Y. Y. Ü. Fen Bilimleri Enstitüsü Dergisi
Yıl : 2010, Cilt : 15, Sayı : 1, Sayfa :40-45

Eksik Gözlemlerin Tahmin Edilmesinde Hot Deck Atfı Yöntemi Kullanılarak Farklı Uzaklık Ölçütlerin Karşılaştırılması[1]

Yılmaz KAYA[1], Abdullah YEŞİLOVA[2], M.Nuri ALMALI[3]

[1]Yüzüncü Yıl Üniversitesi, Van Meslek Yüksekokulu, Bilgisayar Teknolojileri ve Programcılığı, 65080 Van
[2]Yüzüncü Yıl Üniversitesi, Ziraat Fakültesi, Zootekni, Biyometri Genetik Anabilimdalı, 65080 Van
[3]Yüzüncü Yıl Üniversitesi, Mühendislik Mimarlık Fakültesi, Elektrik ve Elektronik Anabilimdalı, 65080 Van

Özet : Bu çalışmada birden fazla değişken için farklı oranlarda rasgele eksik gözlemler üretilmiştir. Eksik gözlemlerin tamamıyla şansa bağlı (missing completely at random=MCAR) olması durumunda eksik gözlemlerin tahmin edilmesi için Hot Deck (HD) atfı yöntemi kullanılmıştır. HD atfı yöntemi uygulanabilirliği kolay bir yöntemdir. HD atfında uzaklıkların hesaplanması için Öklid, Manhattan ve Minkowski uzaklık ölçütleri kullanılmıştır. Farklı uzaklık ölçütleri ile elde edilen veri setlerin gerçek veri seti ile karşılaştırılmasında korelasyon katsayısı ile standart hata değerleri kullanılmıştır. Yapılan analizler sonucunda eksik gözlemleri tahmin edilmesinde HD atfı ile Manhattan uzaklık ölçütünün Öklid ve Minkowski uzaklık ölçütlerine göre daha etkin olduğu saptanmıştır.
Anahtar Kelimeler: Hot Deck atfı, Uzaklık Ölçütleri, Eksik gözlem

Comparing Distance Metrics Via Hot Deck Imputation Method For Estimating Missing Values

Abstract: In this study, missing values in different proportions and belonging to more than a variable were produced randomly. When observation were considered within a context which is missing completely at random (MCAR) Hot Dect (HD) Impuation method were used for estimated missing values. HD method an easy method. Oclid, Manhattan and Minkowski distance metrics used for calculated distances In HD method. Correlation and standart error were used for comparing data set that optained by different distance metrics. As a result of analysis, using Manhattan distance metric in Hot Decl Imputation method was found more effective than Oclid and Minkowski metrics in the estimation pf missing value.
Key Words: Hot Deck Imputation, Distance Metrics, Missing value

Giriş

Yapılan çalışmalarda doğru sonuçların elde edilebilmesi, verilerin eksiksiz bir şekilde elde edilmesine bağlıdır. Eksik gözlemler, uygulanacak istatistiksel analiz sonuçların yanlış olmasına neden olabilir. Ancak farklı nedenlerden dolayı çalışmanın belirli bir kısmı gözlemlenemeyebilir ve bunun sonucunda eksik gözlemler meydana gelebilir. Eksik gözlemler bir değişkene ait olabileceği gibi birden fazla değişkene de ait olabilir. Bazı çalışmalarda eksik gözlemler işlem dışı bırakılarak analizler yapılmaktadır. Bununla birlikte eksik gözlemlerin analiz dışında bırakılması yanlı parametre tahminlerine neden olabilir. Ayrıca veri setinden çıkarılan eksik gözlemler çalışmanın geçerliliğini ve genellenebilirliğini etkilemektedir (Draper, 1998). Veri setinde eksik gözlem olması durumunda, Hot Deck atfı, regresyon atfı, çoklu atıf, EM (Expection-Maximization), yapay sinir ağları, yerine atıf (ortalama, minimum, maksimum ve medyan) gibi yöntemler kullanılarak eksik gözlemlerin tahmin edilmesi mümkündür.

Eksik gözlemlerin tahmin edilmesi için farklı yöntemler geliştirilmiştir. Uygun yöntemin seçilmesi eksik gözlem mekanizmalarına bağlıdır. Literatürde bu mekanizmalar üç kategoride değerlendirilmektedir. Bunlar tamamıyla rasgele olarak eksik (missing completely at random, MCAR), rasgele olarak eksik (missing at random, MAR) ve ihmal edilemez (non ignorable, NI) (Draper, 1998; Pelckmans, 2005; Jerez ve ark., 2006) olarak sıralanabilir.

Bu çalışmada eksik gözlemlerin tamamıyla şansa bağlı olması (MCAR) durumu incelenmiştir. MCAR'da, bir X değişkeninde eksik gözlem olması diğer herhangi bir değişkene ve X değişkenin kendisine bağlı olmadığı varsayılmaktadır (Mohamed, 2005; Schafer, 1999).

Materyal

Çalışmada kullanılan veri seti 2003-DPT-MİM1 numaralı proje kapsamında yapılan çalışmalardan elde edilmişti. Bu çalışmada Yüzüncü Yıl Üniversitesi yerleşkesine konulmuş mikroişlemci kontrollü ölçüm cihazları, standartlara uygun şekilde 30 m ve 10 m yükseklikte rüzgar hızları ölçülmüştür. Rüzgar hızı verileri 10'ar dakika arayla kaydedilmiş Nisan-2004 ile Mart-2005 tarihleri arasında bir yıllık dönemi kapsamaktadır. Bu çalışmada Şubat 2005 tarihine ait 4032 gözlemden oluşan veri seti kullanılmıştır Veri setindeki değişkenler Çizelge 1'de verilmiştir.

Çizelge1: Veri setindeki değişkenler

Kod	Değişken
S1	30 Metre Yükseklikte Rüzgar Hızı
S2	10 Metre Yükseklikte Rüzgar Hızı
S3	Yön
S4	Sıcaklık
S5	Pyranometre
S6	Basınç
S7	Nemlilik

[1]19.İstatistik Araştırma Sempozyumunda sözlü bildiri olarak sunulmuştur.

Yöntem

Hot Deck Atfı: Hot Deck atfı yöntemi, çok fazla matematik ve istatistik bilgisi gerektirmeden veri setindeki eksik gözlemleri tamamlayan önemli bir yöntemdir (Kalton ve Kish,1984). Hot Deck atfı yöntemi, teorik özelliklerinin basitliğinden dolayı tercih edilmektedir. Hot deck atfında, veri matrisindeki eksik gözlemler benzer gözlenen gözlemlerle doldurulur (Joseph ve ark., 2002). Çizelge 2'de bu amaçla verilmiş bir örnek durumu içermektedir.

Çizelge 2: Eksik gözlemli veri seti.

Sırano	X1	X2	X3	X4	X5
1	1	4	3	5	5
2	1	4	-	-	7
3	2	-	-	4	5
4	4	3	6	7	5

Çizelge 3: Hot Deck atfı yöntemi ile tamamlanmış veri seti.

Sırano	X1	X2	X3	X4	X5
1	1	4	3	5	5
2	1	4	3	5	7
3	2	3	6	4	5
4	4	3	6	7	5

Çizelge 2 incelendiğinde, X2 değişkeni için 3 durum, X3 değişkeni için 2. ve 3. durumlar, X4 değişkeni için 2. durumun değerlerinin eksik gözlem içerdiği görülmektedir. Hot Deck atfı, veri setinde değerlerin tam olduğu satırları araştırır ve eksik gözlemler için en çok benzer olduğu tam satırlardan eksik gözlemleri tamamlar. Örnek için tam gözlem değerine sahip durumlar 1 ve 4' dir. Bu 2 ve 3. durum değerleri incelendiğinde, durum 2 için değerlerin, durum 1' e durum 3 için değerlerin durum 4'e daha benzer olduğu sonucuna ulaşmaktayız. Eksik gözlem içeren satırların benzedikleri satırlara göre Çizelge 3 oluşmuştur. Hot Deck atfı yöntemi, eksik gözlemlerin yerine değişkenin ortalamasının konulmasına göre ilgili değişkenin varyansını artırmaktadır (Schimert ve ark., 2001; Fuller, Wayne ve Jae, 2005). Hot Deck atfı uzun bir kullanım tarihine sahiptir. Bu atıf, liste bazında veri silme, çiftler bazında veri silme, yerine ortalamayı koyma yöntemlerinden üstün bir tekniktir. Hot Deck atfının avantajları arasında kavramsal basitliği, değişkenlerin ölçüm düzeylerini koruması (kategorik değişkenler kategorik olarak, sürekli değişkenler sürekli olarak kalır) ve tamamlanmış veri matrisi elde edilmesi sayılabilir. Farklı Hot Deck atfı metotları kullanılmaktadır. En yakın k komşu Hot Deck atfı yöntemi en çok tercih edilen yöntemlerden biridir. Eksik gözleme sahip satırların tam olan satırlara olan uzaklıkları hesaplamayı dikkate alan bir yöntemdir (Juned ve Thomas, 2008).

En yakın komşu Hot Deck atfında en uygun satırın bulunması için k en yakın komşu (K Nearest Neighbor) algoritması kullanılır. Eksik gözleme sahip veri satırının tam olarak gözlenmiş satırlara olan uzaklıkları hesaplanarak en yakın olunan tam satırdan eksik gözlemler tamamlanır. Algoritma aşağıda adımlarla anlatılmıştır (Stefano ve Giuseppe, 2007).

1) Veri seti eksik gözlemli veri seti (Incomplete Data Set) ve tam veri seti (Complete Data Set) şeklinde bölünür.

2) X_i tamamlanmış veri setini belirten veri matrisi ve x_{ij} J. değişkenine ait i.'ci gözlem olsun. Y_i tamamlanmamış veri setini belirten veri matrisi ve y_{ij} J. değişkenine ait i.'ci gözlem olsun.

3) Her eksik veri içeren satırlar için öklid, Manhattan veya Minkowski uzaklıkları hesaplanır.

$$\text{Öklit}(d) = \sqrt{\sum_{j=1}^{n}(y_{kj} - x_{ij})^2} \quad (1)$$

$$\text{Manhattan}(d) = \sum_{j=1}^{n}(|y_{kj} - x_{ij}|) \quad (2)$$

$$\text{Minkowski}(d) = \left[\sum_{j=1}^{n}(|y_{kj} - x_{ij}|^m)\right]^{\frac{1}{m}} \quad (3)$$

Burada m veri setindeki değişken sayısıdır. m=2 olması durumunda öklid uzaklık bağıntısı elde edilmiş olur.

4) En yakın k komşu sayısına göre uzaklıklar belirlendikten sonra uygun tam satır bulunup, eksik gözlem içeren tamamlanmamış veri seti için eksik gözlemler bulunur. Bu yönteme K en yakın komşu atfı da denilmektedir

Uzaklık Ölçütlerinin Karşılaştırılması:

Standart Hata: Standart hata gerçek gözlemler ile tahmin edilen gözlemler arasındaki farkın ortalama sapmasını belirtmek için kullanılabilir (Draper ve Smith, 1998). Gerçek gözlem değerleri (X_i) ile tahmin değerler ($\hat{X}_i$) için standart hata,

$$\sigma = \frac{\sum_{i=1}^{n}(X_i - \hat{X}_i)^2}{n} \quad (4)$$

olarak hesaplanır. Standart hatanın küçük olması tercih edilir. Standart hata büyüdükçe gözlemler arasında bir sapmanın olduğunu belirtir.

Korelasyon Katsayısı: Gerçek gözlem değerleri (X_i) ile tahmin değerler ($\hat{X}_i$) arasındaki korelasyon katsayısı,

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_i)(\hat{X}_i - \bar{\hat{X}}_i)}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X}_i)\sum_{i=1}^{n}(\hat{X}_i - \bar{\hat{X}}_i)}} \quad (5)$$

olarak hesaplanmaktadır.

Bulgular

Bu çalışmada kullanılan veri seti için öncelikle %5, %13, %20 oranında tamamıyla şansa bağlı olarak eksik gözlemler oluşturuldu. Daha sonra Hot Deck atfı ile Öklid, Manhattan ve Minkowski uzaklık ölçütleri kullanılarak eksik gözlemleri tamamlanmış veri setleri elde edildi. Uzaklık ölçütlerin etkilerini araştırmak için farklı uzaklık ölçüleri kullanılarak elde edilmiş olan veri setlerindeki değişkenler ile gerçek veri setindeki değişkenler arasındaki korelasyon katsayıları kullanıldı. Elde edilen korelasyon katsayıları Çizelge 4'te verilmiştir.

Çizelge 4: Farklı uzaklık ölçülerine göre elde edilen veri setleri ile gerçek veri seti arasındaki korelasyon katsayıları.

Değişken	Öklid	Manhattan	Minkowski
Eksik Gözlem Oranı %5			
S1	0.991	**0.993**	0.991
S2	0.990	**0.990**	0.989
S3	0.974	**0.974**	0.972
S4	0.987	**0.987**	0.985
S5	0.990	**0.991**	0.988
S6	0.977	**0.980**	0.977
S7	0.988	**0.988**	0.988
Eksik Gözlem Oranı %13			
S1	0.967	**0.971**	0.962
S2	0.971	**0.972**	0.968
S3	0.931	**0.936**	0.925
S4	0.956	**0.958**	0.953
S5	0.967	**0.972**	0.965
S6	0.945	**0.946**	0.938
S7	0.963	**0.969**	0.960
Eksik Gözlem Oranı %20			
S1	0.940	**0.944**	0.933
S2	0.939	**0.943**	0.930
S3	0.849	**0.849**	0.842
S4	0.920	**0.922**	0.917
S5	0.946	**0.946**	0.946
S6	0.902	**0.903**	0.898
S7	0.925	**0.932**	0.921

Çizelge 5'te, HD atfı ile Öklid, Manhattan ve Minkowski uzaklık kullanılarak elde edilmiş olan veri setleri ile gerçek veri seti arasında güçlü bir ilişkinin olduğu saptanmıştır. Korelasyon katsayılarına bakıldığından tüm değişkenler için HD atfı ile Manhattan uzaklık ölçütünün kullanılması daha iyi sonuçlar verdiği saptanmıştır.

Gerçek veri seti ile farklı uzaklık ölçülerine göre elde edilen veri setlerine ait ortalama değerler Çizelge 6'da verilmiştir.

Çizelge 6: Farklı uzaklık ölçülerine göre elde edilen veri setleri ve gerçek veri seti için ortalama değerler.

Değişken	Öklid	Manhattan	Minkowski	Gerçek Veri Seti
%5 Eksik Gözlem				
S1	4.3988	**4.3995**	4.3999	4.3993
S2	3.4221	**3.4208**	3.4224	3.4145
S3	**87.87**	88.04	87.94	87.79
S4	-2.7440	**-2.7438**	-2.755	-2.7390
S5	151.15	**151.30**	151.29	151.36
S6	101.58	**101.58**	101.58	101.58
S7	65.938	65.960	**65.924**	65.915
%13 Eksik Gözlem				
S1	4.4132	**4.4054**	4.4102	4.3993
S2	3.4328	**3.4312**	3.4312	3.4145
S3	86.40	**87.37**	86.54	87.79
S4	-2.7723	**-2.7921**	-2.7660	-2.7390
S5	**151.58**	150.94	151.66	151.36
S6	101.57	**101.57**	101.57	101.58
S7	65.968	**65.91**	65.964	65.915
%20 Eksik Gözlem				
S1	4.4109	**4.4006**	4.4114	4.3993
S2	3.4254	**3.4152**	3.4324	3.4145
S3	86.56	**86.48**	86.29	87.79
S4	-2.7959	**-2.7838**	-2.7845	-2.7390

S5	152.78	152.85	152.22	151.36
S6	101.57	101.57	101.57	101.58
S7	65.984	65.981	65.950	65.915

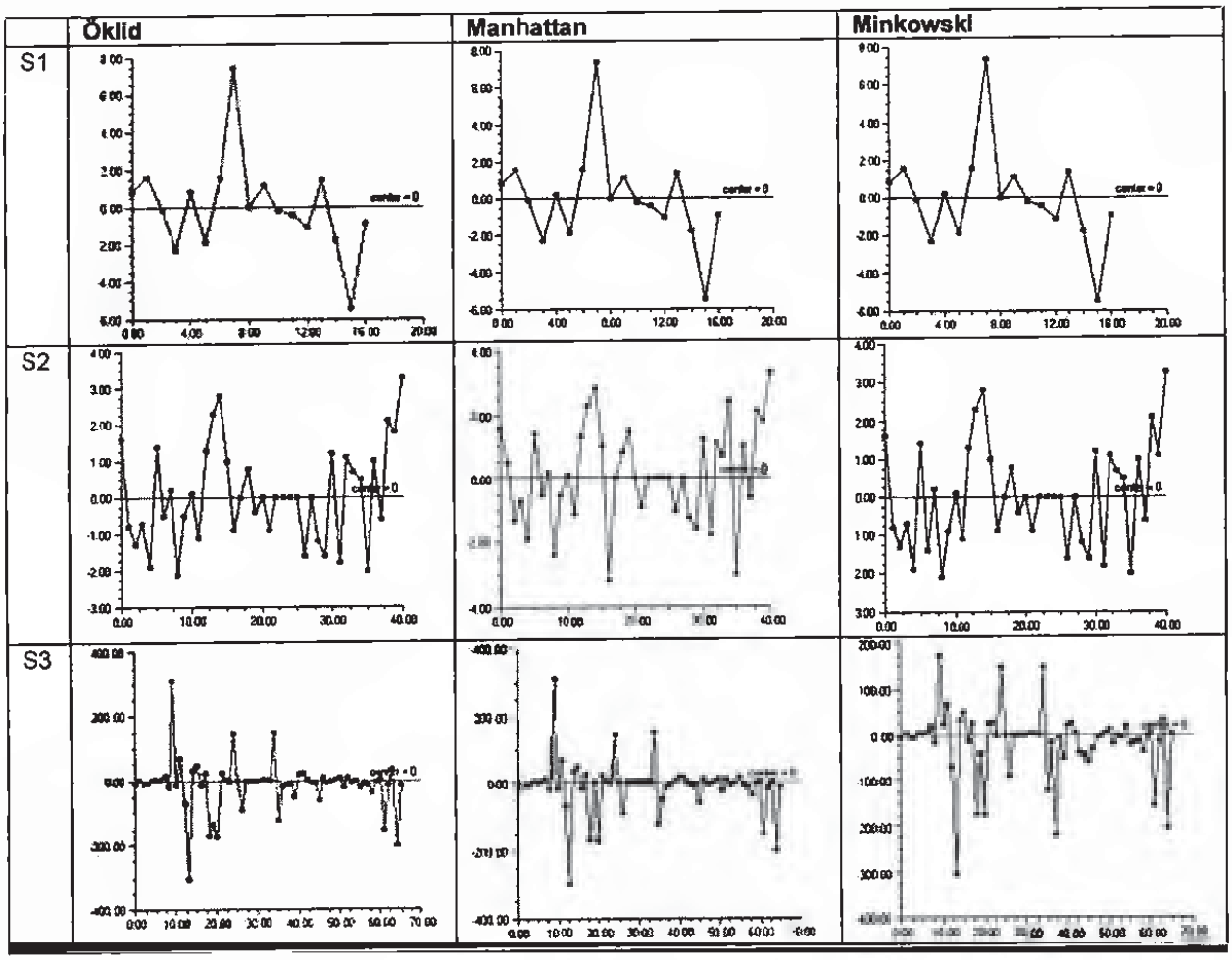
Çizelge 6'ya bakıldığında HD atfı ile Manhattan uzaklık ölçüsü kullanılarak elde edilen veri setindeki değişkenlere ait ortalama değerler gerçek veri setindeki değişkenlere ait ortalama değerlere daha yakın olarak elde edilmiştir.

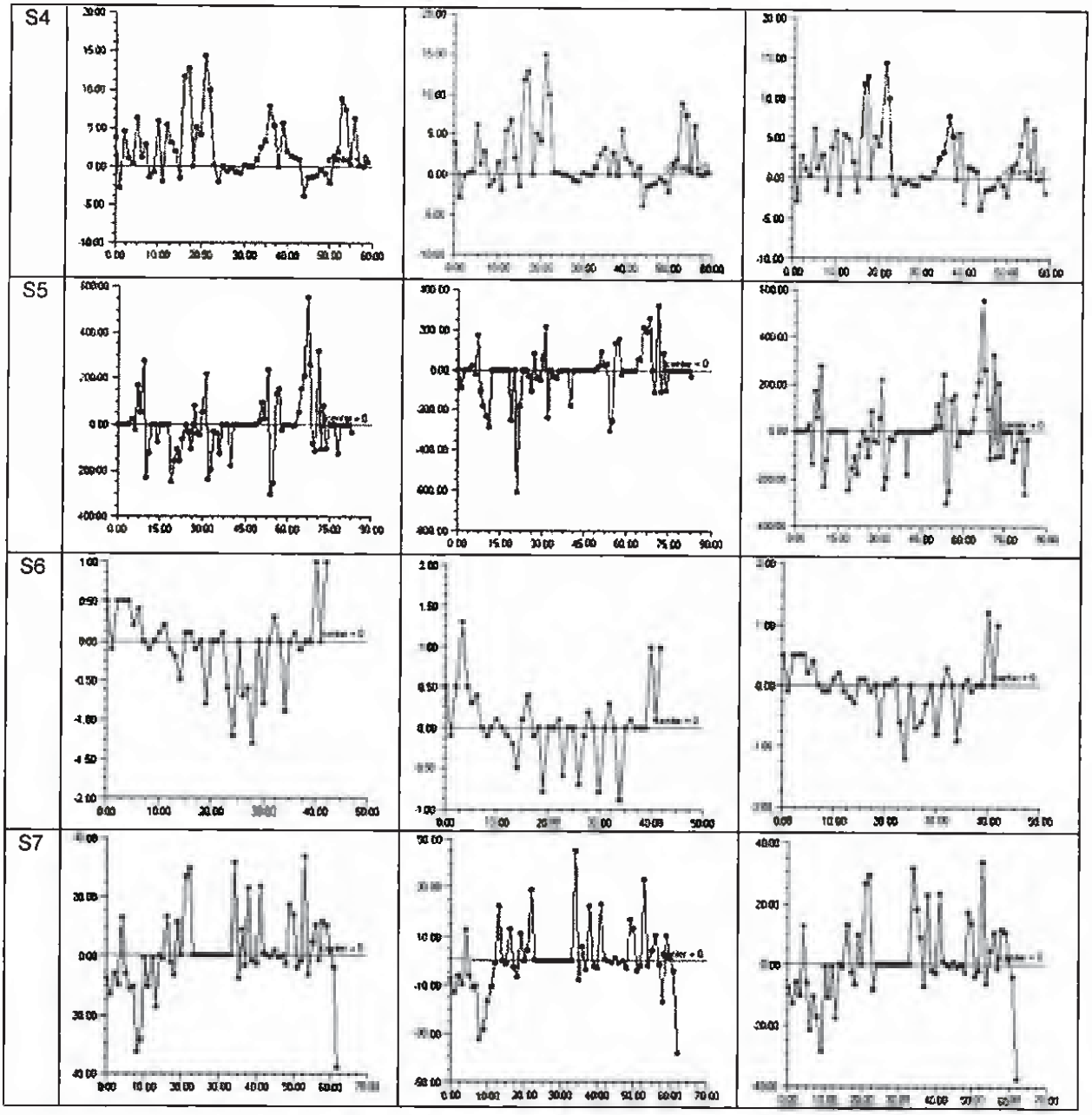
Farklı uzaklık ölçülerine göre elde edilen veri setleri ile gerçek veri seti arasındaki standart hata değerleri Çizelge 7'de verilmiştir.

Çizelge 7: Farklı uzaklık ölçülerine göre elde edilmiş veri setleri ile gerçek veri seti arasındaki standart hata değerleri.

Model	S1	S2	S3	S4	S5	S6	S7
% 5 Eksik Gözlem							
Öklit	0.08	0.08	342.8	0.59	1109.87	0.01	5.94
Manhattan	**0.07**	**0.07**	**373.58**	**0.58**	**989.47**	**0.01**	**5.86**
Minkowski	0.09	0.08	403.29	0.66	1274.09	0.01	6.05
%13 Eksik Gözlem							
Öklit	0.32	0.21	956.87	1.94	3506.2	0.03	18.53
Manhattan	**0.28**	**0.2**	**894.11**	**1.86**	**2981.76**	**0.03**	**15.43**
Minkowski	0.37	0.24	1042	2.07	3773.1	0.03	20.06
%20 Eksik Gözlem							
Öklit	0.59	0.45	2105.67	3.49	5614.47	0.04	37.97
Manhattan	**0.55**	**0.42**	**2109.85**	**3.39**	**5772.39**	**0.04**	**34.36**
Minkowski	0.66	0.52	2203.81	3.61	5765.64	0.05	39.69

Çizelge 7'ye bakıldığında hemen hemen tüm değişkenler için en küçük standart hata değerleri HD atfı ile Manhattan uzaklık ölçütü kullanılarak elde edilmiştir. %5 eksik gözlem içeren veri seti için farklı uzaklık ölçülerinin kullanılması ile elde edilen eksik gözlemler ile gerçek gözlemler arasındaki hatalara ait grafikler Şekil 1'de verilmiştir.





Şekil 1: Farklı uzaklık ölçütleri kullanılarak elde edilmiş olan veri setleri ve gerçek veri seti arasındaki hata değerlerine ait grafikler.

Elde edilen hataların sıfır noktasına yakın olması eksik gözlemlerin gerçek değere ne kadar yakın olduğunu göstermektedir. Grafiklere bakıldığında tüm değişkenler için Manhattan uzaklık ölçütünün kullanılmasıyla elde edilen veri seti ile gerçek veri seti arasındaki hata değerlerinin sıfır noktasına daha yakın olduğu saptanmıştır.

Sonuç

Çalışmada hem eksik gözlemlerin tahmin edilmesinde hem de analizler için MS Visual C# programlama dili ile geliştirilen program kullanılmıştır. Bu çalışmada eksik değerlerin tahmin edilmesinde Hot Deck atfı yöntemi kullanılarak Öklid, Manhattan, Minkowski gibi uzaklık ölçütlerinin etkinlikleri karşılaştırılmıştır. HD atfı ile Manhattan uzaklık ölçütü kullanılarak elde edilen eksik gözlemlerin Öklid, Minkowski uzaklık ölçütlerine göre gerçek değerlere daha yakın sonuçlar verdiği saptanmıştır. Ayrıca, edilen korelasyon katsayıları, standart hata değerleri ve grafiklere bakıldığında HD atfında uzaklıkların hesaplanması aşamasında Manhattan uzaklık ölçütünün öklid, Minkowski gibi uzaklık ölçütlerine göre eksik gözlemleri tahmin etmede daha etkin olduğu saptanmıştır.

Kaynaklar

Draper N., ve Smith, H., 1998. Applied regression analysis. J. Wiley, New York, third edition.

Fuller, W. ve Jae., K., K., 2005. Hot deck imputation for the response model. Statistics Canada, 31(2):139–149, 2005.

Jerez, J. M., Molina, I., Subirats, J. L., Franco, L., 2006. Missing Data Imputation In Breast Cancer Prognosis. Processing of the 24th

IASTED International Multi-Conference Boimedical Enginerring. February 15-17. Innsbruck, Austria.

Joseph L., Schafer, J.,Graham, W., 2002. Missing Data: Our View of the State of the Art, Psychological Methods, 7: 147-177.

Juned S., Thomas R., B., 2008. Multiple imputation using an iterative hot-deck with distance-based donor selection. Statistics In Medicine. 27: 83-102.

Kalton, G., Kish, L., 1984. Some efficient random imputation methods. Commun. Statist.-Theor. Meth., 13(16), 1919–1939.

Mohamed, S., Marwala, T., 2005. Neural Network Based Techniques for Estimating Missing Data in Databases, 16th Annual Symposium of the Patten Recognition Association of South Africa, Langebaan, 2005, pp.27-32.

Pelckmans, K., Brabanter, J., D., Suykens, J.,A.,K., Moor, B., D., 2005. Handling missing values in support vector machine classifiers. Neural Networks. 18: 684-692.

Schafer, J., 1999. Multiple Imputation: A Primer, Statistical Methods in Medical Research, 8: 3-15.

Schimert, J., Schafer, J. L., Hesterberg, T., Fraley, C., Clarkson, D., 2001. Analyzing missing values in S-PLUS. Seattle, WA: Insightful.

Stefano, M., I., ve Giuseppe P., 2007. Missing data imputation, matching and other applications of random recursive partitioning. Computational Statistics & Data Analysis. 52: 773- 789

Wayne A., F., Jae, K., K., 2001. Hot Deck Imputation fort he response model. Proceedings of Statistics Canada Symposium

Yenduri, S., 2007. Performance Evaluation of Imputation Methods for Incomplate Datasets. International Journal of Software Engineering and Knowledge Engineering. 17,1-26.

Ek 1: C# ile geliştirilen program

Program eksik gözlemlerin tahmin edilmesinde kullanılan Hot Deck, Rassal Hot Deck ve Yerine Atıf yöntemleri için geliştirilmiştir. Yaklaşık 2500 kod satırından oluşmaktadır. Program Microsoft Visual Studio 2005 C# ile geliştirilmiştir.