2020

2(1), 73-87

Kod Atama Sistemi (KASİS) ile Otomatik Kod Atama

Levent AHİ*[1], Ebru KILIÇ ÇAKMAK[2]

Öz

Bu çalışmada, Kod Atama Sistemi (KASİS) ve bu sistemde kullanılan iki farklı kod atama yöntemi tanıtılarak yöntemlerin ve sistemin etkinliği değerlendirilmiştir. KASİS, tutarlı, güvenilir ve sistematik bir kod ataması yapabilmek için geliştirilmiştir. Kod atama, var olan metinsel tanımı, standart olarak oluşturulmuş sınıflama sözlüğünde yer alan tanıma ilişkin en uygun koda dönüştürme işlemidir. Sistemin etkinliği, Türkiye İstatistik Kurumu'nun (TÜİK) uyguladığı Hanehalkı Bütçe Araştırması (HBA) veri seti kullanılarak değerlendirilmiştir. HBA, hanehalkı tüketim harcamalarının ana veri kaynağıdır. HBA' da tüketim harcamaları sınıflaması olarak COICOP kullanılmaktadır. Manuel yöntemlerle anketörler tarafından atanan kodların doğruluğu 2016-2018 yıllarında yapılmış HBA veri seti kullanılarak sistem tarafından kontrol edilmiştir. Daha sonra, kod ataması, yetersiz ve şüpheli olarak sınıflanan kayıtlara sistem aracılığıyla iki farklı yöntemle tekrar kod ataması gerçekleştirilmiştir. Her iki yöntemde de bulanık eşleştirme teknikleri kullanılmıştır. Bulanık eşleştirme teknikleri, iki metnin benzerliğini ölçebilmek amacıyla geliştirilen algoritmaları kullanmaktadır. KASİS' in istatistik üretim aşamasında kullanılması sonucunda veri seti kamuoyu ile paylaşılmadan önce anketörlerin kodlamada yaptığı hata ve eksikliklerin giderilmesi sağlanmış olacaktır.

Anahtar Sözcükler

İstatistiksel sınıflama

COICOP

HBA

Otomatik kodlama

Bulanık eşleştirme

Makale Hakkında

Gönderim Tarihi
10 Mayıs 2020
Kabul Tarihi
19 Haziran 2020
Yayın Tarihi
22 Haziran 2020

Makale Türü
Araştırma Makalesi

Automatic Code Assignment with Code Assignment System (KASIS)

Abstract

In this study, the Code Assignment System (KASIS) developed to make a consistent, reliable and systematic code assignment and two different code assignment methods have been introduced and the effectiveness of the methods and the system have been evaluated. The code assignment process consists of converting the textual definition into the most appropriate code in the classification. The system's effectiveness has been evaluated using the Household Budget Survey (HBS) data set implementing by Turkey Statistical Institute (TURKSTAT). HBS is the main data source of household consumption expenditures. COICOP is used as the classification of consumption expenditures in HBS. The accuracy of the codes assigned by the interviewers has been checked by the system using the HBS 2016-2018 data. Then, the code assignment has been re-assigned to the records classified as insufficient and suspicious by two different methods through the system. Fuzzy matching techniques has been used in both methods. Fuzzy matching techniques use algorithms developed to measure the similarity of the two texts. As a result of the use of KASIS at the stage of statistics production, before the data set is shared with the public, the mistakes and deficiencies made by the interviewers in the coding will be eliminated.

Keywords

Statistical Classification

COICOP

HBS

Automatic Coding

Fuzzy Matching

Article Info

Received
May 10, 2020
Accepted
June 19, 2020
Published
June 22, 2020

ArticleType
Research Paper

Atıf/Cite: Ahi, L., & Kılıç-Çakmak, E. (2020). Kod Atama Sistemi (KASİS) ile Otomatik Kod Atama [Automatic Code Assignment with Code Assignment System (KASIS)]. *Bilgi ve İletişim Teknolojileri Dergisi/Journal of Information and Communication Technologies, 2*(1), 73-87.

* **Sorumlu Yazar/Corresponding Author:** leventahi@tuik.gov.tr
[1] Türkiye İstatistik Kurumu, Ankara/Turkey, leventahi@tuik.gov.tr, https://orcid.org/0000-0002-7415-1173
[2] Prof. Dr., Gazi University, Ankara/Turkey, ekilic@gazi.edu.tr, https://orcid.org/0000-0002-3459-6290

Extended Abstract

Introduction

In order for everyone to understand the same thing from the statistical indicators and achieve the same result, it depends on the fact that these indicators meet certain conditions and standards. In this sense, when the classifications used to produce statistical indicators are considered to be a language that enables communication between people living in different countries, they must meet certain standards. There are many types of statistical classifications created within this scope.

Statistics can be produced depending on administrative records or can be produced by conducting a survey. Although it is avoided to ask questions with textual answers as much as possible in the surveys, it is not always possible to ask closed questions (Schierholz, 2014). Therefore, it is inevitable to correctly classify and code textual answers in every statistical research. In this context, coding means translating a textual expression consisting of one or more words into a classification code related to a terminology.

Coding can be seen as an activity used in the statistical process. It can be considered as a very special task and is a very difficult task to do. The purpose of coding is to assign an existing text to an appropriate class. The goal in this process is to group the appropriate text in the appropriate class by selecting the most suitable among the many classes corresponding to the text written (Hacking & Willenborg, 2012). Coding is also very similar to how a doctor diagnoses a patient who presents him with various complaints and symptoms. The doctor's job is to diagnose and choose a treatment on this basis, based on several observations, patient responses, and possibly additional tests (e.g. blood tests) (Hilden & Habbema & Bjerregaard, 1978a, 1978b, 1978c). With a similar approach, the coding process is the process of assigning the most suitable code for textual recognition in the light of the information available.

Method

It is the Household Budget Survey (HBS) is the main data source of consumption expenditures, which is applied to households determined monthly by Turkey Statistical Institute (TURKSTAT). In this survey, the answer to the question of *What is the name, type and detailed description of the spending?* is compiled from households. The conversion of these expenditure definitions into Classification of Individual Consumption According to Purpose (COICOP) codes is carried out by the interviewer. COICOP is the international classification of consumption expenditures.

The Code Assignment System (KASIS) has been developed, which analyses the accuracy of the given COICOP codes. The system outputs the conclusion that the coding made by the interviewer is *correct*, *suspicious*, *inadequate* and *accepted*. Another feature of KASIS is whether it can find the most appropriate code for textual recognition, whether it is coded or not, based on the classification dictionary.

The purpose of this study and the main motivation source is to evaluate the code assignment performance of KASIS, which has been developed as a solution to the problems in coding and classification. Although the developed system is available for all classifications with standard classification dictionary, this study is only on COICOP classification.

In this context, firstly interviewer codes were checked in the 2016-2018 HBS micro data sets by KASIS. Then, the performance of the system was tested by comparing the results with the codes assigned by the interviewer, by reassigning the records classified as inadequate and suspicious by two different methods with fuzzy matching techniques.

Jaro-Winkler distance algorithm is widely used in the record linkage to calculate the similarity between the two texts. Other popular methods are Jaro, Levenshtein, generalized edit distance and n-gram distance algorithms (Ariel, 2014). These algorithms were used to calculate the similarity between the definitions written by the interviewer and the definitions in the classification dictionary, and the most appropriate code was determined based on these calculated results.

The first of the methods used in code assignment is to narrow down the code list that may be an alternative to the words in the definition, and then to assign code to this collapsed list using fuzzy matching algorithms. The other is to make code assignment using the fuzzy matching algorithms to the wide list of all definitions directly in the dictionary without narrowing the list of possible codes.

Findings

Considering three years together, the insufficient code rate in all records is 3%. These rates are 5.3% in 2016, 3% in 2017 and 0.9% in 2018. The suspicious code rate among all records is 0.1%. These rates are 0.2% in 2016, 0.2% in 2017 and 0% in 2018.

In the code assignment method from the collapsed list; KASIS and the interviewer reach the same result in 29% of the records. This rate remains at 15.6% in code assignment method from the wide list. The accuracy of the code assigned by the interviewer is 68% for the first method and 38.4% for the second method.

The number of three, four or five different codes suggestion were made by the system is 12% of the records in the first method and 45.3% in the second method. In these records, it was concluded that the correctness of the interviewer coding should be confirmed with expert opinion or one of the codes suggested by KASIS should be selected by looking at the codes proposed by KASIS with five different fuzzy matching algorithms.

The total number of records in 2016-2018 controlled by KASIS is 6723593. In the first method, it was concluded that the expert opinion will come into play in 04% of these records and 1.4% in the second method. Considering which record is correctly classified, which record is classified incorrectly and how to handle these records by the system; the system has performed an impossible task that can be done by manual methods.

Discussion and Conclusion

Classifications are tools that provide a common language for harmoniously compiling, processing, comparing, presenting and analyzing data over time. Since the textual expression received from the person who responded to the survey was converted into codes during the conversion of the human factor, the assigned classification code may be incorrect even if the received textual expression is correct. Therefore, the results should not be evaluated only with the initiative of the code-giver, it is necessary to check whether the given codes are given in accordance with the classification dictionary and to analyze the results. Performing this control manually will not be very productive considering timeliness, cost and quality with the increase in the number of records to be controlled.

Today, surveys can be done on paper (PAPI), computer assisted personal interview (CAPI), computer assisted telephone interview (CATI) and computer assisted web interview (CAWI). This study will improve the coding quality, especially coding consistency, coding precision, reduce survey costs and reduce the interview load that the interviewer creates during coding, by confirming the accuracy of the coding of the statistical classification codes coded by interviewers or coders, no matter what survey method is used.

It was demonstrated that the first method applied to the narrowed list of these methods gave better results. However, in cases where you cannot narrow the list according to the definitions in the dictionary, code verification using fuzzy matching algorithms must always be a method to be applied.

Giriş

Sınıflama, etimolojik anlamda belirlenmiş standartlara göre karşılık gelen nesneyi tanımlayan ve Yunanca *clasis* kelimesinin Latinceye uyarlanması sonucu ortaya çıkan bir terimdir (Simões, Freitas and Rodríguez-Bravo, 2016). Sınıflamalar, farklı coğrafyalarda yaşayan ve farklı diller konuşan insanlar arasında iletişimi sağlayan ortak bir dildir. Bu yüzden belirli standartları sağlamaları zorunlu olmaktadır. Bu sayede, aynı standartlara sahip sınıflamalara bağlı olarak üretilen istatistikler ülkeler arasında karşılaştırabilir olma özelliğini kazanmaktadır. Bu kapsamda oluşturulmuş çok sayıda sınıflama bulunmaktadır. Faaliyet sınıflamaları, ürün sınıflamaları, dış ticaret sınıflamaları, amaca göre sınıflamalar, coğrafi sınıflamalar, çevre sınıflamaları, eğitim sınıflamaları, sağlık sınıflamaları ve meslek sınıflamaları bu sınıflamalardan bazılarıdır (TÜİK, 2006).

Sınıflamalar, alanlarında uzmanlaşmış Birleşmiş Milletler İstatistik Bölümü (UNSD), Avrupa Birliği İstatistik Ofisi (Eurostat), Uluslararası Çalışma Örgütü (ILO), Uluslararası Para Fonu (IMF), İktisadi İşbirliği ve Kalkınma Teşkilatı (OECD), Birleşmiş Milletler Eğitim, Bilim ve Kültür Örgütü (UNESCO) gibi uluslararası kuruluşlar tarafından geliştirilmektedir (TÜİK, 2006). Sınıflamalar ile ilgili üretilecek istatistiklerden her kullanıcının aynı bilgiyi elde edebilmesi ve istatistiklerin karşılaştırılabilir olması için ilgili sınıflama konusunda uzman uluslararası kuruluşlar el kitapları ve ayrıntılı dokümanlar ile sınıflamalar hakkında düzenli olarak yönlendirmeler yapmaktadır. Bu yönlendirmeler içerisinde soruların ankette nasıl sorulması gerektiğinden sınıflamaların nasıl kullanılması gerektiğine kadar her türlü bilgi bulunmaktadır.

Sınıflamaların kullanıldığı istatistikler idari kayıtlardan elde edilebileceği gibi anket yapılarak da üretilebilmektedir. Anketlerde, cevap veren kişi sınıflama konusunda uzman olmadığı için sınıflamalar ile ilgili seçenekli sorular sormak her zaman mümkün olmamaktadır (Schierholz, 2014). Ankette alınan cevap bir sınıflamadaki koda dönüştürülecek ise cevabın metinsel olarak alınması zorunlu olmaktadır. Alınan metinsel tanıma, ilgili sınıflamadaki en uygun kod anketör tarafından veya kodlayıcı tarafından atanmaktadır. Bu yüzden, sınıflama kullanılan her istatistiksel araştırma veya ankette metinsel tanımlara en uygun kodu atamak kaçınılmaz bir iş olmaktadır. Kod atama süreci, metinsel tanımın sınıflamadaki en uygun koda dönüştürülme işlemlerinden oluşmaktadır (Hacking and Willenborg, 2012).

Verilen eğitimlere ve yönlendirmelere rağmen anketör hatalı kod atama potansiyeline her zaman sahip olmaktadır. Metinsel ifadelerin kodlara dönüştürülmesi esnasında insan faktörü devreye girdiği için alınan metinsel ifade doğru olsa ve bu ifade doğru kodu bulmaya yetecek kadar ayrıntıya sahip olsa bile atanan kod hatalı olabilmektedir. Anketör tarafından kod ataması yapılmış kayıtlarda kodlama hatası yapıldığını belirleyebilmek için ciddi bir çalışma yapılması gerekmektedir. Anketör tarafından kod ataması yapılmış kayıtların doğruluğu, TÜİK' te gözle manuel olarak sağlanmaktadır. Ancak, kayıt sayısı arttıkça bu işlemin bu yöntemle verimli bir şekilde yapılabilmesi mümkün değildir. Bunun için gerek daha önce kod ataması yapılmamış kayıtlara otomatik bir şekilde kod ataması yapabilecek gerekse daha önce anketör tarafından kod ataması yapılmış kayıtların doğruluğunu kontrol edebilecek bir sisteme ihtiyaç duyulmaktadır.

Ulusal ve uluslararası alanda standart bir sınıflamanın kullanılması karşılaştırılabilir istatistik üretmek için gerekli olmaktadır. Bunun yanında her istatistiğin üretim sürecinin insandan bağımsız yöntemler ve otomasyon ile yapılması maliyet ve doğruluk bakımından önemlidir. Kodlama faaliyetini yerine getirmek için farklı yöntemler mevcuttur: manuel kodlama, bilgisayar destekli kodlama ve otomatik kodlama. Kodlama faaliyeti için hangi seçeneğin en iyi olduğu kodlamanın karmaşıklığına bağlı olmakla birlikte bu sistemlerin kombinasyonları da uygulamada kullanılabilmektedir (Schierholz, 2014). Otomatik kodlama, herhangi bir müdahale gerektirmeyen bir kodlama algoritması anlamına gelmektedir.

Clarke ve Brooker (2011), doğrudan insan katılımı olmadan ve bilgisayar tarafından yapılan metinsel bir tanıma kod atama işlemini otomatik kodlama olarak tanımlamaktadır. Bilgisayar programı hangi kodun metne en uygun olduğunu seçebilmelidir. Bu aşamada, yazılı metinle ilgili aşağıdaki sorunlar ortaya çıkabilmektedir (Hacking and Willenborg, 2012):

1. Yazım sorunları
2. Dil bilgisi problemleri (kelimeler arasındaki ilişkiler, söz dizilimleri)

3. Anlamsal problemler (kelimelerin anlamı, kavramlar, cümle parçaları, tek bir cümle, birkaç cümle)
4. Yorumlama problemleri (sınıflamadaki hangi kod metne en iyi şekilde uyar).

Metinle ilgili diğer bir sorunda, sınıflama açısından bakıldığında, bir metnin eksik ifadeler içermesi veya cevabın iki veya daha fazla kodla ilgili olmasıdır. Bu sorun, metnin beklenenden daha ayrıntılı ifade içermesi nedeniyle de ortaya çıkabilmektedir (Hacking and Willenborg, 2012).

Otomatik kodlamada kullanılacak sistemin tüm bu sorunlara çözüm bulması beklenmektedir. Sınıflama probleminin çözümü, son yıllarda makine öğrenmesi ve veri madenciliğinin önemli çalışma alanlarından biri olmuştur (Aggarwal and Zhai, 2012).

Bethmann vd. (2014), Alman panel araştırmalarında otomatik meslek kodlaması için iki tür olasılıksal denetimli makine öğrenme algoritması uygulamış ve eğitim verisi olarak yaklaşık 300000 adet manuel kodlanmış meslek kodu ve tanımlarını kullanmışlardır. Yazarlar, algoritmanın girdi bilgisi olarak kullanılan eğitim verisinin kaliteli olması durumunda, meslek kodlarının otomatik olarak yüksek başarı ile kodlanabileceği sonucuna varmışlardır.

Belloni vd. (2016), meslek kodlarındaki hataları incelemek amacıyla Hollanda'daki Avrupa'da Sağlık, Yaşlanma ve Emeklilik Anketi (SHARE) verilerindeki son ve şu anki mesleğe ilişkin açık uçlu sorulara verilen cevapları bir program kullanarak yeniden kodlamışlardır. Daha önce yapılan kodlamada hatalar tespit etmişlerdir. Anketlerde, kodlama kalitesinin önemli olduğunu ve genellikle ihmal edildiğini belirtmişlerdir. Yanlış kodlamalar nedeniyle hataların istatistiksel analizler yapılırken veya ekonometrik modellerde dikkate alınması gerektiğine vurgu yapmışlardır.

TÜİK tarafından aylık olarak belirlenen hanelere uygulanan ve tüketim harcamalarının ana veri kaynağı olan Hanehalkı Bütçe Araştırması'nda (HBA), *Harcamanın adı, cinsi ve ayrıntılı tanımı nedir?* sorusuna cevap olarak yazılan tanıma uygun Bireysel Tüketimin Amaca Göre Sınıflaması (Classification of Individual Consumption According to Purpose - COICOP) kodları anketörler tarafından verilmektedir. COICOP, tüketim harcamalarının uluslararası sınıflamasıdır (TÜİK, 2006). Anketör tarafından ataması yapılan bu COICOP kodlarının doğruluğunu analiz ederek bunun sonucunda anketör tarafından yapılan kodlamanın *doğru*, *şüpheli*, *yetersiz* ve *kabul* olduğu sonucunu çıktı olarak veren ve bu kayıtlara yeniden kod ataması yapan SAS 9.3 W32_7PRO platformunda istatistiksel analiz için kullanılan bir bilgisayar programlama dili olan SAS dilinde Kod Atama Sistemi (KASİS) geliştirilmiştir.

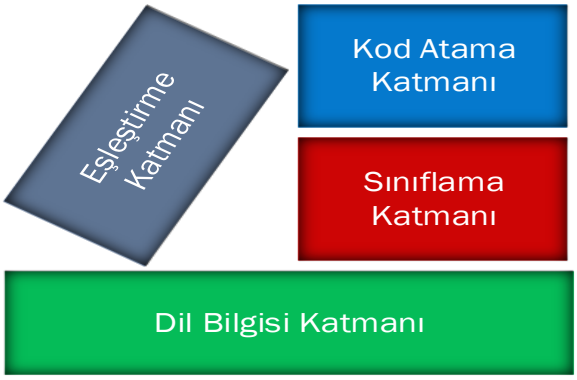
Bu çalışmanın amacı ve temel motivasyon kaynağı, kodlama aşamasında yaşanan sorunlara çözüm olarak geliştirilen KASİS' in kayıtları eşleşen, yetersiz, şüpheli ve kabul olarak sınıflaması ve bu kayıtlara kod atama performansını değerlendirmektir. Geliştirilen sistem, standart sınıflama sözlüğü olan tüm sınıflamalar için kullanılabilir olmakla birlikte bu çalışma sadece COICOP sınıflaması üzerinedir. Bu kapsamda, çalışmada öncelikle 2016-2018 yılında TÜİK tarafından uygulanan HBA mikro veri setlerindeki anketör tarafından kod ataması yapılan kayıtlar KASİS tarafından kontrol edilmiştir (TÜİK, 2016, 2017, 2018). Daha sonra sistemin yetersiz ve şüpheli olarak sınıflanan kayıtlara iki farklı yöntemle bulanık eşleştirme teknikleri kullanılarak yeniden kod ataması yapılmış ve sistemin atadığı kodlar ile anketörün atadığı kodlar karşılaştırılarak sistemin kod atama performansı değerlendirilmiştir. Bu çalışmada, kod ataması yapılan kayıtların doğru sınıfta yer alıp almadığının kontrolünü gerçekleştirilecek ve yanlış sınıfta yer alan kayıtlara uygun kod ataması yapabilecek bir sistemin tanıtımı ve performansı değerlendirilmiştir. Metinsel tanım kullanılarak kodların otomatik olarak atanabileceği gösterilerek literatüre bu alanda katkı sağlanmıştır.

Yöntem

Bu bölümde öncelikle geliştirilen KASİS' in katmanları tanıtılmıştır. Daha sonra kod atama katmanında kullanılan kayıt bağlantısı yöntemlerinden bulanık eşleştirme ve algoritmaları anlatılmıştır. Son olarak kod atama katmanında kullanılmak üzere iki farklı yöntem açıklanmıştır.

Katmanlar

KASİS' in, Şekil 1'de görüleceği üzere dört farklı katmanı bulunmaktadır.



Şekil 1. KASİS' in katmanları

Dil Bilgisi Katmanı: Bu katmanların birincisi, dil bilgisi katmanıdır. Bu katmanda anketör tarafından yazılmış harcama tanımı büyük harfe dönüştürülmekte ve Türkçe karakterlerden arındırılmaktadır. Aynı işlemler sınıflama sözlüğündeki tanımlara da uygulanarak sözlükte yer alan ve anketör tarafından yazılmış harcama tanımları kelimelerine ayrılmaktadır.

Eşleştirme Katmanı: Katmanlardan ikincisi eşleştirme katmanıdır. Bu katmanda kelimelerine ayrılmış harcama tanımı ile kelimelerine ayrılmış sınıflama sözlüğündeki tanımlar karşılaştırılmaktadır. Anketörün yazdığı harcama tanımının kelimelerinin sözlükteki hangi kod veya kodların tanımının kelimelerini maksimum olarak içerdiği bulunmaktadır.

Sınıflama Katmanı: Katmanlardan üçüncüsü, sınıflama katmanıdır. Bu katmanda maksimum kelime eşleşmeleri bulunan harcama tanımları ve sözlük tanımları sınıflanmaktadır. Burada dört durum ortaya çıkmaktadır:

1. *Eşleşen kod:* Anketörün yazdığı harcama tanımındaki kelimeleri maksimum sayıda içeren COICOP sınıflama sözlüğündeki tanım tekse ve bu tanıma karşılık gelen kod anketörün atadığı harcama kodu ile aynıysa kayıt *eşleşen* olarak sınıflanmaktadır.
2. *Şüpheli kod:* Anketörün yazdığı harcama tanımındaki kelimeleri en çok içeren COICOP sınıflama sözlüğündeki tanım tekse ve bu tanıma karşılık gelen kod anketörün atadığı harcama kodu ile farklıysa kayıt *şüpheli* olarak sınıflanmaktadır.
3. *Yetersiz kod:* Anketörün yazdığı harcama tanımındaki kelimeleri içeren COICOP sınıflama sözlüğündeki tanım birden fazlaysa kayıt *yetersiz* olarak sınıflanmaktadır.
4. *Kabul kod:* Anketörün atadığı kod sistem tarafından önerilen alternatif kodların arasında varsa, anketörün atadığı kod ile alternatif olarak önerilen kodların ilk beş basamağı aynı ve anketörün tanımı ile verilebilecek alternatif kod sayısı beşten azsa kayıt *kabul* olarak sınıflanmaktadır.

Kod Atama Katmanı: Katmanlardan sonuncusu, kod atama katmanıdır. Bu katmanda, sınıflaması yetersiz ve şüpheli olarak yapılmış kayıtlara bulanık eşleştirme teknikleri ile kod ataması yapılmaktadır.

Kayıt Bağlantısı

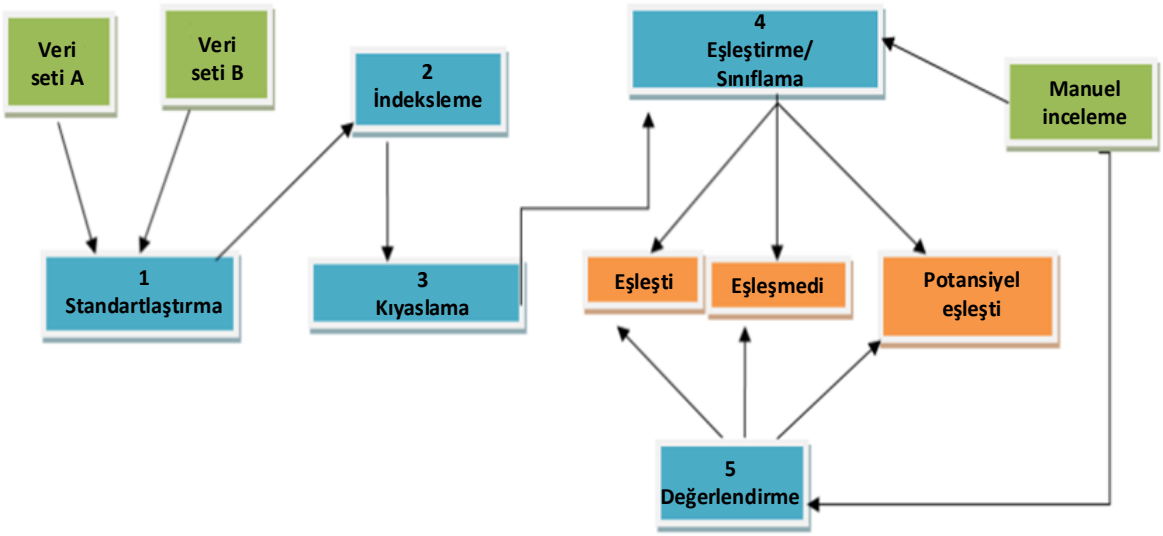
Kayıt bağlantısı, bir veya daha fazla veri tabanından gelen kayıtların aynı kayıt olduğuna karar verme işlemidir (Wright, 2011). Kayıt bağlantısı, bu çalışma kapsamında anketörün yazdığı tanım ile standart COICOP sınıflama sözlüğündeki tanımın aynı olup olmadığına veya eşleşip eşleşmediğine karar verilmesi aşamasında kullanılmaktadır.

Kayıt bağlantısı, nesne tanımlama (Tejada, Knoblock and Minton, 2001), veri temizleme (Do and Rahm, 2001), yaklaşık bağlantı veya yaklaşık birleştirme (Gravano et al., 2001), bulanık eşleştirme (Ananthakrishna, Chaudhuri and Ganti, 2002) ve varlık çözümlemesi (Benjelloun, Garcia-Molina, Su and Widom, 2005) kavramları ile de tanımlanmaktadır.

Temel olarak deterministik ve olasılıksal olmak üzere iki tür kayıt bağlantı yöntemi bulunmaktadır (Statistics Canada, 2016). Deterministik veya olasılıksal yöntem seçimi, bağlantı değişkenlerinin mevcudiyeti ve kalitesine bağlıdır. Değişkenler yüksek kalitede olduğunda, olasılıksal yöntemlere göre deterministik bir yaklaşım sıklıkla tercih edilmektedir (Gu, Baxter, Vickers and Rainsford, 2003). Daha düşük bir veri kalitesi olduğunda, olasılıksal bir yaklaşım sıklıkla seçilmektedir (van Herk-Sukel, 2012).

Uygulamada, veri hacmi arttıkça deterministik ve olasılıksal yöntemlerin kombinasyonu da birlikte kullanılmaktadır (Ariel, 2014). Bulanık eşleştirme yöntemi genellikle benzersiz bir tanımlayıcı mevcut olmadığında veya değişkenler yetersiz kalitede olduğunda kullanılmaktadır. Yöntem, adını Fellegi ve Sunter tarafından geliştirilen olasılık çerçevesinden almakta ve bu yöntemde hesaplamaları gerçekleştirmek için gelişmiş bir yazılım gerekmektedir (Statistics Canada, 2016).

Şekil 2'de, çoğu istatistik ofisi tarafından kullanılan bulanık eşleştirme sürecinin şematik gösterimi yer almaktadır. Uygulamada, iki veri tabanının bağlanması zorunlu değildir, aynı anda birden fazla veri tabanı da bağlanabilmektedir (Statistics Canada, 2016).



Şekil 2. Bulanık eşleştirme süreci

Şekil 2'ye göre; kayıt bağlantısında otomatik bir yöntem kullanılıyor olsa bile sistemin performansını test etmek amaçlı olarak son adım her zaman manuel bir inceleme süreci olmaktadır. Jaro-Winkler mesafe algoritması, iki metin arasındaki benzerliği hesaplamak için olasılıksal kayıt bağlantısında yaygın olarak kullanılmaktadır. Diğer popüler yöntemler; Jaro, Levenshtein, genelleştirilmiş düzenleme mesafesi ve n-gram mesafe algoritmalarıdır (Ariel, 2014).

Anketör tarafından yazılan harcama tanımları ile sınıflama sözlüğündeki tanımlarının benzerliğini hesaplamak için bu algoritmalar kullanılmış ve ataması yapılacak en uygun kod hesaplanan bu algoritma sonuçlarına göre belirlenmiştir.

Kullanılan Yöntemler

Kod atama katmanında, daraltılmış listeden kod atama ve geniş listeden kod atama olmak üzere iki farklı yöntem kullanılmıştır.

Daraltılmış Listeden Kod Atama: Bu yöntemde; KASİS' in dilbilgisi, eşleştirme ve sınıflama katmanlarından geçirilerek yetersiz ve şüpheli olarak sınıflanmış olan kayıtlara kod atama katmanında kod ataması yapılmıştır. Sınıflama katmanında yetersiz veya şüpheli olarak sınıflanmış kayıtlar için eşleştirme katmanında kaç tane alternatif kodun ve tanımın karşılık geldiği ve bunların neler olduğu sistem tarafından bilinmektedir. Bu yöntemde bulanık eşleştirme algoritmaları eşleştirme katmanında belirlenen bu kod ve tanımlar ile sınırlı kalmak koşuluyla uygulanmıştır.

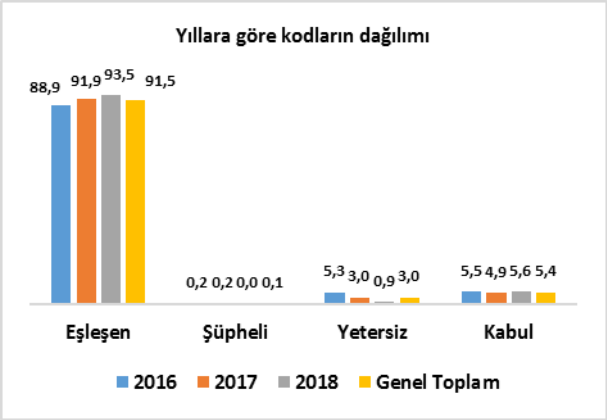
Geniş Listeden Kod Atama: Bu yöntemde; KASİS' in dilbilgisi, eşleştirme ve sınıflama katmanlarından geçirilerek yetersiz ve şüpheli olarak sınıflanmış olan kayıtlara kod atama katmanında kod ataması yapılmıştır. Birinci yöntemin aksine bu yöntemde bulanık eşleştirme algoritmaları, listede daraltma uygulanmadan sözlükte bulunan tüm kod ve tanımlar için uygulanmıştır.

Bulgular

Kod ataması yapılırken kullanılabilecek iki değişik yaklaşım bulunmaktadır. Bu yaklaşımlardan ilki, elimizdeki tüm kayıtlara tekrar kod ataması yapmak ve bu kodlar ile anketörün atadığı kodları kıyaslamaktır. Bu yaklaşımın seçilmesi durumunda anketör tarafından kodlaması doğru olarak yapılmış kayıtlar için gereksiz yere kod ataması yapılmış olacaktır. Bu çalışmada, bu yöntem tercih edilmemiştir

Bu çalışmada tercih edilen diğer yaklaşımda ise, kod ataması doğru olan kayıtları diğerlerinin içerisinden ayırarak kalan kayıtlara kod ataması yapmaktır. Bunun için öncelikle anketör tarafından hangi kaydın hatalı olarak kodlandığının bilinmesi gerekmektedir. Bu anlamda, KASİS iyi bir çözüm sağlamaktadır.

Anketör tarafından kod ataması yapılmış kayıtlar KASİS aracılığıyla kontrol edilmiş ve kayıtların eşleşen, yetersiz ve şüpheli olarak sınıflamaları yapılmıştır. 2016-2018 yıllarındaki toplam kayıt sayısı 6723593'tür. Yıllara göre elde edilen sonuçlara Şekil 3'te yer verilmiştir.



Şekil 3. Yıllara göre kodların dağılımı

Şekil 3'e göre; eşleşen ve kabul olarak sınıflanmış kayıtlar dikkate alındığında tüm kayıtların %96,9'unda anketör kodlaması ile aynı sonuca ulaşılmıştır. Diğer bir ifadeyle, anketör kodlamasının doğruluğu teyit edilmiştir. Tüm kayıtlar içerisinde yıllara göre yetersiz ve şüpheli kodların yüzdesel dağılımına Çizelge 1'de yer verilmiştir.

Tablo 1. Yıllara göre yetersiz ve şüpheli kodların yüzdesel dağılımı

Yıl	Yetersiz	Şüpheli	Toplam
2016	5,3	0,2	5,5
2017	3,0	0,2	3,2
2018	0,9	0,0	0,9
Toplam	**3,0**	**0,1**	**3,1**

Tablo 1'e göre; yetersiz ve şüpheli kodların toplamına bakıldığında en başarılı yıl %0,9 ile 2018 yılı olurken en başarısız yıl %5,5 ile 2016 yılı olmuştur. Bu sonuçlara bakarak, anketörler tarafından tanımların her geçen yıl daha iyi yazıldığını ve doğru kodlama yapıldığını söyleyebiliriz. Sınıflaması yetersiz ve şüpheli olarak yapılmış kayıtlara, anketörün yaptığı kodlamanın doğruluğunu teyit edebilmek veya yeni bir kod önerebilmek amacıyla anketörün yazdığı tanımdan yola çıkarak yeniden kod ataması yapılmıştır. Yetersiz ve şüpheli olarak sınıflaması yapılan kayıt sayısı 212056'dır. KASİS ile kod ataması yapılırken iki ayrı yöntemde beş farklı bulanık eşleştirme algoritması kullanılmıştır. Elde edilen sonuçlar karşılaştırmalı olarak Tablo 2'de verilmiştir.

Tablo 2. İki farklı yönteme göre yapılan KASİS kod atama sonuçları

Açıklama	Daraltılmış listeden kod atama	%	Geniş listeden kod atama	%
Kod ataması yapılan kayıt sayısı	212056	100,0	212056	100,0
Anketörün atadığı 10 basamaklı kodun doğru olduğu kayıt sayısı	61512	29,0	33126	15,6
Anketörün atadığı 10 basamaklı kodun yanlış olduğu kayıt sayısı	150544	71,0	178930	84,4
KASİS tarafından 1 kod önerisi yapılan kayıt sayısı	124967	58,9	82891	39,1
KASİS'in önerdiği kodun ilk beş basamağı ile anketörün atadığı kodun ilk beş basamağının aynı	*82785*	*39,0*	*48255*	*22,8*
KASİS'in önerdiği kodun ilk beş basamağı ile anketörün atadığı kodun ilk beş basamağının farklı	*42182*	*19,9*	*34636*	*16,3*
KASİS tarafından 3 kod önerisi yapılan kayıt sayısı	20928	9,9	39671	18,7
KASİS tarafından 4 kod önerisi yapılan kayıt sayısı	4332	2,0	51064	24,1
KASİS tarafından 5 kod önerisi yapılan kayıt sayısı	317	0,1	5304	2,5

Tablo 2'ye göre; daraltılmış listeden kod atama yönteminde, kayıtların %29'unda, KASİS ile anketör aynı sonuca ulaşmıştır. Bu oran, geniş listeden kod atama yönteminde %15,6'da kalmıştır. Bu kayıtlarda anketörün atadığı 10 basamaklı COICOP kodlarının doğruluğu teyit edilmiştir.

Anketörün atadığı 10 basamaklı COICOP kodlarının hatalı olduğu kayıtların oranı, ilk yöntem için %71 iken ikinci yöntem için %84,4'dür. Bu kayıtlar için, KASİS' in tek kod önerdiği kayıt oranı ilk yöntemde %58,9 iken ikinci yöntemde bu oran %39,1'de kalmıştır.

Anketörün atadığı 10 basamaklı kodun hatalı olduğu kayıtlar incelendiğinde; ilk yöntemde kayıtların %39'unda KASİS' in önerdiği kodun ilk beş basamağı ile anketörün atadığı kodun ilk beş basamağının aynı olduğu görülmüştür. İkinci yöntemde, bu oran %22,8 seviyesinde kalmıştır. Bu kayıtlarda anketörün atadığı kodun doğru olarak kabul edilebileceği sonucuna ulaşılmıştır.

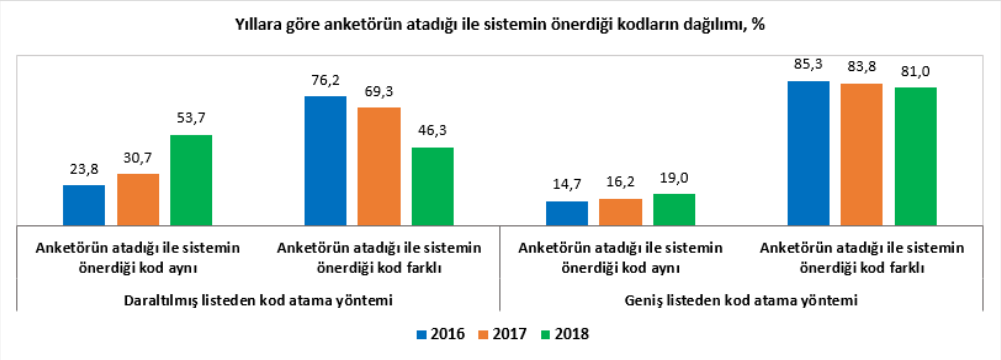
Bu sonuçlar neticesinde, anketörün atadığı kodun doğruluğu sistem tarafından teyit edilen kayıtların oranı ilk yöntem için %68 (%29+%39), ikinci yöntem için %38,4 (%15,6+%22,8) olmuştur.

Sistemin önerdiği kodun ilk beş basamağı ile anketörün atadığı kodun ilk beş basamağının farklı olduğu kayıt sayısı ilk yöntemde 42182 ve ikinci yöntemde 34636'dır. Bu kayıtlarda KASİS' in önerdiği kodun kullanılması önerilmektedir.

İlk yöntemde kayıtların %12'sine ve ikinci yöntemde %45,3'üne sistem tarafından üç, dört veya beş farklı kod önerisi yapılmıştır. Bu kayıtlarda, KASİS' in beş farklı bulanık eşleştirme algoritması ile önerdiği kodlara bakılarak anketör kodlamasının doğruluğunun uzman görüşü ile onaylanması gerektiği veya KASİS tarafından önerilen kodlardan birisinin seçilmesi gerektiği sonucuna varılmıştır.

KASİS ile kontrol edilen 2016-2018 yıllarındaki toplam kayıt sayısı 6723593'tür. İlk yöntemde bu kayıtların %04'ünde ve ikinci yöntemde %1,4'ünde uzman görüşünün devreye gireceği sonucuna ulaşılmıştır.

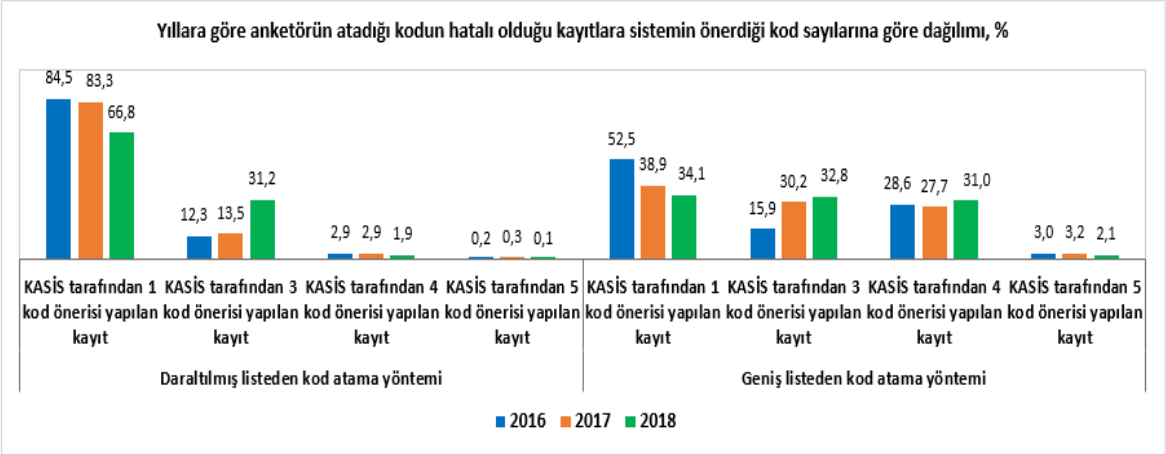
Yıllara göre anketörün atadığı kod ile sistemin önerdiği kodların dağılımı Şekil 4'te verilmiştir.



Şekil 4. Yıllara göre anketör atadığı ile sistemin önerdiği kodların dağılımı

Şekil 4'e göre, 212056 kayda uygulanan daraltılmış listeden kod atama yönteminde anketörün atadığı kod ile uyum oranı en yüksek yıl %53,7 ile 2018 yılı olmuştur. Uyum oranı bakımından ikinci yıl %30,7 ile 2017 ve üçüncü yıl %23,8 ile 2016 yılı olmuştur. Geniş listeden kod atama yönteminde, başarılı yıl sıralaması değişmemiştir. %19 ile 2018 yılı birinci, %16,2 ile 2017 yılı ikinci ve %14,7 ile 2016 yılı üçüncü yıl olmuştur.

Yıllara göre anketörün atadığı kodun hatalı olduğu kayıtlara sistemin önerdiği kod sayılarına göre dağılımı Şekil 5'te verilmiştir.

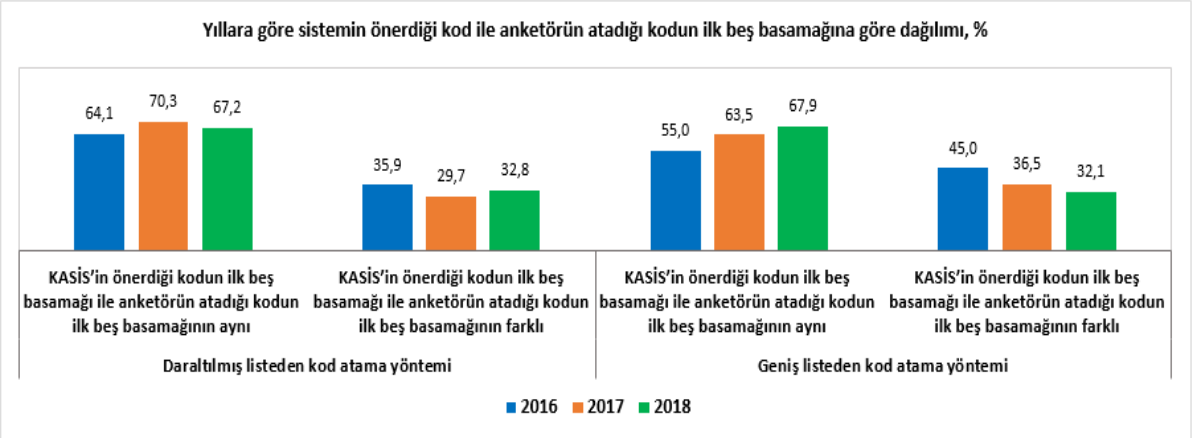


Şekil 5. Yıllara göre anketörün atadığı kodun hatalı olduğu kayıtlara sistemin önerdiği kod sayılarına göre dağılımı

Şekil 5'e göre, daraltılmış listeden kod atama yönteminde bir adet kod önerisi yapılan oransal olarak en yüksek yıl %84,5 ile 2016 yılı olmuştur. İkinci yıl %83,3 ile 2017 ve üçüncü yıl %66,8 ile 2018 yılı olmuştur. Geniş listeden kod atama yönteminde, başarılı yıl sıralaması değişmemiştir. %52,5 ile 2016 yılı birinci, %38,9 ile 2017 yılı ikinci ve %34,1 ile 2018 yılı üçüncü yıl olmuştur.

Sistemin 3 farklı kod önerisi yaptığı yüzdesel dağılımlara bakıldığında; 2017 yılında daraltılmış listeden kod atama yöntemi lehine yöntemler arasında farklılık bulunmaktadır. Aynı farklılık, sistemin 4 farklı kod önerisi yaptığı üç yılda da bulunmaktadır.

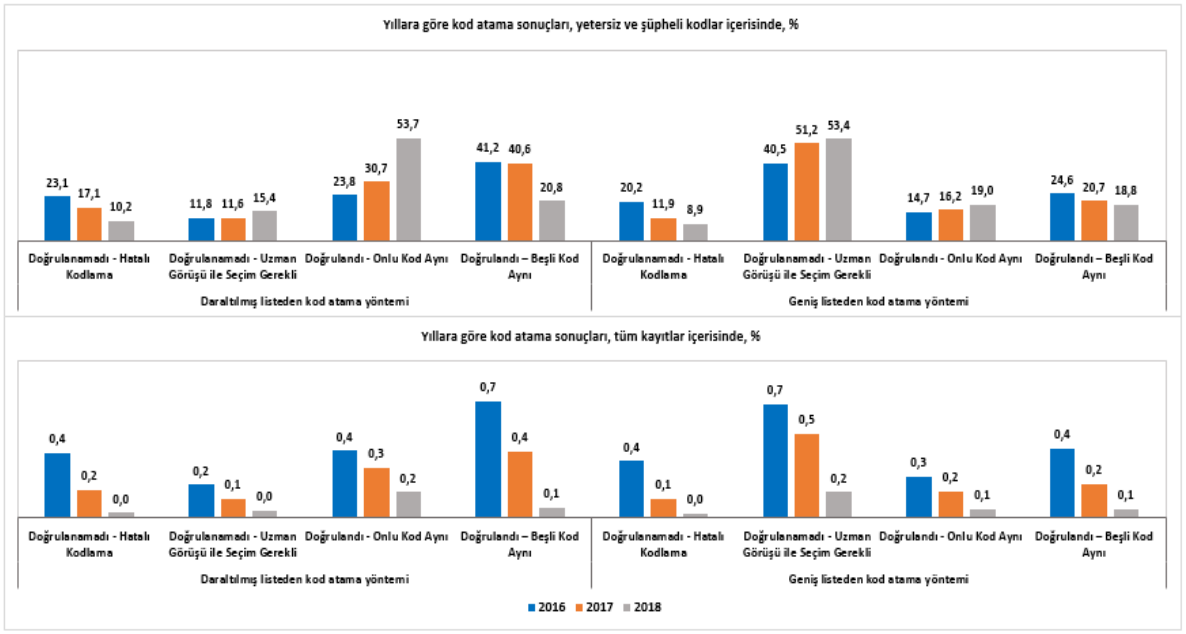
Yıllara göre sistemin önerdiği kod ile anketörün atadığı kodun ilk beş basamağına göre dağılımı Şekil 6'da verilmiştir.



Şekil 6. Yıllara göre sistemin önerdiği kod ile anketörün atadığı kodun ilk beş basamağına göre dağılımı

Şekil 6'ya göre, daraltılmış listeden kod atama yönteminde sistemin önerdiği kodun ilk beş basamağı ile anketörün atadığı kodun ilk beş basamağının aynı olduğu en yüksek yıl %70,3 ile 2017 yılı olmuştur. İkinci yıl %67,2 ile 2018 ve üçüncü yıl %64,1 ile 2016 yılı olmuştur. Geniş listeden kod atama yönteminde, %67,9 ile 2018 yılı birinci, %63,5 ile 2017 yılı ikinci ve %55 ile 2016 yılı üçüncü yıl olmuştur.

Yıllara göre yetersiz ve şüpheli kodlar ile tüm kodlar içerisindeki kod atama sonuçları Şekil7'de verilmiştir.



Şekil 7. Yıllara göre kod atama sonuçları, yetersiz ve şüpheli kodlar içerisinde ve tüm kodlar içerisinde

Şekil 7 incelendiğinde; daraltılmış listeden kod atama yöntemine göre 2016 yılında şüpheli ve yetersiz kodların %11,8'lik kısmı için uzman görüşünün gerekli olduğu sonucuna ulaşılmıştır. Bu oran; 2017 yılında %11,6 ve 2018 yılında %15,4 olmuştur. Bu kayıtlar için, KASİS' in beş farklı bulanık eşleştirme algoritması ile önerdiği kodlara bakılarak uzman görüşü ile anketör kodlamasının doğruluğunun onaylanması gerektiği veya KASİS tarafından önerilen kodlardan birisinin seçilmesi gerektiği sonucuna ulaşılmıştır.

Daraltılmış listeden kod atama yöntemine göre 2016 yılında tüm kayıtların %0,2'lik kısmı ve 2017 yılında %0,1'lik kısmı için uzman görüşüne başvurulması gerektiği sonucuna ulaşılmıştır. Geniş listeden kod atama yönteminin seçilmesi durumunda ise, ilk yöntemden biraz daha fazla kayıtta uzman görüşüne başvurulması gerektiği ortaya çıkmıştır.

Tartışma ve Sonuç

Günümüzde anketler, kâğıt üzerinde (PAPI), bilgisayar destekli yüz yüze görüşme (CAPI), bilgisayar destekli telefon görüşmesi (CATI) ve bilgisayar destekli web görüşmesi (CAWI) şeklinde yapılabilmektedir.

Ankete cevap veren kişiden alınan metinsel ifadelerin kodlara dönüştürülmesi esnasında insan faktörü devreye girdiği için alınan metinsel ifade doğru olsa bile atanan sınıflama kodu hatalı olabilmektedir. Bu bakımdan, farklı toplumlarda farklı değer yargılarına ve algılarına sahip insanların yaşadığı da göz önünde bulundurularak ve insandan uzaklaşan ve insan etkisinin olmadığı sistematik olarak çalışan bir kod atama sisteminin geliştirilmesi önem arz etmektedir. Bu yüzden sadece kod verenin inisiyatifi ile sonuçlar değerlendirilmemeli, verilen kodların sınıflama sözlüğüne uygun olarak verilip verilmediğinin kontrol edilmesi ve sonuçlarının analiz edilmesi gerekmektedir. Bu kontrolün manuel olarak yapılması kontrol edilecek kayıt sayısının artması ile birlikte harcanacak zaman, oluşacak maliyet ve elde edilecek kalite düşünüldüğünde verimli olmayacaktır.

Otomatik kodlamada genellikle bir listedeki kayıtlara kodlama yapılmaktadır. Bu türden bir algoritma başarılı bir şekilde çalıştığında, atanan kod daima tek bir kod olmaktadır. Otomatik kodlama uygulamalarında genellikle denetimli makine öğrenmesi yöntemleri kullanılmaktadır. Denetimli makine öğrenmesinin başarısı büyük ölçüde bağımsız değişkenlerin tahmin gücüne ve eğitim veri setinin boyutuna bağlıdır. Yani, eğitim veri setinin hacmi ne kadar büyük ise modelin öğrenmesi o denli iyi olmaktadır (Keogh, 1995). Bu tip uygulamalarda modelin iyi öğrenebilmesi için kullanılacak eğitim veri setinin de doğruluğunun teyit edilmesi gerekmektedir. Aksi halde, yanlış öğrenen model yanlış sonuçlar üretecektir. Geliştirilen bu sistemin diğer sistemler üzerindeki en önemli üstünlüğü eğitim veri setine ihtiyaç duymamasıdır. Sistem, sıfır noktasından itibaren kayıt sayısı arttıkça öğrenmesini de artırmaktadır. Bu sistemin, diğer sistemlere göre üç üstünlüğü bulunmaktadır. İlki, bu sistem makine öğrenmesinde kullanılacak eğitim veri setinin doğruluğunun teyit edilmesinde ve temizlenmesinde kullanılabilir. İkincisi, bu sistem diğer sistemlerin yaptığı gibi otomatik kod atama sistemidir. Sonuncusu ise, bu sistemin eşleşen olarak sınıfladığı kayıtlar bir başka denetimli makine öğrenme uygulamasında eğitim veri seti olarak kullanılabilir.

Bu çalışma, hangi anket yöntemi kullanılırsa kullanılsın anketörler veya kodlayıcılar tarafından kodlanmış istatistiksel sınıflama kodlarının doğruluğunu teyit ederek veya gerektiğinde kodlamanın düzeltilmesini sağlayarak kodlama kalitesini, özellikle de kodlama tutarlılığını, kodlama hassasiyetini artıracak, anket maliyetlerini düşürecek ve anketörün kodlama esnasında oluşturduğu görüşme yükünü azaltacaktır.

Çalışma kapsamında daha önce anketör tarafından kodlaması yapılarak sonuçları TÜİK tarafından kamuoyu ile paylaşılmış olan 2016-2018 yılları arasındaki veri üzerinde sistemin etkinliği test edilmiştir. Anketör tarafından kodlamanın doğruluğunun sistem tarafından teyit edilemediği kayıtlara iki farklı yöntem kullanılarak tekrar kod ataması yapılarak sonuçlar ortaya koyulmuştur. Buradaki amaç tanıma göre sistem tarafından kod ataması şüpheli bulunan veya tanımı kod ataması yapabilecek kadar ayrıntı içermeyen kayıtları veri setinden dışlamadan önce araştırmacılara alternatif bir yöntem sunmanın yanında bulunan bu eksiklikleri sonuçları resmi olarak açıklanmış verileri düzeltebilmek için bir yöntem önermektir.

Kod atamasında kullanılan yöntemlerden ilki önce tanımda yer alan kelimelere göre alternatif olabilecek kod listesini daraltmak daha sonra daraltılmış bu listeye bulanık eşleştirme algoritmalarını kullanarak kod ataması yapmaktır. Diğeri alternatif olabilecek kod listesini daraltmadan direkt olarak sözlükte yer alan tüm tanımların olduğu geniş listeye bulanık eşleştirme algoritmalarını kullanarak kod ataması yapmaktır.

Bu yöntemlerden daraltılmış listeye uygulanan ilk yöntemin daha iyi sonuç verdiği ortaya konulmuştur. Ancak sözlükte yer alan tanımlara göre listeyi daraltamadığınız durumlarda bulanık eşleştirme algoritmalarını kullanarak kod ataması veya doğrulamasının yapılması her zaman başvurulması gereken bir yöntem olmak zorundadır.

Sistem tarafından hangi kaydın doğru sınıflandığı, hangi kaydın yanlış sınıflandığı ve yanlış sınıflanan bu kayıtlara nasıl bir işlem yapılacağı bilgisinin verildiği düşünüldüğünde; sistem manuel yöntemlerle yapılabilmesi imkânsız bir görevi yerine getirmektedir. Otomatik kodlama programlarının performansını karşılaştırmak için genellikle kullanılan kriterler; verimlilik, güvenilirlik ve hız olmaktadır. Ancak, bu kriterler mutlak değildir (Riviere, 1995). KASİS ile kayıtların %96,9'una anketör ile aynı kod ataması yapılarak bu kayıtlarda yapılan kodlamanın doğru olduğu sonucuna varılmıştır.

Roessingh ve Bethlehem (1983), aile harcama anketinde üç farklı yöntemle otomatik kod ataması gerçekleştirerek ve %94, %85, %78 oranlarında doğru kodlamaya ulaşmışlardır.

Yeni Zelanda İstatistik Ofisi, Census 2013 verilerinde meslek ve okul sonrası yeterlilik değişkenlerini kodlamak için SVM algoritmasını kullanmışlardır. Her iki değişken için test verilerinde %50 doğruluk oranına ulaşmışlardır (Chu and Poirier, 2015).

Tourigny ve Moloney (1997), 1991 Kanada Nüfus Sayımı' nda yer alan yedi farklı değişken için yapılan otomatik kodlama sonucunda %92'lik doğruluk oranına ulaşmışlardır.

Haslinger (1997), Avusturya Nüfus Sayımı' ndaki çalışılan yer değişkenine %96, eğitim değişkenine %92 ve iktisadi faaliyet değişkenine %50 oranında otomatik kodlama ile kod ataması yapabilmiştir.

Otomatik kodlama sisteminin doğru kodu atama performansı, kullanılan sınıflamanın ve verinin karmaşıklığına bağlı olarak değişebilmektedir. Ancak genel bir perspektif sunabilmesi açısından, KASİS' in kod atama performansı benzer sistemler ile karşılaştırıldığında başarılı olduğu sonucuna ulaşılmış ve KASİS' in atadığı kodlar ile anketörün atadığı kodlar karşılaştırıldığında aralarında bir fark olmadığı belirlenmiştir. Anketörün kod ataması yapması yerine otomatik yöntemlerle kod ataması yapılması önerilmektedir. Bu sayede, kodlama konusunda anketörlerin üzerindeki iş yükü azalmış olacaktır.

Çalışma kapsamında, TÜİK tarafından kamuoyu ile paylaşılmış olan veri üzerinde sistemin etkinliği test edilmiştir. Bunun yerine, KASİS' in direkt olarak istatistik üretim aşamasında kullanılması önerilmektedir. KASİS' in üretim aşamasında kullanılması sonucunda, anket sonuçları kamuoyu ile paylaşılmadan önce kodlamada yapılan hata ve eksiklikler giderilmiş, kodlama kalitesi artırılmış ve sonucunda maliyetlerde azalma sağlanmış olacaktır.

Acknowledgements / Teşekkür ve Bilgilendirme

Makale, Levent Ahi tarafından hazırlanan ve henüz savunması gerçekleştirilmemiş olan *Uluslararası İstatistiksel Sınıflamalara Yönelik Kod Atama Sistemi (KASİS)* isimli doktora tezinden yararlanılarak hazırlanmıştır. / The article has been prepared using the doctorate thesis entitled *Code Assignment System (KASIS) for International Statistical Classifications* prepared by Levent Ahi and has not been defended yet.

Research Ethics / Yayın Etiği Bildirimi

Yazarlar araştırmanın etik dışı bir sorunu bulunmadığını, araştırma ve yayın etiği konusunun gözlemlendiğini beyan etmektedir. / The authors declare that the research does not have an unethical problem and that research and publication ethics are observed

Contribution Rate of Researchers / Araştırmacıların Katkı Oranı

Yazarlar, çalışmanın her aşamasında yer almışlardır. / The authors took part in every stage of the study.

Conflict of Interest / Çıkar Çatışması

Yazarlar çalışmanın herhangi bir çıkar çatışması olmadığını ifade etmektedir. / The authors state that the study has any conflicts of interest.

Funding / Fon Bilgileri

Yazarlar bu çalışma için herhangi bir fonları olmadığını beyan etmektedir. / The authors declare that there is no funding for this study.

Kaynakça/References

Aggarwal, C. C., Zhai, C. A. (2012). *Survey of Text Classification Algorithms, Mining Text Data*, Springer, Boston, MA, 163-222.

Ananthakrishna, R., Chaudhuri, S., Ganti, V. (2002). Eliminating Fuzzy Duplicates in Data Warehouses. Paper presented at Proceedings of the Very Large Databases Conference.

Ariel, A., Bakker, B., de Groot, M., Grootheest, G., Laan, J., Smit, J., Verkerk, B. (2014). Record linkage in health data: a simulation study, Netherlands Central Bureau of Statistics.

Belloni, M., Brugiavini, A., Meschi, E., Tijdens, K. (2016). Measuring and detecting errors in occupational coding: an analysis of share data. *Journal of Official Statistics*, 32(4), 917-945.

Benjelloun, O., Garcia-Molina, H., Su, Q., Widom, J. (2005). Swoosh: A Generic Approach to Entity Resolution, Stanford University technical report, Stanford.

Bethmann, A., Schierholz, M., Wenzig, K., Zielonka M. (2014). Automatic Coding of Occupations Using Machine Learning Algorithms for Occupation Coding in Several German Panel Surveys. Paper presented at Proceedings of Statistics Canada Symposium, Canada.

Chu, K., Poirier, C. (2015). Machine learning documentation initiative (Canada), Workshop on the Modernisation of Statistical Production, Switzerland.

Clarke, F. R., Brooker S. (2011). Use of Machine Learning for Automated Survey Coding. Paper presented at International Statistical Institute Proceedings of the 58th World Statistics Congress, Dublin.

Do, H.H., Rahm, E. (2001). COMA – A system for flexible combination of schema matching approaches. Paper presented at Proceedings of the Very Large Databases Conference.

Gravano, L., Ipeirotis, P. G., Jagadish, H. V., Koudas, N., Muthukrishnan, S., Srivastava, D. (2001). Approximate String Joins in a Database (Almost) for Free. Paper presented at Proceedings of the Very Large Databases Conference.

Gu, L., Baxter, R., Vickers, D., Rainsford, C. (2003). Record linkage: Current practice and future directions, Commonwealth Scientific and Industrial Research Organisation. *Mathematical and Information Science*, 3.

Hacking, W., Willenborg, L. (2012). Theme: Coding; Interpreting Short Descriptions Using a Classification, The Hague/Heerlen: Statistics Netherlands, 4-11.

Haslinger. A. (1997). Automatic Coding and Text Processing using N-grams. In Conference of European Statisticians. Statistical Standards and Studies – No. 48. Statistical Data Editing, Volume No. 2, Methods and Techniques, pages 199-209. UNO, New York and Geneva.

Hilden, J., Habbema, J.D.F and Bjerregaard, B. (1978a). The measurement of performance in probabilistic diagnosis, I. The problem, descriptive tools, and measures based on classification matrices. *Methods of information in medicine*, 17, 217-226.

Hilden, J., Habbema, J.D.F and Bjerregaard, B. (1978b). The measurement of performance in probabilistic diagnosis, II. Trustworthiness of the exact values of the diagnostic probabilities. *Methods of information in medicine*, 17, 227- 237.

Hilden, J., Habbema, J.D.F and Bjerregaard, B. (1978c). The measurement of performance in probabilistic diagnosis, III. Methods based on continuous 54 functions of the diagnostic probabilities. *Methods of information in medicine*, 17, 238-246.

Keogh, G. (1995). Automatic Coding of Occupations: The Irish Experience, New Techniques and Technologies for Statistics II, Bonn.

Riviere, P. (1995). Outline of a theory of automated coding, Paper presented at Conference of European Statisticians, Athens.

Roessingh, M., Bethlehem, J. (1983). Trigram coding in the family expenditure survey in statistics, Netherlands Central Bureau of Statistics.

Schierholz, M. (2014). Automating Survey Coding for Occupation. Yüksek Lisans Tezi. Ludwig Maximilians Universitat Institut fur Statistik, Munchen, 70.

Simões, M.d.G., Freitas, M. C. V. d., Rodríguez-Bravo, B. (2016). Theory of classification and classification in libraries and archives: Convergences and divergences. *Knowledge Organization*, 43(7), 530-538.

Statistics Canada Reports on Special Business Projects an Overview of Selected International Business Record Linkage Programs. (2016). Erişim adresi: https://www150.statcan.gc.ca/n1/pub/18-001-x/18-001-x2016001-eng.htm, Son Erişim Tarihi: 03.05.2020.

Tejada, S., Knoblock, C., Minton, S. (2001). Learning Object Identification Rules for Information Extraction. *Information Systems*, 26 (8), 607-633.

Tourigny, J. Y., Moloney, J. (1997). Statistical Data Editing Volume No. 2 Methods and Techniques, United Nations Statistical Commission and Economic Commission for Europe, New York and Geneva.

Türkiye İstatistik Kurumu Hanehalkı Bütçe Anketi Mikro Veri Seti, CD. (2016). Ankara.

Türkiye İstatistik Kurumu Hanehalkı Bütçe Anketi Mikro Veri Seti, CD. (2017). Ankara.

Türkiye İstatistik Kurumu Hanehalkı Bütçe Anketi Mikro Veri Seti, CD. (2018). Ankara.

Türkiye İstatistik Kurumu Sınıflama Sunucusu. (2006). Erişim adresi: https://biruni.tuik.gov.tr/DIESS/, Son Erişim Tarihi: 10.05.2020.

Türkiye İstatistik Kurumu Sınıflama Sunucusu Amaca Göre Sınıflamalar. (2006). Erişim adresi: https://biruni.tuik.gov.tr/DIESS/SiniflamaSurumListeAction.do?turId=5&turAdi=%204.%20Amaca%20G%C3%B6re%20S%C4%B1n%C4%B1flamalar&guncel=Y, Son Erişim Tarihi: 03.05.2020.

Türkiye İstatistik Kurumu Sınıflama Sunucusu Sınıflama Türleri. (2006). Erişim adresi: https://biruni.tuik.gov.tr/DIESS/SiniflamaTurListeAction.do, Son Erişim Tarihi: 03.05.2020.

van Herk-Sukel, M. P., Lemmens, V. E., van de Poll-Franse, L., Herings, R. M., Coebergh, J. W. (2012). Record linkage for pharmacoepidemiological studies in cancer patients. *Pharmacoepidemiology and Drug Safety*, 21, 94–103.

Wright, G. (2011). Probabilistic Record Linkage in SAS. Paper presented at Proceedings of Western Users of SAS Software, California.