Modeling District-wise Spatial Autocorrelation of Mouth Cancer incidence in Tamil Nadu

P.Sampath[1], P.R.Jayashree[2], R. Srinivasan[3], R.Swaminathan[4]

[1]*Statistical Assistant, Division of Epidemiology and Cancer Registry, Cancer Institute (WIA), Adyar, Chennai-20*

[2]*Assistant professor, Department of Statistics, Presidency College, Chepauk, Chennai-05*

[3]*Senior Technical Assistant, National Institute for Research in Tuberculosis, ICMR, Chetpet, Chennai-31*

[4]*Professor and Head, Division of Epidemiology and Cancer Registry, Cancer Institute (WIA), Adyar, Chennai-20*

Abstract

Background:

The Spatial analysis of mouth cancer study is essential since the spatial pattern is unidentified for different districts in Tamil Nadu. The spatial classic model, spatial error model and spatial lag models are examined in this study for the mouth cancer data along with its associated factors.

Methods:

The spatial correlation of mouth cancer incidence was examined using the spatial autocorrelation method and the factors associated with it were also analyzed using a spatial regression model. A comparison was then made for the prediction of mouth cancer incidence between classical model, spatial error model, and spatial lag model using the GEODA Software.

Results:

The Local spatial LISA-significant map identifies statistically significant (95% confidence level; $p < 0.05$) spatial clusters of districts with high or low mouth cancer cases. The Clusters of districts with high mouth cancer incidence high-high(HH) are considered as “hotspots” were also identified from the spatial map.

Conclusion:

This study explained the spatial patterns and spatial clusters of the mouth cancer incidence for all districts of Tamil Nadu. The study also investigates the factors associated with higher cases of mouth cancer incidence in the districts of Tamil Nadu by finding spatial correlation of disease using bivariate OLS regression and the spatial regression model.

Keywords: Clusters, Hotspots, LISA, **Mouth cancer incidence,** OLS bivariate regression,
Spatial autocorrelation and regression Model

INTRODUCTION

Cancer is the third most common cause of death in India during the period 2010-2013[4] reported by vital statistical division, census of India. According to the International Agency for Research on Cancer (IARC), a cancer research wing of the World Health Organization (WHO) approximates 18 million individuals affected by cancer globally in 2018[8]. The Age-Standardized Rate (ASR) is197.9 per one lakh population, India shares 6.8% of the world cancer incidence and Tamil Nadu shares 5.3% to the India's cancer burden. However Cancer incidence shows a rapid increase unevenly, hence the study spatial knowledge is essential at present. Cancer registries are the "nuclei" for cancer control aimed at reducing incidence and mortality through record-based surveillance [8,20,21,23,24]. Among all cancers, the mouth cancer incidence was predominantly increasing over the decade in Tamil Nadu and so the spatial modeling is studied in this paper. The Tamil Nadu Cancer Registry Project (TNRCP), a population-based cancer registry which reported mouth cancer data during 2014, was utilized in this study to explore the spatial pattern and summarize the impact of mouth cancer in all districts in Tamil Nadu, India. The distribution of mouth cancer is invariably high in many districts of Tamil Nadu. The TNCRP registered a total of 61,402 cancer cases during the year 2014. Among those, 27,008 (44.0%) were male and 34,394 (56.0%) were female. Mouth cancer is the third most common cancer, accounting for 7.6% of the total cancers in males and 3.9% in females. So far, no study has explored the geographic distribution of mouth cancer in Tamil Nadu.

Spatial studies have proven a valuable etiological hypothesis in cancer research [3]. The spatial autocorrelation method is an epidemiological research method used to explore the dependence of the variables at one geographic location on variables in nearby locations [18,19]. It is a unique modeling technique in that it violates the assumption of independence but still reveals the scale over which spatial patterns occur [18,19]. A variety of studies have been conducted using spatial autocorrelation to identify the spatial pattern on specified cancer types using regression models[9].Rosenberg et al. used this method to study the cancer mortality distribution in Western Europe[18]. La Vecchis, C., Decrali, A. on their study found that cancer mortality are highly correlates with spatial and concluded that the spatial autocorrelation technique is useful for exploring maps[11]. The study by Yomralioglu T et al., have shown that there is a strong relationship between the health status and location where people are living [25]. Mandal et.al studied the correlation between female breast cancer and male prostate cancer in the United

States region during 2000 and 2005, utilizing ordinary least square regression (OLS) and geographically weighted regression (GWR) analysis [12]. They were able to find that there is a risk factor associate between these two cancers.

This study will project the spatial pattern of mouth cancer incidence by exploring disease clusters, using spatial autocorrelation, in all districts of Tamil Nadu. In this study Moran's I Index technique is performed and assess the spatial correlation of mouth cancer incidence with district level using the classical regression model. The aim of this study is to explore the spatial autocorrelation of mouth cancer incidence in Tamil Nadu and to compare the OLS regression model with spatial Lag and Spatial Error models.

2. Modeling Spatial Relationships

In the spatial regression estimation, a model is first estimated without incorporating spatial effects, but the results of this estimation (and its residuals) form the starting point for the diagnostics of spatial effects. The diagnostics model will used to distinguish between substantive (lag) and nuisance (error) spatialautocorrelation. For prediction, the regression models are often restricted to the interpretation of the significance and magnitude of the coefficients. In a GIS environment however, the results of spatial regression may also be used to predict values at locations.

2.1 *Ordinary Least Square estimation (OLS)*

From an estimation point of view, the problem with an OLS model specification when spatial autocorrelation is present is that the spatial lag term contains the dependent variables for neighboring observations, which in turn contain the spatial lag for their neighbors, and so on, leading to simultaneity. This simultaneity results in a nonzero correlation between the spatial lag and the error term, which violates a standard regression assumption.

Consequently, ordinary least squares (OLS) estimation will yield inconsistent (and biased) estimates, and inference based on this method will be flawed. Instead of OLS, specialized estimation methods must be employed that properly account for the spatial simultaneity in the model. These methods are either based on the maximum likelihood (ML) principle, or on the application of instrumental variable (IV) estimation in a spatial two-stage, least-squares approach. The most important aspect of spatial modeling will be specification testing. To find out the spatial interaction of spatial lag or spatial error dependence is the model specification, ignoring lag dependence it results in biased and inconsistent estimates for all the coefficients in the model.

2.2. *Spatial Error Model*

The spatial error model, evaluates the extent to which the clustering of an outcome variable not explained by measured independent variables can be accounted for with reference to the clustering of error terms. In this sense, it captures the influence of unmeasured independent variables

The spatial error model described by

$$y = X\beta + \varepsilon$$

$$\varepsilon = \lambda W\varepsilon + u$$

Where

Y is a $N \times 1$ vector of observations on the dependent variable,

X is an $N \times K$ matrix of observations on the explanatory variables,

β is a $K \times 1$ vector of regression coefficients,

ε in an $N \times 1$ vector of spatially auto correlated error terms,

Wε is a spatial lag for the errors,

λ(lambda) is the autoregressive coefficient, and u is another error term

2.2 *Spatial Lag Model*

In this model, incorporates the influence of unmeasured independent variables but also stipulates an additional effect of neighboring attribute values, i.e., the lagged dependent variable.

The spatial lag model will be

$$y = \rho Wy + X\beta + \varepsilon$$

where,

Wy is an $N \times 1$vector of spatial lags for the dependent variables,

ρ(Rho) is spatial autoregressive coefficient,

$X\beta$ is an $N \times K$ matrix of observations on the exogenous explanatory variables multiplied by a $K \times 1$ vector of regression coefficients β for each X,

εis a $N \times 1$ vector of normally distributed random error terms.

Results and Discussions:

In Classical model there are six tests were performed to assess the spatial dependence of mouth cancer incidence in Tamil Nadu. The Moran's I score is 0.3956 which is highly significant. It clearly indicates that there is strong spatial autocorrelation of mouth cancer incidence in all districts during the period 2014. Here literacy is highly significance with probability value 0.00338 but consuming tobacco is not significant. The rest of the function reports the tests chosen among five statistics for testing for spatial dependence in linear models. We can see both simple tests of the lag and error are significant, indicating presence of spatial dependence. The robust tests help us to understand the spatial dependency nature. The robust measure for error is significant, but the robust lag test is not significant, which means that when lagged dependent variable is present the error dependence disappears.

A classic spatial regression model was compared with a spatial error and spatial lag model. We found that the spatial error model gives the lowest Akaike info criterion value (AIC 395.366) when compared with the other two models. The classic model assumes that mouth cancers were independent and identically distributed (i.i.d) but disease occurrence depends on spatial location due to several factors like environmental, socio-economical and other risk factors.

Spatial regression models like the spatial error model and spatial lag model is suitable for estimates of the mouth cancer data. Spatial error model assumes that the uncorrelated error term is violated and suitable for and indicative of omitted covariates. If left unattended, it would affect inference. Spatial lag is suggestive of possible processes; events in one place predict an increased likelihood of similar events in neighboring places. AIC value is almost similar in all the models, but spatial error models give lowest value that indicates that accounts the uncorrelated error between mouth cancers among in all the districts.

The cut-off distance at which the overall degree of clustering was maximized was used to calculate the Anselin local Moran's *I*. In Tamil Nadu, three districts are identified as hotspot districts viz.,Chennai, Thiruvallur and Kanchipuram and those are significantly associated. However, Cuddalore district was surrounded by low risk district and the remaining districts come under cold spot districts within Tamil Nadu.

Figure 1: Global Spatial Autocorrelation of mouth cancer incidence of Tamil Nadu.

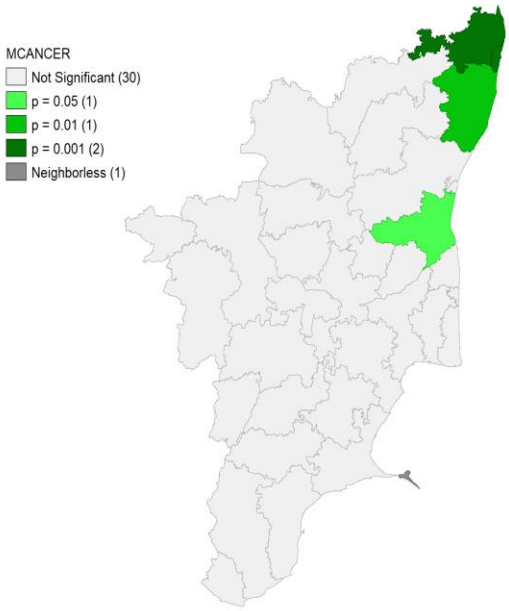


Fig.1. explains the global spatial autocorrelation of mouth cancer incidence of Tamil Nadu. It is found that the autocorrelation of disease was almost 0.329(33%) in Tamil Nadu.

Figure: 2. Local Spatial Autocorrelation of mouth cancer in Tamil Nadu district

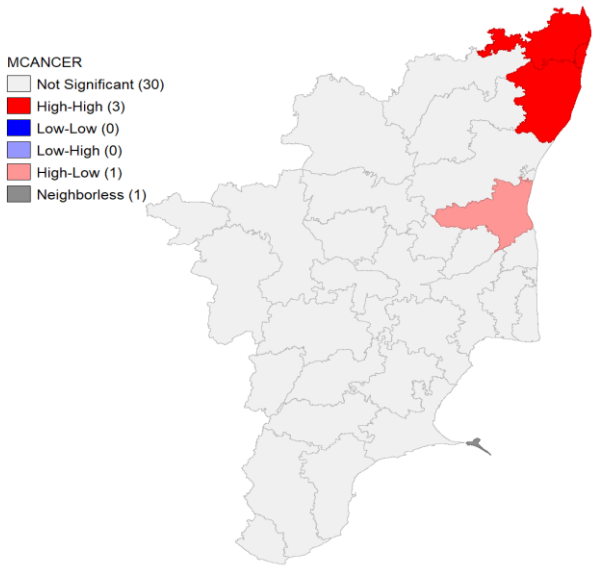
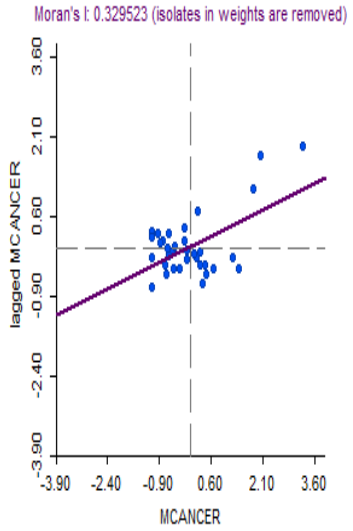


Fig. 2,indicates that the Global spatial autocorrelation explains overall spatial dependence or correlation of the disease, whereas local spatial autocorrelation explains the risk of mouth cancer High High clustering is found in Chennai, Kanchipuram, Thiruvallurdistricts and High low cluster is found in Cuddalore district.

The Local spatial LISA-significant map identifies statistically significant (95% confidence level; $p < 0.05$) spatial clusters of districts with high or low mouth cancer cases. Clusters of districts with high mouth cancer incidence high-high(HH) are considered as "hotspots", whereas clusters of the districts with low mouth cancer incidence low-low(LL) are considered as cold spots. In addition, the local Moran's *I* identifies districts with high mouth cancer incidence that are surrounded mainly by districts with a low incidence known as high-low(HL) spots and districts with a low incidence of mouth cancer that are dominated by districts with a high incidence of mouth cancer is known as low-high(LH) spots.

Figure 3. Local Spatial LISA- Significant Map of mouth cancer in Tamil Nadu district



The mouth cancer incidence rate is in horizontal axis and their spatial lagged counterparts are on the y-axis. The incidence rates have been standardized with mean 0 and standard deviation is 1. The spatial autocorrelation of Moran's I show (33%) spatial correlation existence of mouth cancer in the Tamil Nadu districts were studied. The existence of spatial clusters were found in Northern part of Tamil Nadu, the type of clusters High High is found in Chennai, Kanchipuram, Thiruvallur districts and High low cluster is found in Cuddalore district. Based on the results, our model showed that spatial autocorrelation was not constant in all the districts of Tamil Nadu. Chennai and nearby districts have spatial autocorrelation

of mouth cancer diseases. The spatial model is an emerging model in data sciences and accounts for the correlation among the geographical positions by giving the robust estimates of the value. This spatial model is high dimensional in nature is needed to estimate disease risk in state of Tamil Nadu.

Table1: Comparison of actual mouth cancer cases with all three models of the districts.

Districts.	Mouth cancer	Classic Model		Spatial Error			Spatial Lag Model		
		Predicted	Residual	Predicted	Residual	P.Err	Predicted	Residual	P.Err
TVLR	272	119.6	152.3	122.6	74	149.3	132.9	77	139.1
CHEN	378	128	249.9	130.7	163.1	247.2	140.7	168.1	237.3
VELL	114	112.7	1.2	115.8	-29.3	-1.8	120.3	-27.7	-6.3
KANC	254	120.5	133.4	123.4	84.2	130.5	128.3	87.2	125.6
DHAR	27	92.5	-65.5	96.2	-70.1	-69.2	98.4	-72.4	-71.4
TVMA	80	104.3	-24.3	107.7	-39.4	-27.7	109.7	-39.4	-29.7
SALE	122	100.2	21.7	103.6	18.7	18.3	94.7	26.4	27.2
VLPM	122	101.9	20	105.3	38.5	16.6	105.4	35.5	16.5
EROD	88	100.1	-12.1	103.6	5.3	-15.6	109.7	-1.6	-21.7
CUDD	128	102.7	25.2	106.2	34.7	21.8	80.3	56.9	47.6
NLGR	14	119.9	-105.9	122.9	-113.8	-109	127.4	-115	-113
NMKL	50	104.1	-54.1	107.5	-46.4	-57.5	110.6	-49.5	-60.6
PERA	20	103.2	-83.2	106.6	-85.5	-86.6	103.6	-86.2	-83.6
KARU	43	97.4	-54.4	101	-61.3	-58	100	-65.8	-57
COIM	155	119.8	35.1	122.8	61.9	32.1	127.1	56.6	27.8
TRCY	112	117	-5	120.1	9.9	-8.1	120.8	6.8	-8.8
TANJ	218	120.1	97.8	123	108.9	94.9	106	119.7	111.9
ARIY	43	106.5	-63.5	109.8	-51	-66.8	113.8	-54.9	-70.8
TVAR	106	118.6	-12.6	121.5	-35.4	-15.5	81.8	-4.3	24.1
NAGA	136	114.6	21.3	117.7	20.1	18.2	75.7	46.9	60.2
DIND	88	108.7	-20.7	112	-11.8	-24	118	-16.9	-30
PUDU	56	96.5	-40.5	100.2	-31.7	-44.2	107.9	-42.8	-51.9
SVGN	32	115.1	-83.1	118.1	-70	-86.1	114.3	-64.2	-82.3
MADU	132	116.9	15	120	40.9	12	125.7	33.5	6.3
THEN	37	109.3	-72.3	112.5	-63.7	-75.5	120.5	-69.7	-83.5
RMND	35	115.6	-80.6	118.6	-46.1	-83.6	124.7	-54.3	-89.7

VIRU	70	113.8	-43.8	116.9	-12.4	-46.9	125.7	-22.4	-55.7
TNVL	81	115.3	-34.3	118.4	-27.4	-37.4	132.4	-38	-51.4
THOO	55	125.1	-70.1	127.9	-39.2	-72.9	136.1	-46.9	-81.1
KANY	204	131.4	72.5	134	92.5	69.9	144.4	86.4	59.5

Table 1 shows that the reported mouth cancer cases of Tamil Nadu at district level and predicted for the Classical ordinary least square(OLS) regression, Spatial Lag model and Spatial Error models. From the table, infer that the Chennai District had the highest number of registered mouth cancer cases. Comparatively the spatial lag model had a higher predicted value for all the districts in Tamil Nadu.

Districts Abbreviation.

TVLR- Thriuvallur, CHEN- Chennai, VELL-Vellore, KANC- Kanchipura, DHAR-Dharmapuri, TVMA-Thiruvannamalai, SALE-Salem, VLPM-Villuppuram, EROD- Erode, CUDD-Cuddalore NLGR-Nigiri,NMKL- Namakkal, PERA- Perambalur, KARU-Karur, COIM- Coimbatore, TRCY-Trichy, TANJ-Tanjore, ARIY-Ariyalur, TVAR-Tiruvarur, NAGA-Nagappattinam, DIND- Dindigul, PUDU-Pudukottai, SVGN-Sivagangai, MADU-Madurai, THEN-Theni, RMND-Ramanathapuram, VIRU-Virudunagar, TNVL-Tirunelveli,THOO-Thoothukkudi, KANY-Kanniyakumari. (* Delimited Districtricts of Krishnagiri and Tiruppur are overlapped)

Table2. Comparison of ordinary linear regression, spatial error model and spatial lag model

Variable	Classical Model			Spatial Error model			Spatial Lag Model		
	Co.eff	S.E	Prob.	Co.eff	S.E	Prob.	Co.eff	S.E	Prob.
W_M Cancer	-	-	-	-	-	-	0.52	0.14	0.00
Constant	-16.2	33.5	0.63	-9.4	31.1	0.76	-50.8	28.9	0.07
Tobacco use	0.9	4.4	0.84	0.8	3.1	0.77	0.6	3.6	0.86
Literacy rate	-1.6	0.5	0.003	-1.6	0.4	**0.001***	-1.4	0.4	0.001
Lambda				0.6	0.1	0.001			

Table 3: Goodness of fit for all models

Model Fitness	Classic Model	SpatialError Model	Spatial Lag Model
Log Likelihood	-199.7	-194.7	-195.6
Akaikeinfo criterion	405.4	395.4	399.3
Schwarz criterion	410.1	400.1	405.5

Table 2 and 3, shows that comparison was made on the goodness of fit for the three models. It is observed that the spatial error model gives minimum Akaike Info criterion (AIC 395.366) value when compared with the other two models.

CONCLUSION

In this study employed various spatial models and find out the spatial dependencies, spatial clusters of the mouth cancers in all districts of Tamil Nadu, and AIC of all models were compared , it is found that the spatial error model gives the lowest Akaike info criterion value and concluded this model is more suitable.

Acknowledgments

All the authors of this research paper acknowledge with gratitude of the staff members in the Division of Epidemiology and Cancer Registry at the Cancer Institute (WIA) Chennai with Data facilitation and their tired less work since its beginning of Tamil Nadu Cancer Registry Project (TNCRP).

References:

1. Ajay PR, Ashwinirani S R, Nayak A, Suragimath G, Kamala K A, Sande A, Naik RS. Oral cancer prevalence in Western population of Maharashtra, India, for a period of 5 years. J Oral Res Rev 2018.

2. Al- Ahmed K, Al- Zahrani A. Spatial autocorrelation of cancer incidence in Saudi Arabia. *Int J Environ Res Public Health. 2013.*

3. Boscoe FP, Ward MH, Reynolds P. Current practices in spatial analysis of cancer data: data characteristics and data sources for geographic studies of cancer. *Int J Health Georg.2004.*

4. Census of India, Cause of death in India 2010-2013

5. Fahrmeir L, Tutz G. Multivariate Statistical Modelling Based on Generalized Linear Models. New York; Springer. 2001.

6. FaribaRamezani., MahshidGhoncheh., Reza Pakzad., HamidrezaSadeghi., FereshtehGhorat., Hamid Salehiniya. Epidemiology, incidence and mortality of oral cavity and lips cancer and their relationship with the human development index in the world. in Biomedical Research and Therapy, 2016.

7. Friedman, D. B., & Hoffman= Goetz, L. (2008). Literacy and Helath literacy as defined in cancer education research: A Systematic review. Health Education Jurnal.

8. GLOBOCAN 2018, IARC, Lyon, France

9. Goli A, Oroei M, Jalalpour M, Faramarzi H, Askarian M. The Spatial Distribution of Cancer Incidence in Fars Province: A GIS- Based Analysis of Cancer Registry Data. *Int J Prev Med.2013.*

10. Janbaz KH, Qadir MI, Basser HT ,Bokhari TH, Ahmad B. Risk for oral cancer from smokeless tobacco. ContempOncol (Pozn),2014.

11. LaVecchia, C.; Decarli, A. Correlations between cancer mortality rates from various Italian regions. *Tumori***1985**.

12. Mandal, R.; St-Hilaire, S.; Kie, J.G.; Derryberry, D. Spatial trends of breast and prostate cancers in the United States between 2000 and 2005. *Int. J. Health. Geogr*.**2009.**

13. McCullagh P, Nelder JA. Generalized Linear Models. London: Chapman and Hall, 1989

14. Nelder JA, Wedderburn RWM, Generalized linear models. J.R Stat, Soc, A. 1972.

15. Nirmala CJ et al. Oral cancer and tobacco: a case control study in southern India. *Int J Community Med Public Health. 2017.*

16. Poddar A, Aranha RR, K Muthukaliannan G, et al. Head and neck cancer risk factors in India: protocol for systematic review and meta-analysis BMJ Open 2018.

17. Reshadat, S., Saeidi, S., Zangeneh, A. et al. A Comparative Study of Spatial Distribution of Gastrointestinal Cancers in Poverty and Affluent Strata (Kermanshah Metropolis, Iran) J Gastrointest Can (2018).

18. Rosenberg, M.S.; Sokal, R.R.; Oden, N.L.; DiGiovanni, D. Spatial autocorrelation of cancer in Western Europe. *Eur. J. Epidemiol.* **1999**.

19. Rosenberg, M.S. The bearing correlogram: A new method of analyzing directional spatial autocorrelation. *Geogr.Anal.***2000**.

20. Shanta V, Gajalakshmi CK, Swaminathan R, Ravichandran K, Vasanthi L. Cancer registration in Madras Metropolitan Tumour Registry. *Eur J Cancer 1994*.

21. Sharma DC. Cancer Institute in Chennai: a model for resource-poor countries. Lancet Oncology 2004.

22. SmitaAsthana, MD SatyanarayanaLabani, PhD Uma Kailash, PhD Dhirendra N Sinha, MD Ravi Mehrotra, MD. Association of Smokeless Tobacco Use and oral Cancer: A Systematic Global Review and Meta-Analysis, Nicotine & Tobacco Research, 2018.

23. Swaminathan R, Selvakumaran R, Esmy PO, **Sampath P**, Ferlay J, Jissa V, Shanta V, Cherian M, Sankaranarayanan R. Cancer pattern and survival in a rural district in South India.*CancerEpidemiol. 2009*.

24. Swaminathan R, Shanta V, Balasubramanian S, **Sampath P**. Cancer Incidence in Chennai, India (2008-2012) In: Bray F, Colombet M, Mery L, Pineros M, Znaor A, Zanetti R and Ferlay J, (Eds). Cancer Incidence in Five Continents,Vol XI (electronic version). Lyon: International Agency for Research on Cancer, 2017.

25. Yomralioglu T, Colak EH, Aydinoglu AC. Geo-relationship between cancer cases and the environment by GIS: a case study of Trabzon in Turkey. *Int. J Environ Res Public Health*, 2009.