	1998/05/1	16:30:00	2.7	0.13	0.93	7.25	0.05
207	Closed	PENNYBUI	<10km2	Foyle (witl	243938	418730	1998/06/1	08:50:00	1.2	0.07	5.2	7.28	0.18
208	Closed	PENNYBUI	<10km2	Foyle (witl	243938	418730	1998/07/1	08:00:00	2.61	0.41	4.5	8.49	0.36
209	Closed	PENNYBUI	<10km2	Foyle (witl	243938	418730	1998/08/1	09:30:00	1.43	0.04	0.45	7.19	0.06
210	Closed	PENNYBUI	<10km2	Foyle (witl	243938	418730	1998/09/1	16:30:00	1.54	0.06	0.61	7.17	0.03
211	Closed	PENNYBUI	<10km2	Foyle (witl	243938	418730	1998/10/1	08:45:00	1.8	0.06	1.25	7.43	0.07

4.2 Parameters Selection

In the raw data, it provides 14 water chemistry parameters, for example, phosphorus concentration, ammonia concentration and some other water quality index, showed in Table 2 above. By choosing 4 available parameters (available means there is data for every choosing sample in the cluster), the response variables are set.

5. Data Analysis

5.1 The Relationship Between Land Use Patterns and Urbanization Degree

Although normally, population will be an essential index in evaluating the degree of urbanization, due to the cluster sampling method, which divides data due to different years, we can eliminate the effect of population on urbanization degree if we assume that population during each year has the same value. Therefore, we only need to find the relation between land use patterns and urbanization degree by a certain mechanism, which based on the function of the area, for instance, roads, parks, bridges and so on.

Then getting back to the area covered, land use patterns in this place can be mainly described by two factors, how land is used, and how land is developed. The reason why includes the second part is used to specialize the land used in the same way but in the different places.

5.2 The Relationship Between Urbanization Degree and Water Chemistry Parameters

5.2.1 Variable chose and basic process

Now we can get into the subject of essay, according to above analysis, I will provide some data set examples to show the relation between urbanization degree and water chemistry parameters. In analysing data, I choose concentration of $NO_3, SS(suspended\ solid), NH_4$ and phosphorus to test the relationship between urbanization degree and water quality, below are some examples from each cluster, the data than drew into the form of scatter plot, for showing the best-fit curve to represents the relations. For the chosen four variables, we need to analysis the correlation between each of them in order to assure the independence, fortunately, the correlation coefficient matrix has a maximum value of 0.00273, which means the variables are independent.

5.2.2 Example Cluster 1 Analysis: Foyle Basin in 1998

According to table 3 shows below, after determine the average for each index, we get average concentration of each chemical in the year, in milligram per litter. After rearranging test stations from lowest urbanization degree to highest, and by plotting it into a scatter plot, we can find the relation between urbanization degree and those indexes.

Table 3. Data collection around Foyle Basin in 1998

	NO3	SS	NH4	P
KILFENNAN BURN AT FOYLE BRIDGE	1.924167	2.9	0.085	0.06
KILFENNAN BURN AT MALONE PARK	1.165	9.1	0.405	0.425
PENNYBURN STREAM AT STRAND ROAD	1.9425	9.1	1.520833	0.093333
PENNYBURN STREAM AT NEW ROAD	1.768	12.7	0.075	0.02
CREGGAN BYWASH AT STRAND ROAD	1.228333	14.6	5.076667	0.716667
CREGGAN BYWASH AT GLENOWEN	2.168333	15.5	0.753333	0.113333
ST COLUMBS PARK STREAM AT LIMAVADY ROAD	2.277333	18.7	0.290833	0.050833
ST COLUMBS PARK STREAM AT TRENCH ROAD	4.094182	30.7	0.147273	0.037273
GLENMORNAN RIVER AT BALLYMAGORRY BRIDGE	2.26825	32.5	0.228446	0.149167

According to the research on Google Maps and previous discussion, we arrange these 9 places in series from the lowest urbanization degree to the highest; by using R summary and R plot, I get rid of the effect caused by outlier and influential point, the result is showed in figure 3, and related data is showed in table 4.

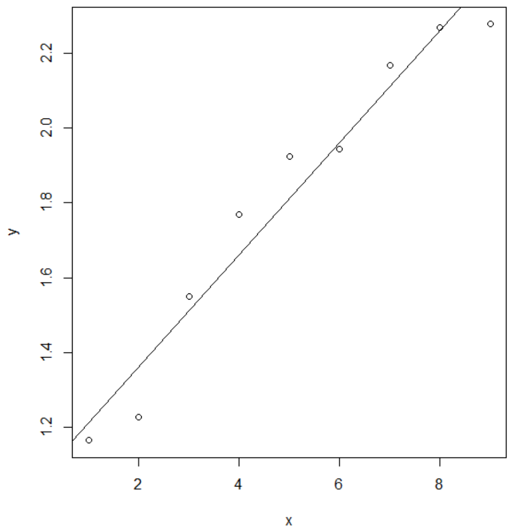

Figure 3. Relation between urbanization degree and NO3 concentration

Table 4. Regression model analysis: NO3

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.133161 -0.046511  0.008939  0.058589  0.113889

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.06186    0.07113   14.93 1.45e-06 ***
x            0.14965    0.01264   11.84 6.96e-06 ***
```

According to the plot and data analysis, we find that there is a positive linear relationship between land use patterns (from infrastructure use to agriculture), it is logical that land for agriculture will cause more leaking of nitrogen trioxide due to the use of fertilizer, and the most urbanized place is around central park, which surrounds by residential area. Moreover, according to the displayed p-value, if we conduct a hypothesis test for point estimator β, the very small of p-value shows we can reject the null hypothesis, indicating that the two variables truly have a linear relationship, in return we determine the relation between urbanization degree and nitrogen trioxide.

Similarly, for other three index, the regression models are built in the same way, showed in figure 4-6.

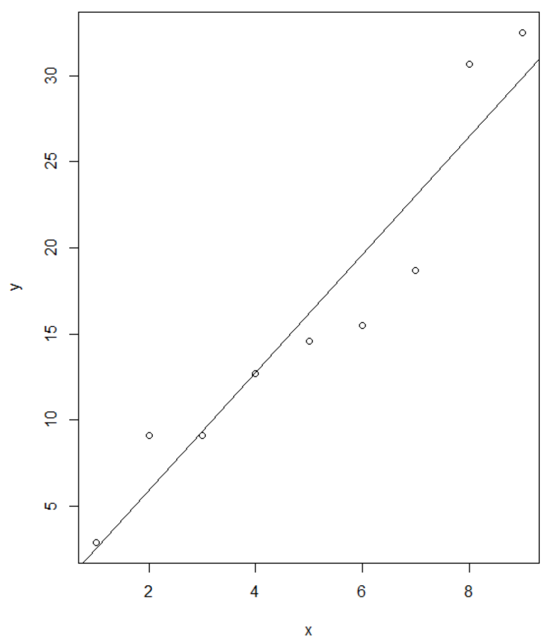

Figure 4. Relation between urbanization degree and SS concentration

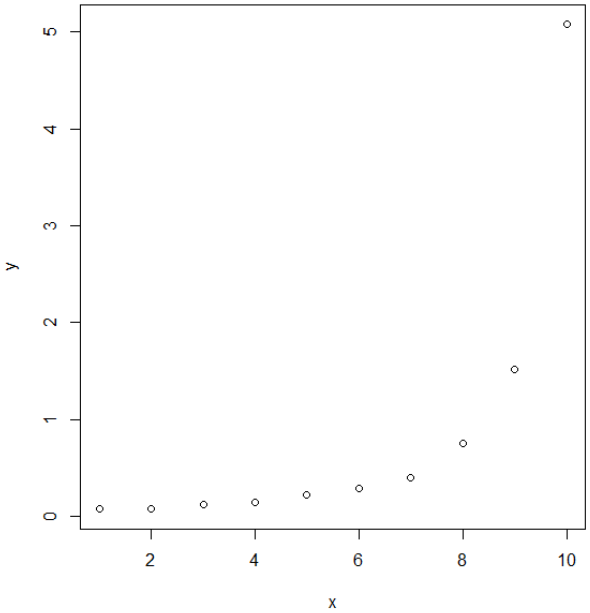

Figure 5. Relation between urbanization degree and NH4 concentration

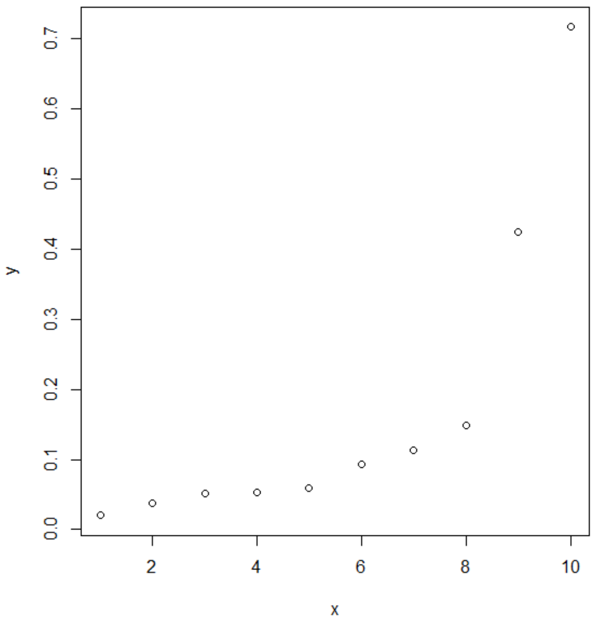

Figure 6. Relation between urbanization degree and P concentration

From these three plots, it is obvious that all three plots have a same pattern, by using quadratic function, data analysis are below in 5-7.

Table 5. Regression model analysis: SS

```
Residuals:
   Min     1Q Median     3Q    Max 
 -4.34  -1.60  -0.08   2.62   4.24 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -0.9000     2.3527  -0.383    0.713    
x             3.4200     0.4181   8.180 7.91e-05 ***
```

Table 6. Regression model analysis: NH4

```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.04937 -0.47825  0.07753  0.41618  1.25446 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  1.36215    0.91778   1.484   0.1813  
x           -0.87188    0.38330  -2.275   0.0571 .
I(x^2)       0.11179    0.03396   3.292   0.0133 *
```

Table 7. Regression model analysis: P

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.13907 -0.04477  0.01188  0.05244  0.09742 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept)  0.185067   0.095245   1.943   0.0931 . 
x           -0.109427   0.039778  -2.751   0.0285 * 
I(x^2)       0.015288   0.003524   4.338   0.0034 **
```

Therefore, from p-value, again we can say that two variables actually have the expected relationship, and therefore we conclude that nitrogen trioxide has a negative linear relationship to urbanization degree, by considering agriculture is the lowest degree of urbanization, and suspended solid has a positive linear relation to the degree of urbanization, while other two indexes experience an exponential as the increasing in the degree of urbanization.

5.2.3 Example Cluster 2 Analysis: Arney Basin in 2005

Similarly, we again determine the average value of each index for each test station in 2005, and use regression model to analysis each data, the following result showed below in figure 7-10, and by comparing to cluster 1, a more accurate relationship will be shown.

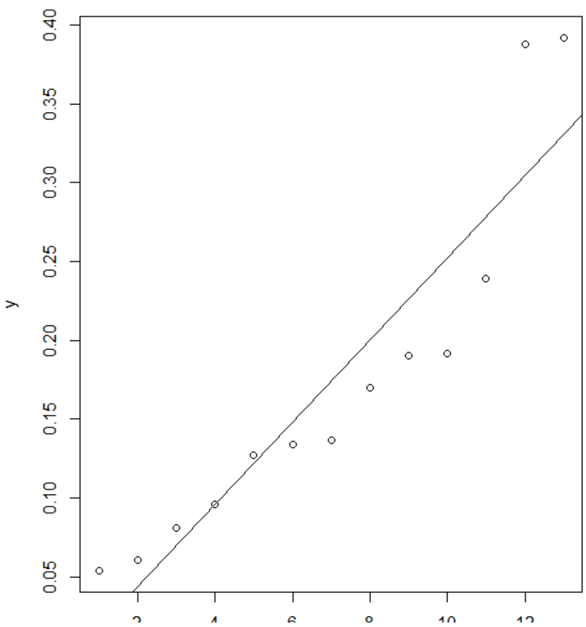

Figure 7. Relation between urbanization degree and NO3 concentration

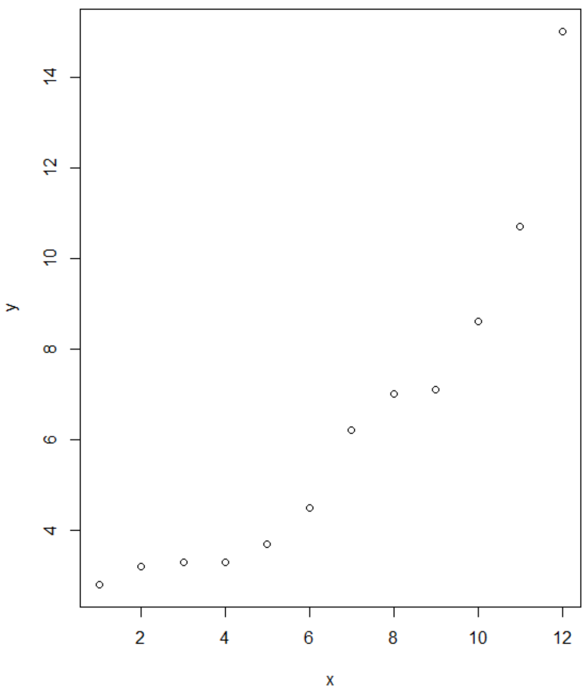

Figure 8. Relation between urbanization degree and SS concentration

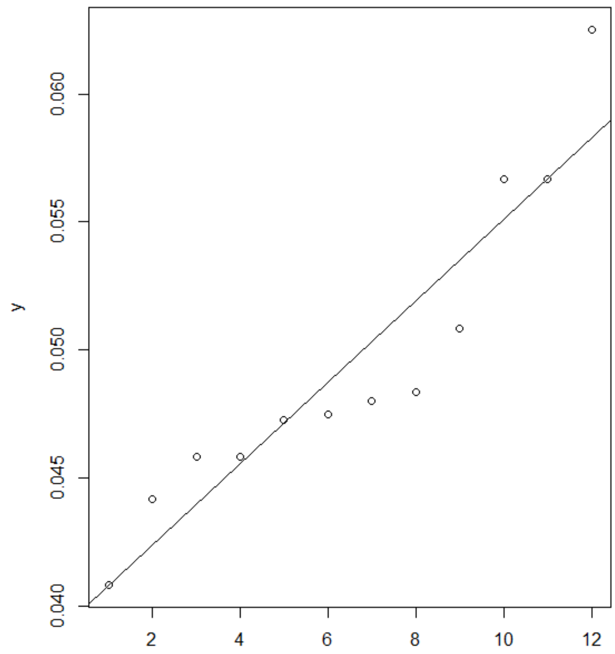

Figure 9. Relation between urbanization degree and NH4 concentration

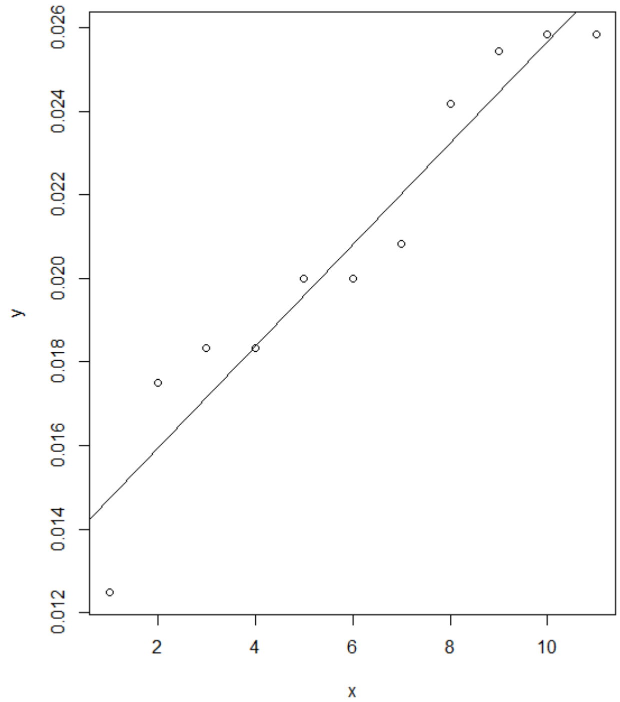

Figure 10. Relation between urbanization degree and P concentration

Related data analysis showed in table 8-11, according to very small p-value, as describes in example cluster 1, a hypothesis test for point estimator may be established, and the null hypothesis will be rejected, which indicates the true relationship between indexes and degree of urbanization.

Table 8. Regression model analysis: NO3

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.060310 -0.035938  0.000092  0.017338  0.083444

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.008417   0.025980  -0.324    0.752
x            0.026039   0.003273   7.955 6.89e-06 ***
```

Table 9. Regression model analysis: SS

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.01988 -0.56791  0.06164  0.31908  1.35275

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.77273    0.81007   4.657 0.001190 **
x           -0.60609    0.28650  -2.115 0.063502 .
I(x^2)       0.11908    0.02145   5.550 0.000356 ***
```

Table 10. Regression model analysis: NH4

```
Residuals:
       Min         1Q     Median         3Q        Max
-0.0035915 -0.0015136  0.0000889  0.0016169  0.0042067

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0391878  0.0014304  27.397 9.72e-11 ***
x           0.0015921  0.0001943   8.192 9.56e-06 ***
```

Table 11. Regression model analysis: P

```
Residuals:
       Min         1Q     Median         3Q        Max
-0.0022148 -0.0009225  0.0001692  0.0009650  0.0015686

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0134982  0.0008056   16.75 4.30e-08 ***
x           0.0012166  0.0001188   10.24 2.93e-06 ***
```

5.3 Fundamental Conclusion

Since the huge amount of raw data, this passage only provides two examples to show the analysis process, they both shows a positive relationship between chosen parameters and urbanization degree, although some of clusters does not show the same patterns, specifically in the difference of linear or exponential, this difference can be ignored in macro view of relation; and by specifically analysing the terrain statistics, this difference can be explained in micro view due to the different in land, water or other factors that affect the deposit or liquidity of those chemicals, for example, the reason suspended solid in second example experience an exponential growth in first example attribute to the special hydrological characteristics, they provides a perfect condition for deposit nitrogen trioxide, resulting in accelerated accumulation, thus showed an exponential growth.

6. Popularization

Since the observational study provides clear and accurate data to show the relation between urbanization degree and the four indexes, and the scope of inference is larger enough, we can use the result to go backward to quantify the urbanization degree rather than just "say" the degree of urbanization by only qualitative analysing maps or predicting from population.

According to the result, the most intuitionistic method to quantify the degree of urbanization is to directly build a linear system to give a score to a specific land in order to tell its degree of urbanization numerically. Therefore, by determine the possible sources of these four deposits, to illustrate, nitrogen trioxide and phosphorus usually comes from farms because of the use of fertilizer, which can be addressed to agriculture, while ammonium and suspended solid mainly caused by the discharge of sewage or leaking of harmful chemical from factories, which can be addressed to urban. Thus, by evaluate the proportion of each chemical in the area, we can determine the degree of urbanization.

Set, concentration of $NO_3 = w, SS = x, NH_4 = y, P = z$, a linear system will be conducted as: $\begin{matrix} aw+bx+cy+dz=T \\ ew+fx+gy+hz=S \end{matrix}$, where T and S are objective unknown numbers, more, $\begin{matrix} a+b+c+d=1 \\ e+f+g+h=1 \end{matrix}$ since they are weight for each indicator which shows the sources of each chemical, therefore given $\begin{matrix} a+d>>b+c \\ e+h<<f+g \end{matrix}$, to indicate the influence of each indicator on the score of the land, according to the discussion, the linear system shows below,

$$\begin{pmatrix} a & b & c & d & -1 & 0 \\ e & f & g & h & 0 & -1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} w \\ x \\ y \\ z \\ T \\ S \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ w(A) \\ x(A) \\ y(A) \\ z(A) \end{pmatrix}$$

where $w(A), x(A), y(A), z(A)$ are function of different places, hence by solving true unknown numbers T and S, we can get the score for a certain land, thus determine the urbanization degree.

7. Conclusion and Discussion

According to the previous discussion, this passage finds a relation between four water chemistry parameters and urbanization degree, which are usually to be positive relationships if setting degree of urbanization from lowest to highest, although nitrogen trioxide may exist an abnormal behaviour as other three parameters, it is completely logical due to the use of fertilizer (at least in this area). Thus, by taking backward track, this passage uses a new way to indicate the urbanization degree, which is form a scoring system by using the linear combination of these four indicators, consequently set a standard to supervise the excess urbanization, which is harmful to the environment.

However, for this observational study, the data analysis is not as accurate as expected since the huge amount of raw data, the interpretation of data and calculation become a hard problem, by solving this, this passage takes average value in each cluster, which cause the decreasing in sample, and this may result in inaccuracy in regression model and popularization.

In further studies, if so, the analysis of data should be more accuracy, particularly when filtering the valid data and processing the data into a clear and accurate form; and this can be done by performing more test to see the relation or rationality of the data.

Reference List

[1] Guan, B.; An, S.; Gu, B. Assessment of ecosystem health during the past 40 years for Lake Taihu in the Yangtze River Delta, China. Limnology 2011, 12, 47–53.

[2] Deng, X.; Xu, Y.; Han, L.; Song, S.; Yang, L.; Li, G.; Wang, Y. Impacts of Urbanization on River Systems in the Taihu Region, China. Water 2015, 7, 1340-1358. https://doi.org/10.3390/w7041340

[3] Zhao, G., Gao, H., & Cuo, L. (2016). Effects of Urbanization and Climate Change on Peak Flows over the San Antonio River Basin, Texas, Journal of Hydrometeorology, 17(9), 2371-2389. https://doi.org/10.1175/JHM-D-15-0216.1

[4] Benjamin, M. M. (2014). Water chemistry. Waveland Press.

[5] River Water Quality Monitoring 1990 to 2018

https://www.data.gov.uk/dataset/a18547a7-bc13-49d9-8be1-59487b61e000/river-water-quality-monitoring-1990-to-2018