Cluster Analysis and Principal Component Analysis using Micropollutant Measurements on the Most Polluted Tributary of Ergene River: Çorlu Stream

CINGIROGLU F.[1*], EMADIAN S.M.[2], TEZEL U.[2] and KAYNAK B.[3]

[1] Eurasia Institute of Earth Sciences, Istanbul Technical University, Istanbul, Turkey

[2] Institute of Environmental Sciences, Bogazici University, Istanbul, Turkey

[3] School of Civil Engineering, Department of Environmental Engineering, Istanbul Technical University, Istanbul, Turkey

[*] corresponding author

e-mail: cingiroglu@itu.edu.tr

Abstract. This study focuses on investigation the spatial distribution of pollutants in Çorlu Stream is the highest industrialized tributary of Ergene River which is one of the most polluted rivers in Turkey. A total of 250 conventional, metal and micropollutants were scanned at eleven sampling locations for four seasons in Çorlu Stream of Ergene River. At these locations, total of 126, 124, 99 and 107 pollutants were detected at least one sampling location in summer, fall, winter and spring, respectively. Four micropollutants which are associated with textile industry were detected in every location in all seasons. Cluster Analysis found four main clusters for sampling locations from most polluted to low polluted, and six cluster for pollutants ranging from conventional to unique pollutants. Principal Component Analysis identified six components for every season explaining more than 80% of the variation in each. PCA results indicated the impact of textile, leather, metal industries and appliance-electronics production in the Çorlu Region.

Keywords: cluster analysis, principal component analysis, micropollutants, Ergene River, Çorlu Stream

1. Introduction

Rivers as the most important freshwater supplies are being contaminated by industrial and domestic activities. Multivariate statistical analysis and source apportionment methods such as Cluster Analysis (CA), Factor Analysis (FA), Principal Component Analysis (PCA) and Absolute Principal Component Scores (APCS) are used for defining pollution sources discharging to rivers in several studies.

There are several implementations of these methods on river systems. One of the most comprehensive CA studies in the Hudson River identified three cluster as core, sewage treatment plant, and diffuse micropollutants using up to 200 micropollutants at seventeen sampling locations (Carpenter and Helbling 2018). CA, PCA, and APCS used 23 variables including metals at 43 sampling locations in Elbe River identified three main clusters with CA, and three major components with PCA as sediment resuspension, water discharge, and mining activities (Baborowski et al. 2012). FA and CA performed for groundwater samples with 16 parameters in Tanzania categorized five clusters and three sources of mixed origin of human wastes and soil in the runoff, dual origin of turbidity (human wastes and soil/organic matter), and natural/geochemical processes in aquifers (Basamba et al. 2013). PCA/FA and APCS analysis of long-term sampling from the Kuantan River with thirteen parameters indicated five factors including fertilizer waste, surface runoff, anthropogenic input, chemical, and mineral changes, and erosion (Nasir et al. 2012). CA, PCA/FA, and APCS were used for 23 water quality parameters including metals collected at 26 sampling locations along Soan River (Nazeer et al. 2016). China; PCA and APCS analyzed samples from 20 sampling sites for 17 water quality variables from Jinjiang River (Chen et al. 2013), FA/PCA and CA analyzed samples from 53 sites measured of thirteen trace elements from Huaihe River (Wang et al. 2017). In Turkey, a study focusing on Çoruh River found two clusters as high heavy metal pollution, and mixed domestic-agricultural wastewater; four factors as mining activities, geochemical structure, agriculture, domestic wastewater from measured 21 water quality parameters and heavy metals from 14 monitoring sites using CA, FA, and PCA (Bilgin and Konanç 2016).

In our study, wide number of micropollutants along with metal and conventional pollutants for four seasons were used in CA and PCA to understand the impacting sources on Çorlu Stream (Ergene River).

2. Methodology

Study Area. Çorlu Stream is located at the beginning of Ergene River; and its watershed has significant number of industrial facilities concentrated in the Çorlu-Çerkezköy region. The sampling locations (*L09-L19*) on Çorlu Stream along with industrial and domestic pollution sources are given in Figure 1. Six Organized Industrial Districts (OID) within the sub-watershed of Çorlu Stream with possibly affected sampling locations are as follows: Çerkezköy (*L09*), Kapaklı (*L09*), Yalıboyu (*L11*), Veliköy

(*L13*), Velimeşe (*L12, L14, L15*), and Çorlu Leather (L17). The textile sector takes the first place in the sectoral distribution of the industries, followed by chemical, leather, metal, mining, food, and glass in Çorlu-Çerkezköy region. Çorlu has an OID solely focusing on leather processing. In addition, Çerkezköy settlement has the highest population density in the region (*L09, L10*), followed by Çorlu (*L16*).

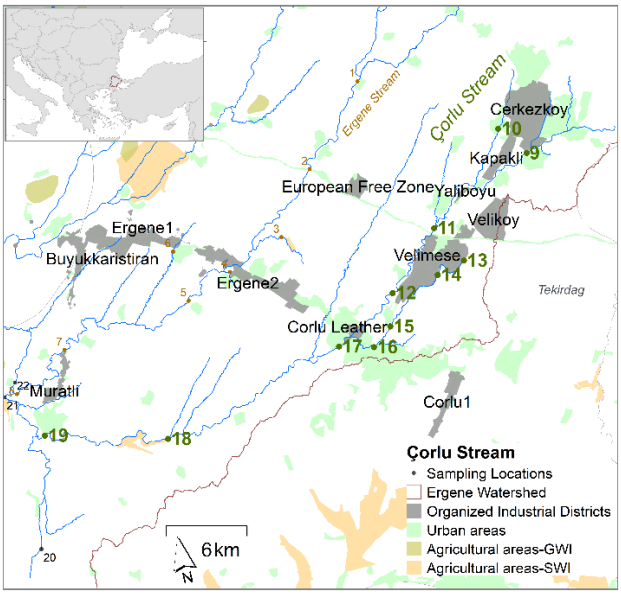

Figure 1. The sampling locations on Çorlu Stream with possible pollution sources and land use around the region.

GIS Analysis. The spatial information of the river, watershed, urban settlements, agricultural areas, location of industrial facilities, and OIDs was processed using ArcGIS. European Digital Elevation Model (EU-DEM 2016) and CORINE Land Cover (CORINE 2018) data was used. OID information were compiled in terms of industrial activity (OIZSO 2018).

Measurements. 75 sampling locations were selected for Ergene River based on intensity, type and spatial distribution of the pollution sources in the sub-watersheds, and 11 of them were on Çorlu Stream. The samples were scanned for 250 pollutants (222 organic micropollutants) for four seasons (August 2017, November 2017, February 2018, and May 2018). Measurements of conventional pollutants, metals and micropollutants such as endocrine disrupters, personal care products, industrial chemicals, pesticides, herbicides, fungicides, hormones, pharmaceuticals, antibiotics, fertilizers, and other highly toxic chemicals such as polychlorinated biphenyls (PCBs) and polycyclic aromatic hydrocarbons (PAHs) were performed (Emadian et al. 2021).

Multivariate Statistical Analysis. Cluster and Principal Component Analysis methods were applied using R software. Ward linkage method with normalized concentration values (0-1, quotient transformation: x/x_{max}) was used in CA to emphasize the impact of relative changes in concentrations instead of magnitudes that changes from chemical to chemical. CA analysis was performed for all sampling locations, all seasons together. Main clusters with respect to sampling locations (*C#-L*) and pollutant species (*C#-P*) were determined.

PCA was applied initially for all sampling locations, all seasons together. However, results showed PCs with low contributions (maximum PC-1 with 13%) indicating seasonal changes in the measurements. Thus, PCA was applied for each season separately to better understand the seasonal PC contributions using detected pollutants at least one measurement above LOD along with color and toxicity measurements.

3. Results and Discussion

Çorlu Stream, usually having the highest pollutant levels, was selected for this study, and the multivariate statistical analyses were performed. Six OIDs located in the region play a substantial role on the pollution levels of Çorlu Stream. *L13* which is the still water body at the start of one of the tributaries can be considered as background pollution with pH, temperature and color measurements being low in all seasons. The values increase significantly following *L13* at locations of *L14, L15, L16*. These measurements indicated a possible significant discharge. Temperature, pH and color measurements were clearly high in locations of *L14, L15, L16* in all seasons. For instance, *L13* temperature of 14.4°C was increased to 29°C at *L14* in fall season; *L13* pH of 7.3 increased to 9 in *L14* in winter season. In winter, with higher rainfall, flowrates reached maximums of 21 m^3/s at *L12*, 18 m^3/s at *L17*, and only to 2 m^3/s at *L14, L15* and *L16*.

Out of 250 scanned pollutants, 161 pollutants (132 micropollutants) were detected at least one location in at least one season. The most commonly detected micropollutants were Benzyldimethyldodecylammonium, Benzyldimethyltetradecylammonium, N,N-Diethyl-m-toluamide, and Hexa-(methoxymethyl)-melamine in every location in all seasons. These chemicals are mainly associated with industrial textile processes. Summer samples have 126 at least one, 26 pollutants (12 micro) in every location; fall samples have 124 at least one, 28 pollutants (12 micro) in every location; winter samples have 99 at least one, 23 pollutants (8 micro) in every location; spring samples have 107 at least one, 32 pollutants (15 micro) in every location.

CA was carried out with 44 sets sampled for 4 seasons at 11 sampling locations with 164 pollutants identified. A heatmap was created to show clustering of both sampling locations and pollutants together (Figure 2).

CA for **sampling locations** indicated two main clusters that were divided into two main subsets. Locations in the most contaminated subset that is *C1-L* are *L14 spr, L15 spr, L15 fall, L16 fall, L16 sum, L14 sum, L15 sum*. Highly contaminated subset that is *C2-L* has mostly summer and fall locations. Moderate contaminated subset that is *C3-L* has only winter locations with other *L13* locations. Because of larger flowrate in winter, pollutant concentrations were decreasing. Nevertheless, *winter L14, L15* and *L16* clustered together. Low contaminated subset that is *C4-L* includes only spring locations.

CA for **pollutants** divided two main and six subset clusters. *C1-P* has the most frequently detected pollutants throughout Çorlu Stream, followed by *C2-P* mainly detected pollutants in the most contaminated location clusters (*C1-L* and *C2-L*). The red bands in the heatmap

clearly shows the unique location-based pollutants in *C6-P* (Figure 2). Specifically, *L13 win* showed metal pollutants; *L09 fall* showed pollutants associated with personal care products (PCPs) and antibiotics, *L17 sum* showed pesticides and sunscreen agents, *L10 sum* showed pesticides and antibiotics, and *L13 spr* showed pesticide and fungicides.

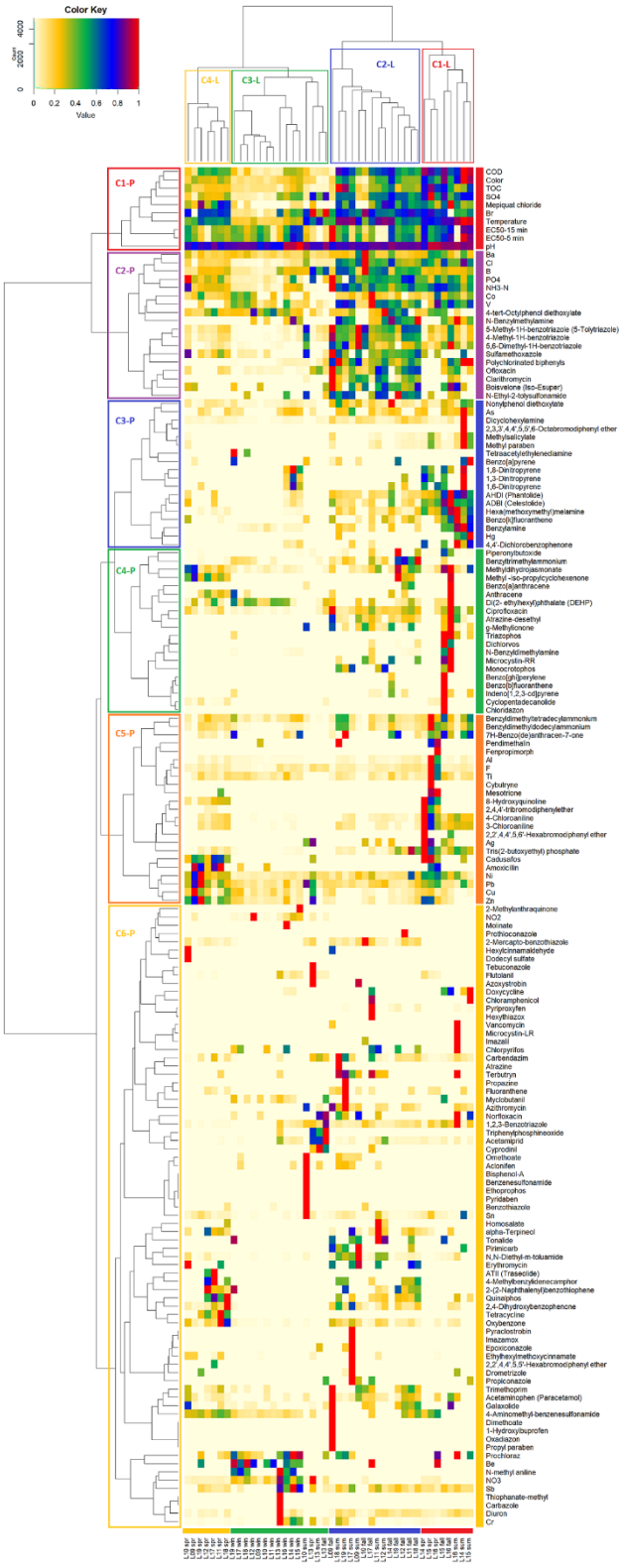

Figure 2. CA results for Çorlu Stream

C1-P showed conventional pollutants along with color, temperature, pH and toxicity parameters, thus indicating the overall pollution. *C2-P* summarizes metal industry and corrosion inhibitors, PCPs and pharmaceuticals. *C3-P* is composed of mixed use chemicals along with PAHs, and PCPs. *C4-P* showed mixed manufacture of plastic and chemicals, PCPs and PAHs. *C5-P* summarizes surfactants, pesticides, metal industry chemicals and flame retardants, and *C6-P* is composed of pesticides, insecticides, PCPs and metal industry chemicals.

PCA was performed initially for the same 44 sets. The six components with an individual variance greater than 5% were found. However, these first six components explained only 47% and the first seven components explained only 50% variance. Pollution levels could not be explained with reasonable number of components due to the seasonal change of pollution sources. Therefore, PCA was performed separately for each season. Individual results showed the total variance was explained by more than 70% with first five, 80% with first six, and 88% with first seven PCs for all four seasons. This indicated that the most important pollution sources vary from season to season. The first six PCs for each season are given in Figure 3. The contributions of sampling locations usually indicate *L14*, *L15* and *L16* on *PC-1* indicating the more polluted region (Figure 3). Other PCs such as *PC-2* in winter and spring, and *PC-3* in winter usually showed one or two locations with very high contributions, thus sources associated with specific discharge locations. *PC-2* in winter has approximately 60% contribution from the last location, *L19* indicating a possible discharge from close-by urban settlement.

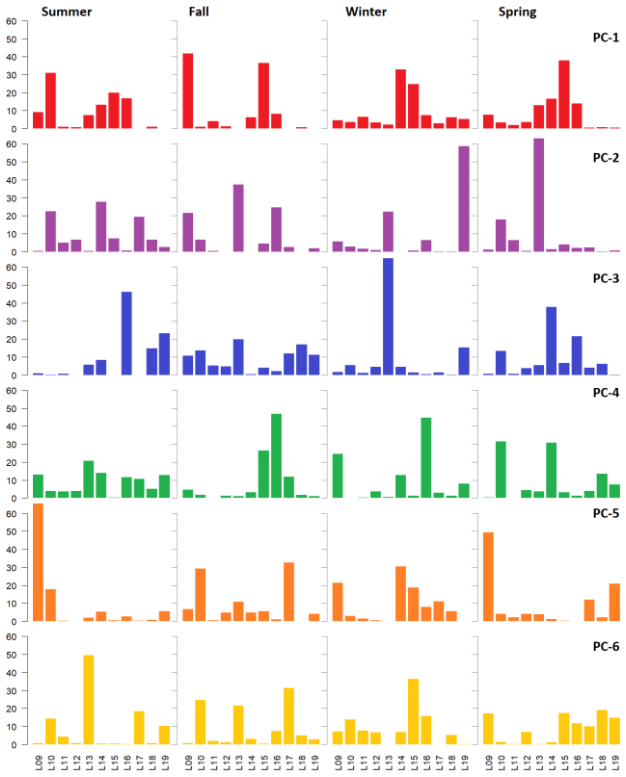

Figure 3. PCA results for four seasons for Çorlu Stream

The region has important sources of textile and leather industries along with large appliance production facilities. The PCA findings were summarized with important pollutants and mostly affected sampling locations (contributions >%20) and possible sources (Table 1). The impact of textile industry was observed in various PCs in

all seasons along with metal industry. The pesticides were also commonly found in various PCs.

Table 1. Summary of seasonal PCA indicating the highest contributing sampling locations (contribution %) and pollutants with possible pollution source.

	Summer	Fall	Winter	Spring
	L10 (31%) *L15 (20%)*	*L09 (42%)* *L15 (37%)*	*L14 (33%)* *L15 (25%)*	*L15 (38%)*
	3-Chloroaniline Color ADBI (Celestolide)	Trimethoprim COD Benzyldimethyldo-decylammonium	SO_4 Color Sb	TOC Cl COD
PC-1	**textile**	**textile antibiotics**	**textile metal**	**textile conventional**
	L14 (28%) *L10 (22%)*	*L13 (37%)* *L16 (25%)* *L09 (22%)*	*L19 (59%)* *L13 (22%)*	*L13 (63%)*
	Oxybenzone Monocrotophos Ofloxacin	EC50-5 min EC50-15 min Benzo[k]-fluoranthene	Benzyldimethyldode-cylammonium Anthracene Oxybenzone	B Cr Flutolanil
PC-2	**PCPs pesticides**	**toxics**	**PAHs dyes**	**textile metal**
	L16 (46%) *L19 (23%)*	*L13 (20%)*	*L13 (65%)*	*L14 (38%)* *L16 (21%)*
	1,2,3-Benzotriazole Al Vancomycin	Al Quinalphos Benzyltrimethyl-ammonium	Carbendazim Carbazole Thiophanate-methyl	As Dicyclohexylamine 2,2',4,4',5,6'-Hexa-bromodiphenyl ether
PC-3	**appliances electronics**	**metal textile**	**textile pesticides**	**mixed industrial**
	L13 (21%)	*L16 (47%)* *L15 (26%)*	*L16 (45%)* *L09 (24%)*	*L10 (32%)* *L14 (31%)*
	5,6-Dimethyl-1H-benzotriazole F Piperonylbutoxide	Di(2-ethylhexyl)-phthalate (DEHP) 2,2',4,4',5,6'-Hexabromodiphenyl ether Benzo[a]anthracene	Ti Molinate N-Ethyl-2-toly sulfonamide	PO_4 Fluoranthene Dodecyl sulfate
PC-4	**pharmaceutical**	**PAHs plastics**	**appliances electronics**	**domestic**
	L09 (65%)	*L17 (32%)* *L10 (29%)*	*L14 (30%)* *L09 (21%)*	*L09 (49%)* *L19 (21%)*
	Trimethoprim Azoxystrobin Di(2-ethylhexyl)-phthalate (DEHP)	Cr Propiconazole Pb	Aclonifen Piperonylbutoxide N-methyl aniline	Ni Pb Benzo[a]pyrene
PC-5	**PCPs antibiotics**	**leather**	**pesticides**	**metal**
	L13 (50%)	*L17 (31%)* *L10 (25%)* *L13 (22%)*	*L15 (36%)*	*L18 (19%)*
	Cyprodinil Triphenyl phosphineoxide Flutolanil	Ba F Cl	2-Methylanthra quinone 4-Aminomethyl-benzenesulfonamide Methyldihydro-jasmonate	Benzyltrimethyl-ammonium Cyclopentade-canolide Prochloraz
PC-6	**pesticides metal**	**appliances electronics**	**PCPs pharmaceutical**	**leather metal**

4. Conclusion

CA was performed using Ward linkage method with pollutants and sampling locations for all seasons with normalized concentrations. CA for sampling locations indicated four clusters; first and second are polluted locations, third one is moderately polluted and last has relatively clean locations for all seasons. Six clusters were categorized according to pollutants. First cluster contains generally conventional pollutants that are detected almost all locations.

PCA was performed to each season explaining more than 80% cumulative variance with six PCs. The impacting industries were textile, leather, metal, appliance production along with various chemicals as pesticides, antibiotics and PAHs. In some of the PCs the domestic discharge impact can be seen by a wide variety of PCP chemicals.

Acknowledgements. This research is supported by TUBITAK under the grant number of 115Y064.

References

Baborowski M., Simeonov V., Einax J.W. (2012), Assessment of Water Quality in the Elbe River at Flood Water Conditions Based on Cluster Analysis, Principle Components Analysis, and Source Apportionment. *Clean - Soil, Air, Water,* **40**, 373–380.

Basamba T.A., Sekabira K., Mary Kayombo C., Ssegawa P. (2013), Application of factor and cluster analyses in the assessment of sources of contaminants in borehole water in Tanzania. *Polish J Environ Stud*, **22**, 337–346.

Bilgin A, Konanç M.U. (2016), Evaluation of surface water quality and heavy metal pollution of Coruh River Basin (Turkey) by multivariate statistical methods. *Environ Earth Sci*, **75**.

Carpenter C.M.G, Helbling D.E. (2018), Widespread Micropollutant Monitoring in the Hudson River Estuary Reveals Spatiotemporal Micropollutant Clusters and Their Sources. *Environ Sci Technol*, **52**, 6187–6196.

Chen H., Teng Y., Wang J., et al (2013), Source apportionment of sediment PAHs in the Pearl River Delta region (China) using nonnegative matrix factorization analysis with effective weighted variance solution. *Sci Total Environ*, **444**, 401–408.

CORINE (2018), CORINE Land Cover (CLC) 2018, version 20. *https://land.copernicus.eu/pan-european/corine-land-cover/clc2018*.

Emadian S.M., Sefiloglu F.O., Akmehmet Balcioglu I., Tezel U. (2021), Identification of core micropollutants of Ergene River and their categorization based on spatiotemporal distribution. *Sci Total Environ*, **758**.

EU-DEM (2016), European Digital Elevation Model (EU-DEM), version 1.1. *https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1*.

Nasir M.F.M., Zali M.A., Juahir H., et al (2012), Application of receptor models on water quality data in source apportionment in Kuantan River Basin. *Iran J Environ Heal Sci Eng*, **9**.

Nazeer S., Ali Z., Malik R.N. (2016), Water Quality Assessment of River Soan (Pakistan) and Source Apportionment of Pollution Sources Through Receptor Modeling. *Arch Environ Contam Toxicol*, **71,** 97–112.

OIZSO (2018), Organized Industrial Zones Supreme Organization OID Inventory. *https://osbuk.org/osb-envanteri/*.

Wang J., Liu G., Liu H., Lam P.K.S. (2017), Multivariate statistical evaluation of dissolved trace elements and a water quality assessment in the middle reaches of Huaihe River, Anhui, China. *Sci Total Environ*, **583**, 421–431.