Data Granulization을 이용한 수송수요예측에 관한 연구

Study on the Demand Prediction for Transportation System Utilizing Data Granulization

이덕규* 홍태화** 김학배§ 우광방+
Lee, Deog-Kyoo Hong, Tae-Hwa Kim, Hag-Bae Woo, Kwang-Bang

ABSTRACT

The demand prediction becomes an essential mean to utilize efficiently finite traffic facilities and to provide the optimized schedules for transportation system. The demand prediction is one of the critical complex management schemes for distibuting resources of transportation service by means of computer system. The construction of a prediction model is based on data granulization, followed by processing the raw input data and evaluating the predicted output values. A large number of ecomomic-social parameters are also to be implemented in conventional prediction models which are only based on a sequence of past data. The proposed prediction models are classified by static and dynamic characteristics and its performances are evaluated utilizing computer simulation.

1. 서론

철도자원의 운영이나 이와 관련된 막대한 데이터의 업무처리, 그리고 수송계획등은 상당한 경험을 가진 실무자나 전문가의 직관에 의해서 진행된다. 그러나 한정된 철도자원에 비해 급속한 증가량을 보이고 있는 수송수요(승객, 화물)를 충족하기 위해 기존의 경험과 직관에 의한 방법들은 많은 문제점을 안고 있다. 따라서 철도운영과 관련되는 수많은 정보들을 보다 신속하고 효율적으로 관리하기 위해서 수송수요예측을 위한 시스템의 개발이 요구된다.

일반적으로 교통관련 수송수요는 시간상의 변동으로 대응되는 stochastic process로 모델링 되어 예측되어질 수 있다. 물론 이를 위한 데이터수집 방법론 및 파라미터화에 있어 경제 사회적인 요소들이 중요한 역할을 하기 때문에 단순한 수리적 모델링이 용이하지 않으나, 기존의 단순한 통계기법이나 인력에 의존한 합산방식보다 유동적이고 통합적인 시스템적 접근방식을 통해 관련 전산시스템의 운영체계에 접목시키는, 엔지니어링 측면의 접근방식과 최신 예측이론의 활용화가 선행되어야 한다.

---

* 연세대학교 기계·전자공학부 박사과정
** 연세대학교 기계·전자공학부 석사과정
§ 연세대학교 기계·전자공학부 조교수
+ 연세대학교 기계·전자공학부 교수

기존의 예측모델은 시간에 대해 stochastic 시계열 함수(time series)로 표시되고, 이의 대표적인 형태는 다항함수 또는 saturation곡선형이다. 그러나 이들은 수송수요의 극변동의 고려가 용이치 못하고, 장기적 변동을 감안한 계절적 효과 등이 구현될 수 없다는 제한점이 있다. 장기적 변동, 주기적 진동, 계절적 변화, 또는 불규칙적 변동들을 위해 계절적 조정기법이 고려되고 있으나 이 방법의 복잡한 구조와 수치적 사항의 현실화가 문제시되고 있다. 시간 변동 데이터를 직접적으로 고려하기 위해 AR(autoregressive), MA(moving average), ARMA(autoregressive moving average), ARIMA(autoregressive integrated moving average)등의 stochastic 모델들이 검토되고 있으나 그 정확도 문제와 함께 교통관련 수송수요예측 적용에 필요한 가정이나 제약에 대한 분석의 결여로 인하여, 이러한 모델들은 수송수요예측을 위한 보완과정을 거쳐야 한다. 물론 EMM/2와 같은 교통계획 S/W package가 수송수요예측을 위해 위와 같은 시계열함수 모델들에 근거해 예측치를 계산하고, 사용자는 관련 필요자료를 적절한 변수를 통해 입력시키는 기법이 개발되어 있으나 이를 직접 국내의 고속철도를 위한 수송수요예측 시스템에 적용시키는 데에는 한계가 있으며 이의 보완은 관련 기본이론 연구 및 알고리즘과 S/W개발을 통해 진행되어 왔다.

본 연구에서는 기존기법의 보완을 위해 데이터 granulization 기법을 이용하여 고속철도 수송수요예측을 위한 최적의 시계열함수를 효율적으로 발생시키는(이를 통해 필요정보를 실시간으로 제공)알고리즘과 이의 효과적 데이터 처리를 위한 시계열 모델, 칼만 필터, 인공지능형 기법(신경회로망기법, 퍼지논리 예측기법)등을 적용토록 하는 것이다. 또한 개발기법의 실용화가 진행되고 철저한 시험과정과 현장 평가를 거쳐 고속전철의 종합정보시스템 엔지니어링의 기반 기술 활용을 목적으로 한다.

2. 수송수요예측 시스템

2.1 수송수요예측 시스템의 개요

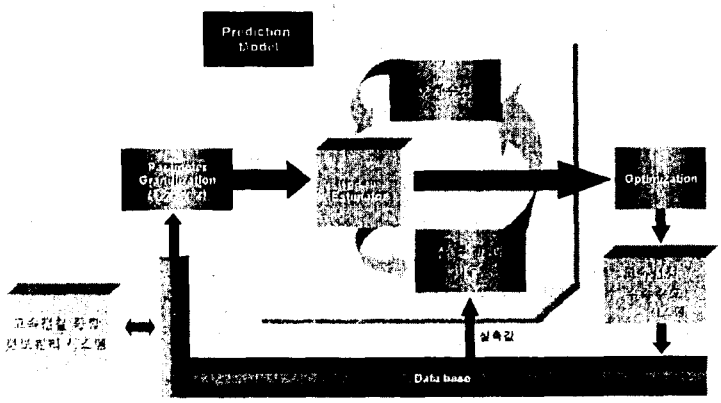

그림 1. 수송수요예측 시스템

수송수요예측시스템의 구조는 1)종합 정보관리 시스템, 2)수송수요예측 시스템, 3)데이터베이스, 4)예측모델의 4가지 부분으로 분류된다. 종합 정보관리 시스템은 철도의 운영과 관련해서 발생된 모든 데이터를 총괄하여 관리할 수 있는 슈퍼컴퓨터라고 할 수 있다. 이 시스템은 데이터베이스와의 실시간적 통신이 가능하여야 한다. 여기서 따로 수송수요만을 위한 예측 시스템을 가정한다. 예측시스템 역시 데이터베이스에 접근하여 실측데이터들을 읽어들이고 예측모델로부터 계산된 예측값들을 임시적으로 저장한다. 이렇게 두가지 시스템이 따로 존재하는 이유는 예측모델이 예측된 값들을 임시적으로 필요로 할 경우 실시간상에서 이 시스템에 접근하여 데이터를 사용할 수 있게 하기 위함이다. 이러한 예측시스템은 컴퓨터의 캐쉬 메모리(cache memory)와 같은 기능을 하고 최종적으로 예측된 데이터들은 데이터베이스를 통해 종합 시스템에 저장되는 것이다. 예측모델은 실측값으로부터 예측값들을 구하기 위한 함수들로 구성되는 소프트웨어를 의미한다. 물론 실측값은 이 모델의 입력, 예측값들은 출력이 될 것이다. 이러한 모델에는 최적화된 모델식의 차수(order)를 결정하고 적절한 계수(coefficient)들을 추정(estimation)하며 피이드백을 위해 실측값과 예측값들의 오차를 최소화하는 최소자승법(least square method)등이 프로그램된 하나의 팩키지가 될 것이다. 그림 1은 이러한 수송수요예측시스템의 구조이다.

2.2 고속철도 수송수요예측기법을 위한 입출력 변수

수송수요예측 시스템에서 출력변수는 승객수(M)[단위:명]만을 고려하고, 기본적인 입력변수는 1)O-D간 파라미터, 2)시간 파라미터, 3)사회경제적 파라미터로 구분된다.

O-D간의 파라미터는 고속전철의 출발지와 도착지를 의미하며 아래의 식 (1)과 같이 정의한다. 이 경우에 O-D간 파라미터의 총 경우의 수는 식 (2)와 같이 구해질 수 있다.

$$i(Origin),\ j(Destination) \in \{서울, 천안, 대전, 대구, 경주, 부산\} \tag{1}$$

$$경우의수 \Rightarrow 6\,(i의\ 개수) \times 5\,(j의\ 개수) = 30 \tag{2}$$

시간적 파라미터는 ① 시계수(hour parameter), ② 일계수(day parameter), ③ 계절계수(season parameter), ④ 년계수(year parameter)로 구분된다. 시계수, T_k는 24시간으로 식 (3)과 같이 정의된다. 일계수, x는 평일과 주말, 그리고 특정일로 구분된다. 식 (4)는 일계수를 정량화한 것이다. 계절계수, y는 봄, 여름, 가을, 겨울로 구분하고 식 (5)에 정량화되어 있다. 년계수는 z로 정의한다.

$$T_k(k=1,2,\cdots,24),\quad T_k : (k-1)시간에서\ (k)시간까지의\ 승객량 \tag{3}$$

$$\begin{aligned} x = 1 &\Rightarrow 특정일\ (명절,\ 연휴) \\ 2 &\Rightarrow 주말\quad (토,\ 일) \\ 3 &\Rightarrow 평일\quad (월\sim 금) \end{aligned} \tag{4}$$

$$y = \begin{array}{ll} 1 & \Rightarrow \text{ 봄} \\ 2 & \Rightarrow \text{ 여름} \\ 3 & \Rightarrow \text{ 가을} \\ 4 & \Rightarrow \text{ 겨울} \end{array} \qquad (5)$$

사회 경제적 파마미터, $E_{ij}(z)$는 인구, 소득, 고용, 교통환경에 의해서 정의된다. O(i)-D(j)간의 $E_{ij}(z)$는 년계수 z의 함수이다.

2.3 Granulization

데이터 granulization은 사회경제적 요인들을 제외한 시변 데이터들을 특성에 따라서 class로 분류하는 기법이다. 만약, 수송수요예측 시스템이라면 계절, 요일, 시간 등에 따른 여러 형태의 granule(최소의 단위를 의미)을 구성하고, 구성된 granule을 유사한 형태들끼리 분류하여 class로 정의한다. 이러한 class 데이터를 모델 함수의 과거 데이터로 사용함으로써 효과적인 예측이 가능하다. 이제부터 간단하게 granulization 과정을 설명할 것이다.

기존의 경부선 정차역을 서울, 천안, 대전, 대구, 경주, 부산의 6개역으로 한정한다. 이러한 경우 총 30가지의 O(Origin)-D(Destination) 구간을 정의할 수 있다. 따라서, 각 O-D간에 따른 시간적 파라미터들을 구성할 수 있다. 시간의 파라미터는 크게 계절단위, 일단위, 시간단위로 구성된다. 계절의 파라미터는 봄, 여름, 가을, 겨울로 하고, 물론 이러한 기준은 항시 변동 가능하다. 일의 파라미터는 공휴일, 주말과 평일의 3가지로 분류한다. 마지막으로 시간의 파라미터는 각 시간을 단위로 하여 24개로 구성된다.

모형화 방법은 접근 방법에 있어서 사회경제적 파라미터들과 다른 시변 파라미터들 사이의 커다란 차이를 인식하여 서로 다른 개념에서 출발한다. 즉 인구, 각 도시의 주민소득, 타교통수단과의 연계성 등과 같은 사회경제적 파라미터들은 변화의 속도가 느리기 때문에 시변파라미터들을 고정시켜 놓고 수집된 사회경제적 통계 데이터(물론 가장 최근의 것)를 바탕으로 기존의 연구방법을 활용하여 모형화한다. 사회경제적 파라미터들이 데이터의 빈번한 수집이 불가능한 반면에 시변 파라미터들은 특성상 잦은 변화로 인해 통계적 데이터의 수집이 용이하다. 1년 주기의 데이터는 총 8760(24시간 * 365일)가 될 것이다. 예를 들어, 어느 해, 봄, 평일이라는 하나의 시간을 생각하면 8760개의 데이터 중 여러 개가 여기에 포함될 것이다. 즉, 이러한 식으로 8760개의 데이터들을 커다란 하나의 클래스 속에 분류하게 되는 것이다. 이러한 데이터들은 각각의 상황에서 현재의 값을 추정하는데 모델함수의 과거데이터로 사용된다.

본 연구에서는 정적모델(static model)과 동적모델(dynamic model)로 구분하여 보다 자세히 graulization과정을 설명할 것이다.

2.4 정적모델

정적모델(Static Model)은 년계수, z와 사회경제적 파라미터, $E_{ij}(z)$는 변화하고 시간적 파라미터들, T, x, y는 고정시킨 모델을 의미한다. 수집된 사회경제적 통계 데이터(물론, 가장 최근

의 변동치)를 바탕으로 O-D간의 승객량, M은 식 (6)에 의해 결정된다. 이 식에서 M은 년계수, z와 사회경제적 파라미터, $E_{ij}(z)$의 함수이다.

$$M = f_{ij}(z, E_{ij}(z)) \tag{6}$$

2.5 동적모델

동적모델(Dynamic Model)은 정적모델과는 반대로 년계수, z와 사회경제적 파라미터, $E_{ij}(z)$는 고정시키고 시간적 파라미터들, T, x, y를 변화시킨 모델을 의미한다. 변동된 사회경제적 통계 데이터의 빈번한 수집이 불가능함을 고려하여 O-D간의 승객량, M은 식 (7)에 의해 계산된다. 이를 위한 필요조건은 아래와 같다.

$$M = f_{ij}(T, x, y) \tag{7}$$

여기서, $f_{ij}(\cdot)$는 예측모델의 식이다.

필요조건

1) 각 구간별(O-D), 시간별로 모든 granule에 대한 충분한 과거 데이터 수집
   1년 : 24(시간) * 365(일) = 8760개의 데이터 수집 필요
2) 분류(partitioning) 및 클러스터링(clustering)
   T, x, y 의 각 경우에 대한 총 경우의 수 계산
   총 경우의 수 = T(24가지) * x(3가지) * y(4가지) = 288(classes)
3) 한 class 안의 주기적/비주기적 요소 고려

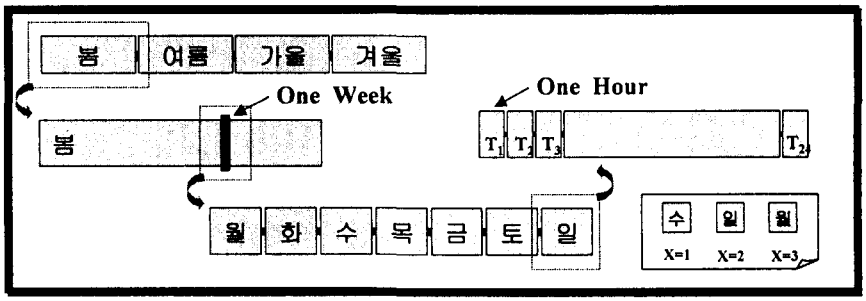

그림 2. Class 분류

그림 2는 특정 클래스의 분류 도식을 나타낸다. 이러한 알고리즘을 기초로 하여 과거의 샘플 데이터들을 이용한 현재 및 미래 데이터 예측을 위해 예측기법을 이용하여 승객량을 예측한다. 즉 단일 클래스내의 요소(element)인 모든 granule을 바탕으로 해당 클래스의 승객량이 예측된다. 특정 클래스의 샘플 granule의 index인 윈도우 크기는 식 (8)와 같다. 그림 3은 하나의 특정 클래스를 나타낸다.

$$\text{Window Size N} : t_k = \{t_1, \cdots, t_{p-1}, t_p\} \tag{8}$$

여기서 t_0 : 첫 번째 클래스가 나타나는 시간,

t_p : 클래스가 나타나는 예측될 시간과 가장 가까운 시간

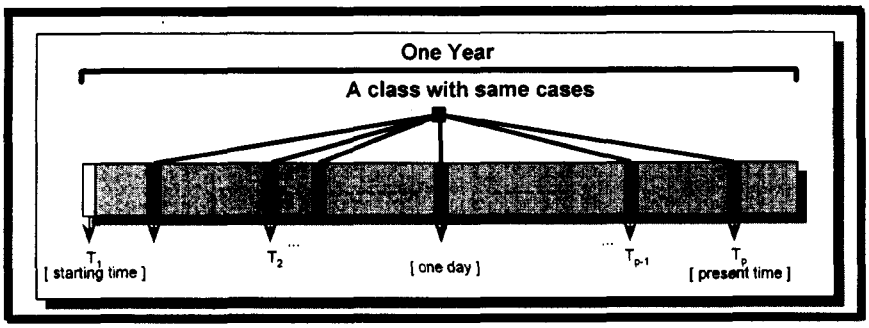

그림 3. 특정 class

모든 O-D에 대해 이러한 인덱스들을 이용하여 입력계수(T, x, y)들에 의해 할당된 모든 클래스에 관한 샘플 데이터를 수집하여 현재 기준시점 t_p의 승객량 $\widehat{M}_1(t_{p1})$가 식 (9)에 의해 승객량이 예측된다. 모델의 결정 요소, $\hat{g}(\cdot)$는 예측모델식이 된다.

$$\widehat{M}_1(t_{p1}) = \hat{g}(M_1(t_1), M_1(t_2), \cdots, M_1(t_{p1}-1)) \tag{9}$$

윈도우의 크기는 데이터 granule의 개수가 된다. 이는 데이터의 종류와 수집 방법에 영향을 받으므로 스케줄링 및 역무 자동화 과정으로부터의 피드백 입력 활용이 가능하다. 또한 정확도와 계산량 간에 상호보완이 고려되어져야 한다.

이 경우의 사용자 입력은 O-D간의 선택, T_k의 선택, 요일 및 계절, 연도의 선택이 된다. 이러한 입력을 바탕으로 데이터의 granulization에 의한 클래스가 결정된다. 하나의 예를 들어 보자. 어느해 가을 주말, 오후 1시부터 오후 3시까지의 수요예측과정은 다음과 같다. 이러한 경우, 승객량의 예측값은 식 (9)에 의해 아래의 4단계 과정을 거쳐 구해진다. [절차 2]에서는 주어진 예측 구간에 대한 클래스 데이터 집합을 탐색한다. [절차 3]에서는 클래스로 분류된 데이터를 이용하여 예측 구간별 샘플데이터를 추출한다. [절차 4]에서는 이러한 샘플데이터를 이용하여 예측 구간의 시간별 예측 승객량을 구한다. 그림 4는 이러한 과정에 대한 운영 흐름도가 된다.

운영흐름도

[절차 1] : 예측 내용/어느해 가을의 주말, 오후 1시부터 오후 3시까지의 수요예측

[절차 2] : 분할된 클래스들의 탐색/

$$\{T_{14}, x=2, y=3\} \cup \{T_{15}, x=2, y=3\}$$

$$= M_1(t_{p1}) \cup M_2(t_{p2})$$

$M_1(t_{p1})$: 클래스 1을 위한 집합, $M_2(t_{p2})$: 클래스 2를 위한 집합

[절차 3] : 클래스 1, 클래스 2에 대한 샘플데이터 수집

클래스 1 : $M_1(t_{p1}) = \{M_1(t_1), M_1(t_2), \cdots, M_1(t_{p1}-1)\}$,

클래스 2 : $M_2(t_{p2}) = \{M_2(t_1), M_2(t_2), \cdots, M_2(t_{p2}-1)\}$

[절차 4] : 각 클래스의 예측(prediction)

$$\widehat{M}_1(t_{p1}) = \hat{g}(M_1(t_1), M_1(t_2), \cdots, M_1(t_{p1}-1))$$

$$\widehat{M}_2(t_{p2}) = \hat{g}(M_2(t_2), M_2(t_2), \cdots, M_2(t_{p2}-1))$$

$$\widehat{M} = \widehat{M}_1(t_{p1}) + \widehat{M}_2(t_{p2})$$

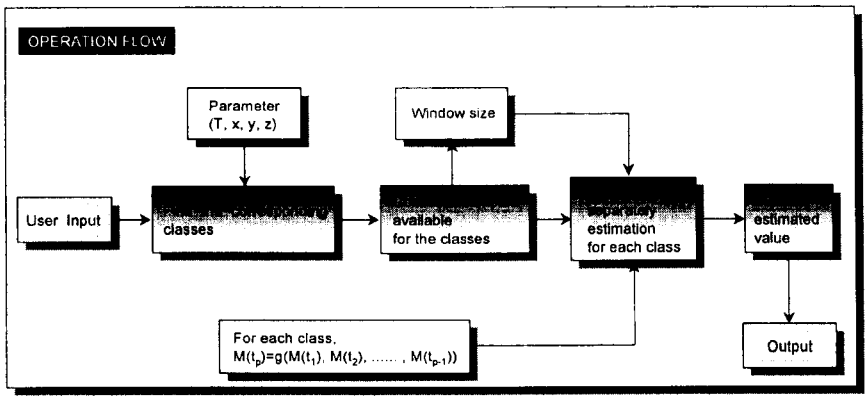

그림 4. 운영 흐름도

3. 결론

본 연구를 통해 수송수요예측을 위한 데이터 처리 방법(granulization)과 기존의 예측기법들에 의한 모델들(정적모델, 동적모델)이 검토되었다.

가장 중요한 운송수단으로서의 철도시스템에서 기존의 예측모델들이 과거 데이터들의 단순한 수학적 모델링에 국한되어 있었을 뿐만 아니라 실측값들의 실질적인 경향을 추출하는데 중요한

역할을 하는 사회경제적 요소들이 무시됨을 감안할 때, 보다 정확한 예측을 위해 이러한 요소들을 고려할 수 있는 새로운 모델들의 제안이 필요하다. 따라서 본 연구에서는 실측데이터의 보다 효율적인 이용을 위해 granulization기법을 제안하였고 변수들의 특성에 따라 분류된 정적모델과 동적모델을 제시하였다. 그러나 실질적으로 이러한 모델들 역시 수학적 모델링이 용이하지 않고 특히, 정적모델에서 사회경제적 변수들을 정량화하는 데 한계점을 드러내고 있다. 따라서 향후 여러 수요예측분야에서 요구되는 사회경제적 변수들에 기존의 예측모델들이 접목되어 출력변수에 대한 다양한 요소들을 고려함으로써 종합적인 수송수요예측 관리시스템개발이 추진될 수 있을 것이다.

참고문헌

1. Masafumi Tsutsemi, and Takeshi Chishaki(1992), "A Study On Time Series Prediction System and AROP Model for Transportation Demand Analysis", Memoirs of the Faculty of Engineering, Kyushu Univ., Vol.52, No.2, pp.145-169
2. James V. Hansen, and Ray D. Nelson(1997), "Neural Networks and Tranditional Time Series Methods:A Synergistic Combination in State Economic Forecasts", IEEE Transaction on neural networks, Vol.8, No.4, pp.863-873
3. Daniel Brand, Thomas E. Parody, Poh Ser Hsu, and Kevin F. Tierney, "Forecasting High Speed Ridership", Transportation research record, pp.12-18
4. Tara J. Weidner, "Hubbing in U.S. Air Transportation System:Economic Approach", Transportation research record, pp.28-37
5. Y. Liu, and X. Wang, "The Analysis of Train Transportation Simulation System", Railway operations, pp.297-304
6. Ardeshir Faghri, and Sandeep Aneja, "Artificial Neural Network-Based Appoach to Modeling Trip Production", Transportation research record, pp.131-136