Original Article

대한영양사협회 학술지
J Korean Diet Assoc 25(1):44-58, 2019
https://doi.org/10.14373/JKDA.2019.25.1.44

기계학습방법을 활용한 대형 집단급식소의 식수 예측: S시청 구내직원식당의 실데이터를 기반으로

전종식 · 박은주[1] · 권오병†
경희대학교 경영학과 · [1]경남대학교 식품영양학과

Predicting the Number of People for Meals of an Institutional Foodservice by Applying Machine Learning Methods: S City Hall Case

Jongshik Jeon · Eunju Park[1] · Ohbyung Kwon†
School of Management, Kyung Hee University, Seoul 02447, Korea
[1]Dept. of Food & Nutrition, Kyungnam University, Changwon 51767, Korea

ABSTRACT

Predicting the number of meals in a foodservice organization is an important decision-making process that is essential for successful food production, such as reducing the amount of residue, preventing menu quality deterioration, and preventing rising costs. Compared to other demand forecasts, the menu of dietary personnel includes diverse menus, and various dietary supplements include a range of side dishes. In addition to the menus, diverse subjects for prediction are very difficult problems. Therefore, the purpose of this study was to establish a method for predicting the number of meals including predictive modeling and considering various factors in addition to menus which are actually used in the field. For this purpose, 63 variables in eight categories such as the daily available number of people for the meals, the number of people in the time series, daily menu details, weekdays or seasons, days before or after holidays, weather and temperature, holidays or year-end, and events were identified as decision variables. An ensemble model using six prediction models was then constructed to predict the number of meals. As a result, the prediction error rate was reduced from 10%～11% to approximately 6～7%, which was expected to reduce the residual amount by approximately 40%.

Key words : institutional foodservice, meal forecasting, classification model, machine learning, big data analytics

본 논문은 박사학위 논문 중 일부임.
This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education(NRF-2017S1A3A2066740).
접수일 : 2019년 1월 7일, 수정일 : 2019년 1월 22일, 채택일 : 2019년 1월 23일
† Corresponding author : Ohbyung Kwon, School of Management, Kyung Hee University, 26 Kyungheedae-ro, Dongdaemun-gu, 02447 Seoul, Korea
Tel : 82-2-961-2148, Fax : 82-2-961-0515, E-mail : obkwon@khu.ac.kr, ORCID : http://orcid.org/0000-0001-9686-6586

서 론

예측은 과거 발생한 일에 대해서 이와 관련된 변수의 알려진 과거값 및 데이터를 기반으로 하여 우리가 알고 싶어 하는 미래에 발생할 수 있는 사건을 미리 알고자 하는 것으로 정의된다(Makridakis 등 1998). 기업뿐만 아니라 조직이나 단체의 운영, 계획, 전략 등과 같은 여러 주요한 조직 관리 기능들 중에서 모든 활동에 선행되는 예측은 데이터를 기반으로 하기 때문에 조직 운영과 실무적인 차원에서 경영 및 활동을 하는데 있어서 기초가 되는 중요한 의사결정의 활동이라고 할 수 있다.

기숙사・학교・공장・사업장・후생기관 등에서 특정한 사람들을 대상으로 지속적으로 음식을 공급하는 집단급식에서는 정확한 식수 예측이 발주, 구매, 조리, 배식 관리에 있어서 매우 중요한 의사결정요인이다(Kim 2014b; Ministry of Food and Drug Safety 2016). 집단급식소에서의 현장에 맞는 식수 예측 알고리즘의 개발 및 적용을 통한 정확한 식수 예측은 집단급식소의 미배식 잔반의 감소를 통한 원가 절감, 노동 시간 절감, 조리 작업 업무의 효율화 등의 효과가 있으며, 집단급식소의 시설 투자 및 운영에 있어서도 여러 가지 긍정적인 영향을 미친다(Chung 2001). 그러나 많은 집단급식소에서는 제한된 예산 등의 이유로 체계적인 식수 예측이 어려워 잔식 발생, 음식 및 서비스 품질 저하와 음식 품절이 발생되고, 이로 인해 민원 및 생산 과부족으로 인한 비용이 발생하여 고객 불만의 상승 등으로 이어지고 있다(Kim 2014a; Hur 2017).

집단급식소의 식수 예측 대다수가 경험에 의한 예측에 의존하는 경향이 있어 관공서, 산업체, 병원, 대학교 집단급식소를 중심으로 실무에 적용할 수 있는 수준의 식수 예측 모델의 필요성이 계속 제기되어왔으나(Chung 2001; Lim 2008; Lim 2016), 기존의 식수 예측 모델은 실무적으로 적용할 수 있는 정도로 성과가 나오는 식수 예측 모델과는 거리가 멀었다. 또한 집단급식소의 식수 예측은 병원, 학교, 산업체, 관공서, 공공기관 등 유형에 따라 영향을 주는 요인이 다르므로 기존 연구결과를 그대로 적용하기 어렵고(Cullen 등 1978; Miller 등 1991; Lin 등 1992), 정확한 식수 예측 알고리즘에 대한 현장에서의 필요성(Miller & Shanklin 1988; Repko & Miller 1990; Miller 1990)에도 불구하고 충분한 연구가 이루어지지 않고 있는 실정이다(Kim 2000; Chung 2001; Lim 2008).

한편, 기계학습은 통상적으로 데이터 수집, 알고리즘에 의한 학습, 테스트용 데이터 셋을 활용한 알고리즘 성능 측정 등 3단계로 이루어지고 있는데, 각 단계에서의 알고리즘과 모델링의 결합을 통해서 모델의 성능을 보완하고 향상시킴으로써 예측 목적에 맞게 진행을 하여 이를 통해 여러 응용 영역에서 성공적으로 적용되기 시작하고 있다(Lee 등 2016). 그러나 현재까지 기계학습기법을 사용하여 집단급식소의 식수 예측 모델을 수립하여 실증 검증한 것은 거의 전무한 상황이다(Park 2017).

더욱이 우리나라 한식 식단 구성은 서구식과는 달리 복잡하고 경우의 수가 너무 많아서 한식 식단 메뉴를 회귀 모델링 등 전통적인 통계기법을 통해서 모델링을 하는 데에는 많은 제약이 발생한다. 또한 식수에 영향을 미치는 변수가 너무 많고 상황에 따라 다르기 때문에 실제적으로 식수 예측 모델링을 진행하는데 있어서 예상치 못한 경우가 발생하며 데이터 확보의 어려움 등 많은 제약요소가 발생하고 이를 반영해야 하기 때문에 전통적인 통계기법으로 진행을 하는 경우 양질의 실데이터를 기반으로 예측 모델링을 진행하는데 있어서 많은 제약과 애로사항이 있다.

이에 실제적으로 식수 예측에 관련 있을 것으로 보이는 실제 데이터를 S시로부터 제공받아 청사 내 집단급식소의 식수 예측 모델링을 진행하게 되었다. 연구대상인 S시청의 집단급식소에서는 연간 약 11만 L의 잔반이 발생하고 있고, 이 잔반을 줄이기 위해 경험을 통한 식수 예측에서 진일보한 데이터 기반의 식수 예측 모델이 필요한 상황이어서 실데이터를 기반으로

청사 내 집단급식소의 잔반량 감소를 위해 기계학습을 활용한 식수 예측 모델을 연구 개발하였다. 또한 본 연구를 기반으로 영양사 등 현장 실무자들의 경험을 통해 식수 예측이 진행되고 있는 대학교, 산업체, 공공기관, 병원 등의 집단급식소에서 데이터 과학에 기반한 맞춤형 식수 예측 기법이 확산되는 데에 기여하고, 실데이터를 기반으로 한 집단급식소 맞춤형 식수 예측 모델링 제공을 통해 집단급식소에서의 식수 예측과 관련된 실무적인 활용도를 높일 수 있는 기틀을 마련하고자 한다.

이에 본 연구의 목적은 집단급식소에서 발생하는 잔식을 포함하여 잔반을 경감하는 집단급식소의 식수 예측 기법을 제안하고, 실제 사례를 통해 그 기법의 활용 가능성을 보이는 것이다. 이를 통해 메뉴를 기획하는 단계와 음식 조리를 위한 재료 발주 단계에서부터 적합한 식수를 예상하고, 잔식과 잔반을 줄이는 등 집단급식소 운영의 효율성을 제고하고자 한다. 이를 위해 모집한 실제 집단급식소의 데이터를 기반으로 회귀분석 등을 통해 어떠한 변수가 식수 예측에 유의한 영향을 미치는지 파악한 후에, 기계학습기법을 응용한 식수 예측 모델을 제안하였다.

연구방법

1. 전체적인 구조

본 연구에서는 실제 집단급식소에서 적용 가능한 식수 예측 방법을 제안하기 위해 S시에서 제공받은 데이터와 날씨 데이터로 식수 예측 모델링 연구를 진행하였으며, 본 연구의 식수 예측 모델링 프로세스는 Fig. 1과 같다. 식수 예측을 위해서 먼저 식사 가능 인원 변수, 시계열 인원 변수, 메뉴 특성 변수, 요일 변수, 전후 휴일 여부 변수, 날씨 변수, 연휴 연말 변수, 이벤트 변수 등 8개 카테고리 63개의 변

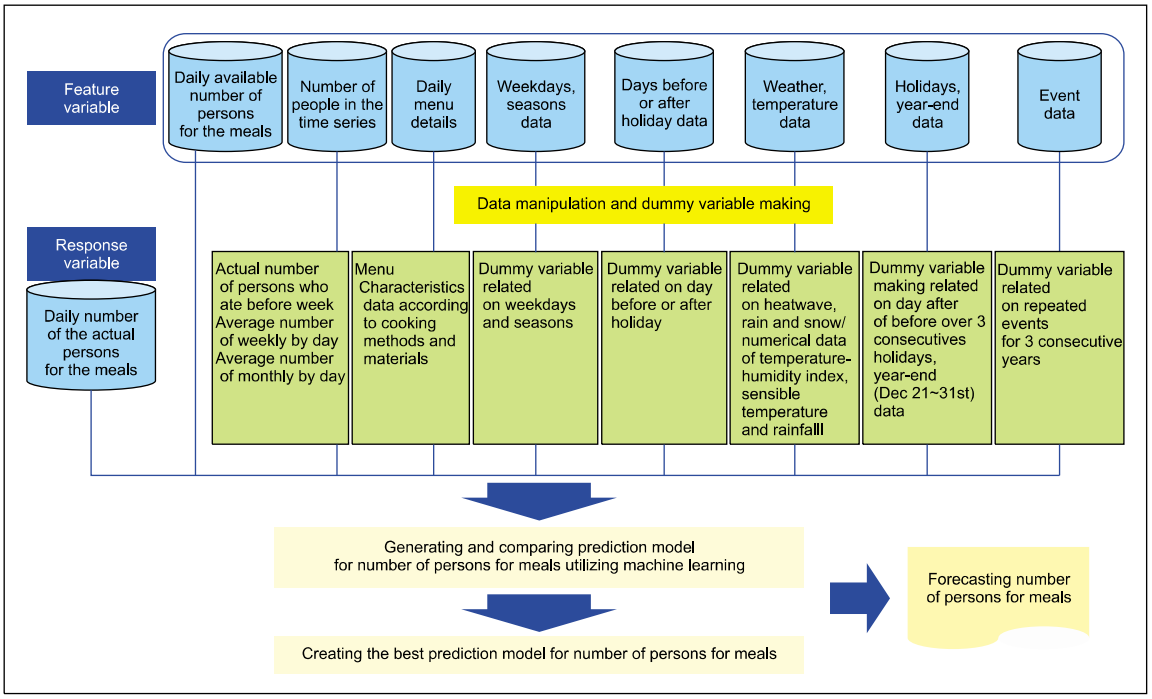

Figure 1. Research model for forecasting number of persons for the meals.

수를 선정하였다. 다음으로 63개의 변수 관련 데이터를 전처리한 후에 6개의 회귀분석 및 기계학습기법을 활용하여 식수 예측 모델을 생성하였다. 이후 생성한 모델에 대해서 실제 식수와 식수 예측치와의 오차율 비교를 수행한 후 확보한 3년치의 데이터 중 70%에 해당하는 데이터를 학습세트(Training Set)로, 나머지 30%에 해당하는 데이터를 검증세트(Test Set)로 진행하였다. 마지막으로 6개의 모델을 비교 평가한 결과 중에서 성능이 우수한 일부 모형의 앙상블로 식수 예측 모델을 최종 생성하였다.

2. 데이터

본 연구는 식수 예측 모델링을 위해 S시청 구내식당에서 제공한 메뉴, 식사 가능 인원 및 날씨 등에 대한 데이터를 대상으로 연구를 진행하였다. 먼저 문헌 연구를 통해 식수 예측에 영향을 미치는 요인을 파악하고자 하였다. 문헌 연구에 따르면(Lim 2016) 산업체 집단급식소에서 식수 예측 시 고려하는 요인으로 요일, 제공 메뉴 선호도, 날씨, 전주 동일 요일 식수, 운영 끼니, 고객사 행사, 복수 메뉴의 선호도, 전월 식수 자료, 계절, 동일 메뉴 과거 식수, 이벤트, 전년 해당 월(주) 식수, 기온, 주변 식당 이용 가능성, 동일 메뉴 과거 잔반량 등이 식수 예측 향상을 위한 요인으로 파악되었다. 이러한 문헌 연구를 토대로 식수 예측에 영향을 미치는 변수들을 1차적으로 파악함으로써 식수 예측에 영향을 미치는 요인을 파악하고, 이를 통해 보다 중요한 변수를 확인하고자 S시청 구내식당에서 제공한 메뉴, 식사 가능 인원 등의 실제 데이터 3년치(S시 제공, 2015년 5월 1일~2018년 4월 30일 기준)와 날씨 등의 실제 데이터 3년치(기상청 제공, 2015년 5월 1일~2018년 4월 30일 기준)를 대상으로 중요 변수를 파악하였으며, 이를 바탕으로 식수 예측 모델링을 진행하였다(Table 1).

3. 구성 변수

식수 예측에 미치는 변수 파악을 위해 크게 내부적 요인과 외부적 요인의 2가지 요인으로 구분을 하였으며, 사용한 변수들에 대한 내용을 요약하면 Table 2와 같다. 내부적 요인으로 고려를 할 변수로는 식사 가능 인원(출근 인원, 출장자, 휴가자 등), 시계열 인원 변수(전주 식사 인원, 요일별 평균 식사 인원, 월 평균 식사 인원, 일일 식사 가능 인원), 메뉴 특성(일일 메뉴) 등 3가지 범주를 고려하여 분류를 하였다. 외부적 요인으로는 요일 및 계절, 전일・익일의 휴일 여부, 날씨, 이벤트 여부 및 연휴・연말 등 5가지 범주로 분류를 하였다. 또한 식수 예측 모델링을 위해 S시와 기상청에서 제공받은 변수와 특성추출

Table 1. Data used for forecasting modeling.

Division	Data provider	Explanation	Summary	Period
Menu details of restaurant for employees and actual number of persons for the meals	S City	S City's menu details and actual number of persons for the meals	Menu details and actual number of persons for the meals	May 1st 2015~ April 30th 2018 (3 years)
Daily available number of persons for the meals	S City	Daily available number of persons for the meals (number of the attendance, number of business trip persons, number of off-duty persons)	Date, daily available number of persons for the meals	
Weather data	Korea meteorological administration	Precipitation type, rainfall, wind speed, humidity, temperature	Date and time, place, precipitation type, temperature, wind speed, humidity	

(Feature Engineering)에 의한 변수 생성 등을 통해 적용한 데이터를 요약 정리하면 Table 3과 같다.

S시청 구내식당의 당일 식사 가능 인원을 예측하기 위해 S시로부터 제공받은 과거 3년치의 출근 인원, 출장자, 휴가자 등을 고려한 식사 가능 인원의 데이터를 사용하였으며, 산출하는 식은 다음과 같다.

Table 2. The summary of the variables used for model.

Feature variable		Variable explanation	Response variable
Internal variables	Daily available number of persons for the meals	Number of the attendance, number of business trip persons, number of off-duty persons	Daily number of the actual persons for the meals
	Number of people in the time series	Actual number of persons who ate 1 week before, average number of persons who ate weekly, average number of persons who ate monthly by day (the variable related on number of persons with constant cycle and characteristics)	
	Daily menu characteristics	Menu characteristics according to cooking methods and materials	
External variables	Weekdays, seasons	Monday, tuesday, wednesday, thursday, friday, spring, summer, autumn, winter	
	Days before or after holiday	Whether the following day or the previous day is a holiday or not	
	Weather	Heatwave, rain, snow, temperature-humidity index, sensible temperature and precipitation	
	Events	Repeated events for 3 consecutive years	
	Holidays, year-end	Day after or before over 3 consecutives holidays, year-end (Dec 21st～31st)	

Table 3. The summary of data used for model.

Division	Data provider	Explanation	Feature summary
Actual number of persons for meals in S City's restaurant for the employees	S City	S City's status of restaurant for employees	Date, weather, temperature, menu, number of persons for the meal, leftover
Number of the attendance, number of business trip persons, number of off-duty persons	S City	Daily number of available persons	Date, number of the attendance, number of business trip persons, number of off-duty persons
Time series personnel variable	Feature engineering	The variable related on number of persons with constant cycle and characteristics	Actual number of persons who ate 1 week before average number of persons who ate weekly average number of persons who ate monthly
Menu characteristics data	S City	3 years lunch menu details	Menu details
Days before or after Holiday variable	Variable by feature engineering	Whether the following day or the previous day is a holiday or not	Weekdays related on days before or after holiday
Weather data	Korea meteorological administration	Temperature, precipitation type (rain/snow/sleet)	Temperature, heatwave, rain, snow, temperature-humidity index, sensible temperature and precipitation
Event variable	Variable by feature engineering	Repeated events for 3 consecutive years	Any one of the three "dog days", free rice cake soup, the 15th of January by the lunar calendar etc.
Holidays, year-end variable	Variable by feature engineering	Day after or before over 3 consecutives holidays, year-end (Dec 21st～31st)	Day after or before over 3 consecutives holidays, year-end (Dec 21st～31st)

식사 가능 인원=본청 출근 인원-행정국 예외 직원-본청 출장자-본청 휴가자

다음으로 식수 예측에 영향을 미치는 변수를 파악하기 위해 현장 조사를 수행하였다. S시 청사 내 집단급식소의 경우 영양사 및 조리사는 경험에 의해 식수를 예상하고 있었는데, 요일별 식수를 주요 결정요인으로 인식하고 있었다. 식수는 요일 별로 일정한 패턴이 있는 것으로 알려져 있다.

본 연구에서 사용한 시계열 인원 변수는 전주 식사 인원, 요일별 식수와 월별 식수 변수를 특성추출기법으로 새로 생성하였으며, 요일별 식수는 요일평균으로, 월별 식수는 월평균으로 산출하여 요일평균과 월평균 변수를 사용하여 분석과 식수 예측 모델링을 진행을 하였다.

특성추출기법에 의한 새로운 변수 생성과 관련하여 요약한 결과는 다음과 같다.

1) 요일별 식수

현장에서 경험치로 식수를 예측함에 있어 중요한 요인으로 삼고 있는 것은 요일별 식수였다. 이는 현장에서 요일별로 다르게 식수를 파악하고 식사 준비를 진행하고 있다는 점에 착안하여 본 연구에서도 요일별 식수 변수를 생성하여 진행하였다.

2) 전주 식수

요일별 식수 변수 외에 식수 예측 모델링의 정확도에 기여할 추가 변수를 확인하고자 지난주와 같은 요일의 식수를 파악하여 전주 식수라는 변수를 생성하여 진행한 결과 식수 예측 모델링의 정확도가 향상되었다. 요일별 식수와 요일별 전주 식수를 동시에 고려함으로써 모델링의 오차율을 줄이고 정확도를 더욱 높일 수 있는 계기를 마련하였다.

3) 월별 식수

1년 중 월별로 식수가 항상 고르게 분포하는 것이 아니라 연말과 같이 월별로 식수가 달라 같은 요일이라 하더라도 월에 따라 다른 요일이 있다는 것을 파악하여 월별 식수라는 변수를 생성하였다. 요일별 식수, 전주 식수와 함께 월별 식수 변수를 추가적으로 진행함으로써 식수 예측 모델링의 설명력이 더욱 향상되었다.

한편, 본 연구에서는 듀이 십진분류법(Dewey Decimal Classification)에 의한 수정 개정안을 바탕으로 메뉴를 크게 주식, 부식 및 밑반찬, 후식 및 간식 등으로 나누고(Chung & Choi 2011), 메뉴에 있는 개별 음식을 조리 재료와 조리 방법 2가지로 분류하여 1차적으로 조리 방법에 의한 분류를, 2차적으로는 재료에 의한 분류를 하여 진행하였다. 한식 메뉴 분류는 조리 재료와 조리 방식에 따라 듀이 십진분류법에 의한 수정 개정안을 큰 틀로 한 음식 메뉴 분류를 바탕으로 실제적으로 S시청에서 제공한 메뉴를 본 연구에 적용하였다.

4. 비교 모형

본 연구에서는 741개의 관측치를 사용하여 요일, 날씨, 기온, 메뉴 등 8가지 카테고리의 63개의 변수를 사용해 R(Version 3.3.2)에 적용하여 식수 예측 모델링을 하였으며, 기계학습의 분석기법은 회귀분석과 의사결정나무기법을 사용하여 R 프로그래밍을 진행하였다. R 프로그래밍에 의한 회귀분석은 다중회귀분석법 중 후진 단계적 선택법(Backward Stepwise Regression), LASSO회귀(Least Absolute Shrinking Selector Operator Regression), 능형회귀(Ridge Regression) 등 3가지를 사용하였으며, 의사결정나무기법은 랜덤포레스트(Random Forest, RF), 배깅(Bagging), 부스팅(Boosting) 등 3가지 기계학습기법을 적용하여 각각 개발된 모델에 대한 예측력을 비교 평가하고, 비교 평가된 모델을 기반으로 향후 미래의 데이터 예측의 안정성을 위해 예측력이 높은 최적의 모델을 생성하였다.

본 연구에서 고려한 식수 예측 모형은 Table 4와 같다. 첫째, 다중회귀분석법 중 후진 단계적 선택법

은 회귀분석 중 모든 변수들을 포함한 상태에서 유의성이 낮은 변수들을 제외시키는 변수 선택법이다. 후진 단계적 선택법은 유의성과 설명력이 높은 간단한 모델을 형성하고자 할 때와 변수의 수가 많아서 유의성이 높은 변수를 선택하기 어려울 때 유용하다.

둘째, LASSO회귀란 기존의 회귀분석에 제약조건(t)을 주어 중요하지 않은 변수의 회귀계수 값을 축소시켜 0으로 만들어주는 모델이다. LASSO회귀는 변수가 많을 경우 간결함의 원리에 따라 가장 잘 설명할 수 있는 변수만을 선별하여 간단한 설명으로 만들어야 할 때 유용하며, 영향력이 낮은 변수의 계수를 0으로 만듦으로써 어떠한 변수가 모델에 중요한지 알게 되어 모델 해석력이 향상된다는 장점이 있다.

셋째, 능형회귀는 기존의 회귀분석에 제약조건(t2)을 주어 중요하지 않은 변수의 회귀계수 값을 축소시켜 0에 가깝게 만들어 주는 모델이다. 능형회귀는 변수가 많을 경우 오차를 최소화하면서도 최대한 간단한 모델로 만들 수 있으며, 적합이 되는 해(solution)에 제약조건을 부여함으로써 모델의 과적합을 방지할 수 있다는 특징이 있다.

넷째, 랜덤포레스트는 배깅에서의 모형을 의사결정나무로 이용하고 랜덤으로 특정 개수의 변수를 선택해주는 모델링 기법이다. 랜덤포레스트는 데이터를 여러 데이터 묶음들로 분할시켜 과적합 위험을 막아주고, 데이터가 많을수록 성능이 좋아지며, 변수를 랜덤하게 뽑음으로써 비슷한 모델들이 만들어지지 않도록 한다.

다섯째, 배깅은 원본 훈련 데이터를 복원 추출하여 여러 데이터 묶음으로 만들고, 각각 따로 모델링하여 모델들의 평균으로 결과를 도출하는 모델링 기법이다. 배깅은 데이터를 여러 데이터 묶음으로 분할시키면서 과적합될 위험을 막아주며, 데이터가 많을수록 성능이 좋아진다.

마지막으로 부스팅이란 예측이 잘 되지 않은 데이터가 더 많이 뽑히도록 가중치를 부여하여 잘 학습되도록 하는 기법이다. 부스팅의 특징은 이전 모델의 오류를 고려해주기 때문에 보다 더 성능이 좋아진다는 점이다.

6가지의 식수 예측 모델 비교를 위해 S시에서 제공한 3년치 데이터 중 70%에 해당하는 518개의 데이터를 학습세트로, 나머지 30%에 해당하는 223개의 데이터를 테스트세트로 진행하였다. 성능 비교는 오차율로 하였고, 오차율을 산출하는 식은 다음과 같다.

오차율 = 절대값 (실제값 - 예측값) / 실제값

일반적으로 모델링은 어떤 학습 데이터에 특정 편향을 가져오기 때문에 학습 데이터의 일부에 대

Table 4. Comparative models.

Model	Description
Backward stepwise regression	A model that excludes all variables with low significance
LASSO	A model that reduces the regression coefficient of non-important variables to 0 by applying a constraint (t) to the existing regression analysis
Ridge regression	A model that reduces the regression coefficient of non-important variables by giving a constraint (t2) to the existing regression analysis
Random forest	A model that randomly generates a large number of decision trees, deduces them, and collects the results
Bagging	A model that creates multiple sets of data and that produces an average of the modeling results each separately with the original training data which is restored and extracted
Boosting	A model that weights are given so that more unexpected data is picked up and learned

해서는 순조롭게 학습할 수 있지만, 또 다른 데이터에는 문제를 가질 수 있다. 그러므로 다양한 전문가로 구성된 팀을 만드는 것과 비슷한 원리를 이용하는 것처럼 여러 개의 약한 학습기법을 이용해 최적의 답을 찾아내는 학습기법을 앙상블(ensemble)이라고 한다. 즉, 앙상블기법이란 마치 문제 해결의 강건성을 증진시키기 위해 한 전문가의 의견을 따르지 않고 다양한 전문가로 구성된 팀을 만드는 것과 같이 어느 한 예측 알고리즘에 의존하지 않고 복수의 알고리즘을 통해 각각 산출된 결과들을 기반으로 종합적인 결론을 내리는 기법이다. 종합적인 결론을 내리는 방법은 다수결의 원칙이나 평균의 원칙 등 여러 가지가 존재한다. 앙상블기법은 여러 약한 학습기법들을 결합하여 강한 학습기법으로 만드는 개념에 기반을 두고 있다. 본 연구에서도 최종 결과를 앙상블기법으로 진행하였다.

결 과

1. 다중회귀분석

본격적인 식수 예측 모형 비교에 앞서 다중회귀모델을 적용하여 8개 범주의 선정된 변수로 이루어진 다중 회귀식이 어느 정도의 설명력을 가지는지를 파악하였다. 그 결과 Table 5와 같이 F 검정 값은 25.9이며, 조정된 결정계수 기준의 회귀식의 설명력은 66.47%이며, 다중공선성은 관찰되지 않았고 P-value도 매우 낮아(P-value: <2.2e-16) 본 다중 회귀 모형은 유의한 것으로 판단된다.

Table 5에서 보듯이 63개의 변수들로 이루어진 다중 회귀식의 회귀계수 중 식사 가능 인원, 복날, 설날, 휴일 전날, 휴일 다음날, 연말, 난류, 육류, 죽류, 덮밥 및 국밥류, 비빔밥 및 복음밥류, 국탕류, 구이류, 전류, 튀김류, 전주 식사 인원, 월평균 식사 인원 등의 변수가 의미가 있었다. 특히 3년 동안 반복적으로 시행된 이벤트 중에서 복날에 제공되는 특식과 설날에 제공되는 음식은 식수에 긍정적인 영향을 미치는 것이 유의미한 것으로 나타났다.

Table 5. Results of the multiple regression analysis.

Variable name		Estimate Std.	Error	Pr (>| t |)
Feature variable	Variable explanation			
(Intercept)		−1072.000	1198.000	−0.895
Daily available number of persons for the meals	Number of the attendance−(number of business trip persons+number of off-duty persons)	0.117	0.042	2.818**
Event	Dog days of summer	269.300	43.850	6.141***
	Free rice cake soup	69.370	64.440	1.076
	The 15th of January by lunar calendar	159.900	61.470	2.601**
Weather	Temperature-humidity index	1.739	2.045	0.85
	Sensible temperature	−2.079	2.080	−0.999
	Heatwave	20.420	19.810	1.031
	Rainy	15.630	18.140	0.861
	Snowy	−20.000	45.380	−0.441
	Rainfall	−2.571	4.500	−0.571
Seasons	Spring	−10.790	14.920	−0.723
	Summer	−9.495	21.410	−0.443
	Fall	−8.609	15.510	−0.555
	Winter	NA	NA	NA

Table 5. Continued.

Variable name		Estimate Std.	Error	Pr (>| t |)
Feature variable	Variable explanation			
Weekdays	Monday	−96.450	398.700	−0.242
	Tuesday	−79.130	246.700	−0.321
	Wednesday	−88.560	250.700	−0.353
	Thursday	−46.400	179.600	−0.258
	Friday	NA	NA	NA
Holidays, year-end	Day before holiday	−159.100	27.160	−5.857***
	Day before holiday(more than 3 days)	−47.85	28.22	0.090
	Day after holiday	91.130	26.490	3.44***
	Day after holiday(more than 3 days)	39.94	29.49	0.176
	Year-end	−178.400	27.120	−6.578***
Daily menu	Grain	−0.594	8.544	−0.07
	Soybean	−1.255	8.670	−0.145
	Egg	29.120	12.670	2.298*
	Korean jelly	−11.950	14.170	−0.843
	Fish & shell	7.477	7.020	1.065
	Meat	33.070	8.075	4.095***
	Vegetable	−7.084	4.600	−1.54
	Seaweeds	−6.389	8.777	−0.728
	Rice cakes	11.870	13.090	0.907
	Flavoring & fermented soybean products	12.530	6.731	1.862
	Kimchi	−12.000	8.347	−1.437
	Steamed bun	4.387	16.470	0.266
	Flour	7.189	9.495	0.757
	Fruits	−10.660	10.920	−0.976
	Rice	7.442	9.378	0.794
	Rice gruel	94.070	41.530	2.265*
	Bowl of rice served with topping & boiled rice served in soup	46.580	21.710	2.146*
	Bibimbab & fried rice	58.730	17.460	3.364***
	Gimbap and sushi	NA	NA	NA
	Soused seafood	NA	NA	NA
	Noodles	14.960	31.400	0.476
	Soup	−51.220	23.890	−2.144*
	Stew	−22.850	25.620	−0.892
	Meat roasted	33.660	12.270	2.744**
	Korean traditional salad	5.898	9.176	0.643
	Fried food	−2.783	10.780	−0.258
	Pickles	−25.030	19.300	−1.297
	Jeon	29.750	13.500	2.204*

Table 5. Continued.

Variable name		Estimate Std.	Error	Pr (>| t |)
Feature variable	Variable explanation			
Daily menu	Hard boiled food	12.790	11.910	1.074
	Steamed food	7.414	15.050	0.493
	Fried dish	24.750	11.080	2.233*
	Salad	17.060	12.790	1.334
	Single	−5.628	6.948	−0.81
	Milk	−12.930	20.960	−0.617
	Bread & cookie	−42.470	25.760	−1.649
	Beverage & drink	−2.551	22.460	−0.114
Number of people in the time series	Actual number of persons who ate before week	0.183	0.031	5.828***
	Average number of persons who ate weekly	1.065	1.187	0.897
	Average number of persons who ate monthly	0.498	0.090	5.542***
	F value=25.9, Adj. R Square=0.6647			

F-statistic: 25.86 on 59 and 681 DF, P-value: <2.2e-16
*P<0.05, **P<0.01, ***P<0.001

2. 성능 비교

후진 단계적 선택법, LASSO회귀, 능형회귀, 랜덤 포레스트, 배깅 그리고 부스팅의 6가지 모델링 기법을 적용하여 실제 식수와 각각 모델링 기법 예측치의 오차율을 정리하면 Table 6과 같다. 그 결과 배깅과 부스팅을 제외한 상위 4개 모형은 seed에 따라 각 모델마다 오차율이 낮은 특정 데이터 분포가 있는 것을 확인할 수 있다. Table 6에 따르면, 예를 들어 샘플링 시 seed 4, 123으로 진행을 한 경우 오차율이 제일 낮은 것은 랜덤포레스트였으며, seed 1, 3, 6인 경우 오차율이 제일 낮은 모델은 후진 단계적 선택법이었다. seed 5, 8, 9로 진행한 경우는 LASSO 회귀가 오차율이 제일 낮았으며, seed 2, 7, 10인 경우는 능형회귀가 오차율이 제일 낮은 것을 확인할 수 있었다. 이는 어떤 특정 모델이 오차율이 항상 제일 낮은 최상의 모델링을 생성하는 것은 아니라는 의미이다. 즉, 본 연구를 통해 사용되는 데이터에 따라 적용된 각 모델링의 오차율이 낮은 정도가 달라 향후 미래의 데이터를 이용해 식수 예측을 진행하는 경우 최적의 모델을 생성하기 위해 하나의 기법에 따른 특정 모델링을 적용하기에 무리가 따른다는 것을 알 수 있었다. 이를 통해 여섯 가지 기법을 활용한 앙상블을 고려했다.

3. 앙상블

하나의 기법에 의해 특정 모델링을 진행하는 경우 미래의 데이터에 대한 예측 시 오차율이 높아지는 경우가 발생할 수 있어 주의가 요구되며, 1차 모델링 결과를 기본으로 개발된 각각의 모델에 대한 예측력을 비교 평가하고, 비교 평가된 모델을 기반으로 향후 미래의 데이터 예측의 안정성을 위해 예측력이 높은 최적의 모델을 앙상블 기법으로 생성하였다. 최종 모델링은 다음의 2단계를 거쳐 진행하였다.

단계 1 : 1차 모델링을 적용한 기법 중 오차율 상위 4개 선정
단계 2 : 오차율 상위 4개 모델링의 예측값들의 평균값을 최종 모델링의 예측값으로 선정

Table 6에서 각 모델별 오차율의 평균값인 Average의 오차율 기준으로 성능이 좋은 상위 4개 모델은 LASSO 회귀(7.14%), 능형회귀(7.16%), 다중회귀분석법 중 후진 단계적 선택법(7.21%), 랜덤포레스트(7.33%) 순이었다. 여기서 식수 예측 모델의 안정화를 증진시키기 위해 상위 4개 모델의 예측치의 평균값을 산출하여 이를 최종 모델링의 예측값으로 하였으며, 이는 Table 7과 같다.

최종 모델링은 하나의 기법에 의해 특정 모델링을 진행하는 경우 미래의 데이터에 대한 예측 시 오차율이 높아지는 경우가 발생할 수 있어 1차 모델링의 결과를 기본으로 모델별로 특정 샘플링에 강한 것을 활용한 것이다. 각각 개발된 모델에 대한 예측력을 비교 평가하고 비교 평가된 모델을 기반으로 향후 미래의 데이터 예측의 안정성을 위해 예측력이 높은 최적의 모델을 생성 진행하였다.

상위 4개의 각각 모델링의 예측치들의 평균값을 최종 모델링의 예측값으로 진행한 결과 Table 7에는 제시하지 않았지만 오차율은 상위 4개 모델링의 오차율의 각각의 평균(LASSO회귀 7.14%, 능형회귀 7.16%, 후진 단계적 선택법 7.21%, 그리고 랜덤포레스트 7.33%) 최솟값인 7.14보다 낮은 7.04의 최저의

Table 6. The comparison of error rate between the prediction value and actual value according to modelling by seed number.

Modeling	Back ward[1] error rate (%)	LASSO[2] error rate (%)	Ridge[3] error rate (%)	RF[4] error rate (%)	Bagging error rate (%)	Boosting error rate (%)
Seed 1	7.55	7.59	7.63	7.65	8.12	12.10
Seed 2	7.46	7.28	7.23	7.34	7.77	11.30
Seed 3	6.97	7.15	7.18	7.58	8.01	11.80
Seed 4	7.71	7.56	7.61	7.21	7.49	12.00
Seed 5	7.42	7.33	7.40	7.57	8.20	12.90
Seed 6	6.82	6.98	6.97	7.22	7.85	11.80
Seed 7	7.00	6.74	6.69	7.17	7.78	11.50
Seed 8	7.08	6.87	6.98	7.22	7.82	11.60
Seed 9	7.14	7.06	7.20	7.57	8.03	12.00
Seed 10	7.45	7.39	7.32	7.70	8.06	13.40
Seed 123	6.68	6.56	6.57	6.45	6.86	11.00
Average	7.21	7.14	7.16	7.33	7.82	11.90

[1] Backward stepwise regression
[2] Least absolute shrinking selector operator regression
[3] Ridge regression
[4] Random forest

Table 7. Final modeling. (unit: person)

LASSO		Ridge regression		Backward stepwise regression		Random forest		Final	Actual number of persons for the meal
830		819		833		850		833	766
1183		1174		1197		1146		1175	1136
940	+	939	+	938	+	880	=	924	929
1146		1145		1145		1192		1157	1185
---		---		---		---		---	---

오차율의 평균값을 가지면서 특정 샘플링에 따라 예측치의 값이 편중되지 않는 안정적인 예측치를 가지게 되었다. 이로써 1차 모델링을 거쳐 각각 개발된 모델에 대한 예측력을 비교 평가하고 비교 평가된 모델을 기반으로 향후 미래의 데이터 예측의 안정성을 위해 예측력이 높은 최적의 모델을 생성 진행함으로써 오차율 7% 수준의 식수 예측 모델링이 가능하였다.

고 찰

본 연구는 S시청 집단급식소의 실데이터를 기반으로 식수 예측 정확도 향상에 유의한 영향을 미칠 수 있는 다양한 변수를 탐색하고 공공 데이터를 기반으로 기계학습기법을 적용함으로써, 공공기관 집단급식소에 활용 가능한 새로운 식수 예측 알고리즘을 국내 최초로 개발하여 식수 예측 모델링을 제안하고 그 성능을 분석하였다.

국내에서 식수 예측과 관련하여 그동안 진행된 연구 논문을 살펴보면 전통적인 통계 방법을 사용하여 진행을 하였으며(Kim 2000; Chung 2001; Lim 2008; Lim 2016), 한식 메뉴를 한식 분야의 듀이 십진분류법(Dewey Decimal Classification) 수정 분류 전개 방안에 의해 분류하기도 하였으나(Chung & Choi 2011), 복잡한 한식 메뉴를 단순화하여 기계학습기법을 적용하여 진행한 사례 연구는 없는 것으로 파악되었다.

오차율을 줄이기 위해 본 연구에서 적용한 앙상블기법에 의한 식수 예측 알고리즘 모델링의 결과로 예상되는 직접적인 효과는 경험치에 의한 평균 식수 예측 오차율이 10～11%대에서 약 6～7% 이내로 줄어드는 효과일 것으로 예상하며, 이로 인해 잔반량이 약 40% 정도 감축될 수 있는 효과가 있을 것으로 예상한다. 이를 금액으로 환산하면 음식물쓰레기 절감으로 인한 쓰레기 처리 비용과 식재료 구매 비용 절감액이 연간 약 5천만 원으로 예상된다(음식물쓰레기 처리 절감 비용: 연간 쓰레기 절감량 8만 kg × 200원/kg = 16,000,000원, 식재료 구매 비용 절감액: 줄어드는 일평균 잔식량 약 40～50인분 3300원/1인 × 250일 = 33,000,000원～ 41,250,000원).

본 연구는 2가지 측면에서 기존의 다른 연구 논문들과 차별점이 있다고 본다. 첫째, 한식 메뉴 분류 체계에 있어서 학자와 기관마다 다른 분류 체계를 적용하고 있는데, 본 연구에서는 통일되지 않고 복잡한 한식 분류 기법에서 탈피하여 새로운 시각으로 적용한 한식 메뉴 분류 방법은 기존 연구에서 시도하지 않은 방법으로써 복잡한 한식 메뉴를 단순화하여 식수 예측 모델링에 적용할 수 있었다. 둘째, S시청 집단급식소를 대상으로 실데이터 및 기계학습기법을 사용하여 여러 예측 모형들을 비교・결합함으로써 타 문헌에서는 보이지 못한 실무에 적용할 수 있는 수준까지 오차율을 낮출 수 있었다.

결론적으로 본 연구를 통해 적절한 양의 식재료 발주, 이에 따른 적합한 식재료의 전처리, 더 나아가 적합한 식수 예측 인원에 따른 조리 및 운영을 할 수 있도록 함으로써 미배식 잔반량을 줄이고 이에 따른 적합한 식수 예측의 비용 절감을 통해 식사의 질을 개선하는데 사용하여 이용자의 만족도를 높이고 음식물쓰레기를 줄여 환경적 사회적 비용을 줄일 수 있는 등 이른바 선순환 구조의 기틀을 마련해줄 수 있는 모델링을 제시할 수 있었다. 또한 본 연구를 통해 주로 현장 실무자들의 경험치를 기반으로 식수 예측을 하고 있는 학교, 관공서, 지방자치단체 등 집단급식소의 적정 식수 예측의 확산과 향후 활용도를 높일 수 있는 기틀을 마련하였다고 할 수 있다.

본 연구의 차별성은 먼저 첫째, 다양성 및 메뉴 선호도를 고려한 수요 예측 모델링을 하고, 이를 기계학습기법을 적용하여 식수 예측을 수행하는 새로운 접근을 시도했다는 것이다. 국내에서 진행된 식수 예측 관련 연구 사례를 살펴보면 대부분 회귀분석 등 전통적인 통계기법을 활용하고 있다. 그러나 식수와 이에 영향을 주는 요인들 간의 인과성을

규명하여 식수 예측에 적용하는 인과형 예측법의 하나인 다중회귀분석은 다양한 설명 변수들로 인해 변수들 사이에 강한 상관관계가 발생하여 다중 공선성이 존재하거나 모델링을 진행하는데 있어서 모델 그 자체가 지나치게 복잡해지는 문제가 발생한다는 단점이 있다.

둘째, 본 연구에서는 집단급식소의 실데이터를 사용한 식수 예측을 진행했다는 점이다. 집단급식소의 식수 예측은 지금까지 여러 차례 설명한 바와 같이 매우 중요하다고 할 수 있으며, 이를 위해서는 여러 가지 제약이 있는 전통적인 통계기법 외에 기계학습기법 등 다각적인 방법을 이용한 식수 예측에 관한 연구가 보다 활발히 이루어져야 하며, 실데이터를 통해 진행하는 것이 중요하다고 할 수 있다.

셋째, 식수 예측의 정확도를 높이기 위해 진행된 식수 예측 모델링의 다양한 기법의 예측 결과를 비교 적용하여 이를 최종적으로 결합하여 새로운 성능이 좋은 식수 예측 모델링을 개발하는 2단계 접근법을 시도하였다.

반면 본 연구가 가지는 한계점은 다음과 같다. 첫째, 현재 모델링의 한계점은 8개 범주의 전체 설명변수를 사용하였음에도 불구하고 이 변수들의 조합으로 다중회귀모델을 적용하였을 때 70% 이하의 설명력을 보인다는 한계가 있다. 여러 변수가 있음에도 극단적인 실제값을 예측하지 못하고 있어 주어진 데이터만으로는 알 수 없는 사항들이 존재한다고 할 수 있다. 그러므로 주어진 정보 외에도 식수 예측 모델링에 큰 변동을 주는 이벤트 혹은 변수가 있을 것으로 사료되며, 향후 이러한 변수 파악에 대한 연구가 필요한 것으로 생각된다. 특히 예측 가능한 범위의 제한이 발생하고 있어 추가적인 데이터 확보를 통해 이에 대한 추가 연구가 필요한 사항이다.

둘째, 예측 오차가 큰 특이치(outlier)에 해당하는 요일과 날짜에 대해 좀 더 세밀한 조사를 통해 오차를 최소화하는 연구가 필요하다고 할 수 있다. 아웃라이어에 식별을 위한 판별자(classifier)를 만들어 학습을 통해 예상보다 아주 많이 오거나 아주 적게 오는 날을 새로운 특성추출기법을 바탕으로 새로운 설명 변수로 추가하여 아웃라이어의 식별을 통해 아웃라이어에 대한 보정치를 다시 부여함으로써 오차율을 줄이기 위한 노력이 필요하다고 판단된다.

요약 및 결론

집단급식소에서 체계적인 식수 예측이 이루어지지 않을 경우 잔식 과다 발생 또는 음식 품절과 이로 인한 예산 부족으로 급식 서비스 품질 저하의 우려가 있으며, 이로 인해 비용 추가 발생 및 고객 불만 상승도 발생할 가능성이 크다. 이에 본 연구의 목적은 실제 현장에서 사용되고 있는 메뉴와 메뉴 이외에 고려할 수 있는 다양한 요인도 모두 고려한 예측 모델링을 포함한 식수 예측 방법을 수립하고, 실제 공공기관 집단급식소 사례에서 그 유용성을 보이는 것이다. 특히 정확한 식수 예측을 위해서 본 논문에서는 복수의 기계학습 알고리즘을 앙상블기법으로 활용한 방법을 제안하였다.

1. 식수 예측에 영향을 미치는 요인으로 식사 가능 인원 변수, 시계열 인원 변수, 메뉴 특성 변수, 요일 변수, 전후 휴일 여부 변수, 날씨 변수, 연휴 연말 변수, 이벤트 변수 등 8개 카테고리에서 총 63개의 변수들을 추출하였다.
2. 회귀분석법 중 후진 단계적 선택법, LASSO회귀, 능형회귀, 랜덤포레스트, 배깅 그리고 부스팅의 6가지 기계학습기법을 활용하여 식수 예측을 수행한 후에 다시 우수한 성능을 보인 기법에서 추론한 결과를 평균하는 앙상블기법을 활용하여 최종적으로 식수 예측 결과를 획득하였다. 요약하면 복수의 식수 예측 모형을 활용한 앙상블형 식수 예측 모델링을 1차 모델링과 최종 모델링 등 2단계로 진행하였다.
3. 그 결과 경험치에 의한 평균 식수 예측 오차율이 10~11%대에서 약 6~7% 이내로 줄고, 이로 인

해 잔반량이 약 40% 정도 감축될 수 있는 효과를 기대할 수 있었다.

이상의 결과를 종합해보면 제안된 기계학습기법은 집단급식소의 식수 예측에 유의한 기여를 하였으며, 이로 인해 비용 감축의 효과까지 기대할 수 있었다. 향후에는 본 논문에서 대상으로 했던 S시청 사례 외에도 초・중등학교 등 대형 급식소는 기계학습 등 인공지능기법을 활용한 혁신적 식수 관리를 할 가능성을 보였으며, 이러한 합리적 경영의 활성화가 반드시 필요해 보인다. 본 논문의 결과를 활용함에 있어 여타 급식소는 S시청 사례와 다를 것이기 때문에 여기에 소개된 기계학습기법의 우수성을 일반화하는 데에는 주의가 따른다. 특히 식수 관련 데이터는 상당 경우 비공개 자료일 것이기 때문에 해당 기관의 협조 없이는 제3자에 의하여 식수 관리 합리화를 추진하기는 어려울 것이다. 이러한 문제를 해결하기 위한 한 방편으로 영양사 또는 급식소 관리자들이 인공지능기법에 대한 학습 역량을 보유하는 것이 시급할 것으로 보인다.

ORCID

전종식: https://orcid.org/0000-0001-7287-0345
박은주: https://orcid.org/0000-0002-3462-6090
권오병: https://orcid.org/0000-0001-9686-6586

REFERENCES

Chung LN (2001): (The) development of a forecasting model as a management strategy in university foodservices. Masters degree thesis. Yonsei University. pp.1-109

Chung YK, Choi YK (2011): A study on developing modifications to the Dewey decimal classification for Korean foods. J Korean Libr Inf Sci Soc 45(1):29-49

Cullen KO, Hoover LW, Moore AN (1978): Menu item forecasting systems in hospital foodservice. A cost comparison of two- and three-echelon systems. J Am Diet Assoc 73(6): 640-646

Hur B (2017): Current status of parents' monitoring and trust on school lunch programs. Masters degree thesis. Seoul National University. pp.401-412

Kim HY (2000): Comparison of service style for plate waste in industry foodservice operation. J Korean Soc Food Cook Sci 16(5):416-424

Kim JB (2014a): An empirical study on the critical success factors in implementing appropriate demand forecasting model for optimum feeding number of persons in large feeding organization. JDMR 17(5):39-46

Kim JE (2014b): A study on how the catering service menus made with local food effects of consumer's satisfaction -focused on university students in Jeonju- .Int J Tour Manage Sci 29(5):337-355

Lee H, Chung SH, Choi EJ (2016): A case study on machine learning applications and performance improvement in learning algorithm. J Digit Converg 14(2):245-258

Lim HJ (2008): Comparative assessment of forecasting models applicable for business and industry foodservice operations. Masters degree thesis. Yonsei University. pp.1-104

Lim JY (2016): Analysis of forecasting factors affecting meal service in business foodservice. Masters degree thesis. Yonsei University. pp.1-92

Lin BS, Vassar JA, Miller J (1992): Building a strategic forecasting system for hospital foodservice operations. J Am Diet Assoc 92(2):204-207

Makridakis S, Wheelwright S, Hyndman RJ (1998): Forecasting : methods and applications. 3rd ed. John Wiley and Sons. New York. 642 p.

Miller JJ, McCahon CS, Miller JL (1991): Forecasting production demand in school food service. School Food Serv Res Rev 15(2):117-121

Miller JL (1990): Forecasting in foodservice: surveys in three types of operations. National Assoc Coll Univ Food Serv J 15:13-16

Miller JL, Shanklin CW (1988): Forecasting menu-item demand in foodservice operations. J Am Diet Assoc 88(4):443-449

Ministry of Food and Drug Safety (2016). Presidential decree in food sanitation law, paragraph 2. Available from: http://

www.law.go.kr/lsInfoP.do?lsiSeq=184959&efYd=20160804 #0000. Accessed Aug 4, 2018

Park SJ (2017): A study on environmental perception, school food waste status and source reduction education of middle school students in Seoul region. Masters degree thesis. Yonsei University. pp.1-118

Repko CJ, Miller JL (1990): Survey of foodservice production forecasting. J Am Diet Assoc 90(8):1067-1071