대중교통 이용자 수를 통한 상습 정체구간 예측

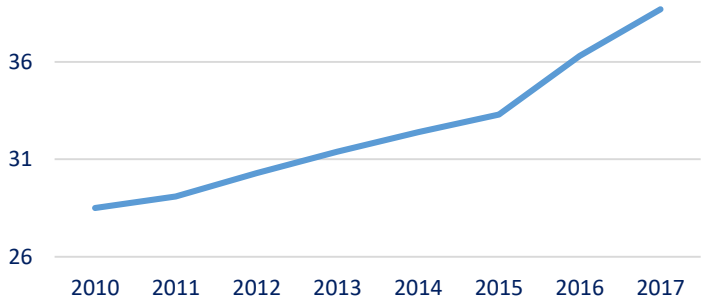
목 차

배경 및 사전조사

- **교통 혼잡 비용**

- 교통체증이 없는 상황에서 정상속도를 냈을 경우, 줄일 수 있었던 불필요한 차량 운행비와 시간 손실 등을 환산한 액수를 말한다.
- 차량 운행 비용 + 시간 가치 비용
- 2017년 기준 약 38.7조원
  → 국내 총생산 (GDP)의 3.4%

교통 혼잡 비용 변화 추이



* 출처 : 통계청

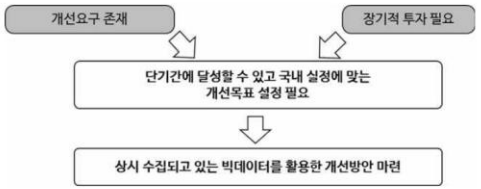
이 비용의 증가 원인은 늘어나고 있는 차량 보유 대수와
유가 인상, 교통체증 심화 등이다.

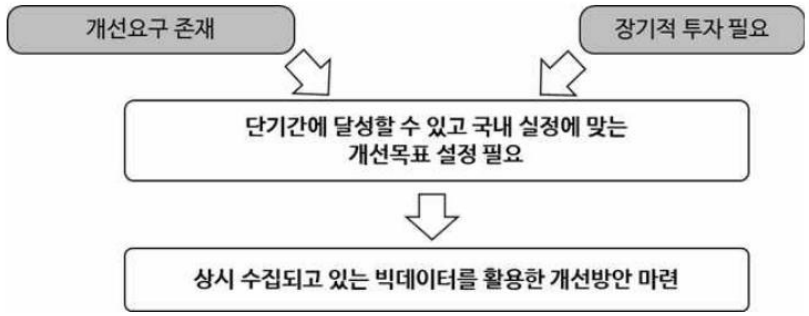
프로젝트 소개
배경 및 사전조사

기존 연구

그림 1-1 연구의 목표 설정

	[4단계 모형] 집계적 모형			[활동기반모형] 마이크로시뮬레이션
	특징	문제점		특징
시간	시간대 구분 (3~4개)	시간에 민감한 정책 테스트 어려움	새로운 모형 개발에 의한 개선 방안	분단위 시뮬레이션
공간	행정동(구) 단위	토지이용의 영향을 세밀히 고려하기 어려움		블록, 필지 단위 시뮬레이션
모형단위	행정동(구) 단위의 집계치	개인, 가구의 의사결정과정을 반영하기 어려움		개인, 가구





1. 가용한 기존의 데이터 분석방법을 활용
2. 상시로 수집되는 서울시 공공데이터 활용

출처: 국토연구원 융합 빅데이터를 활용한 교통수요 추정 개선 연구

배경 및 사전조사

연구 방향 설정

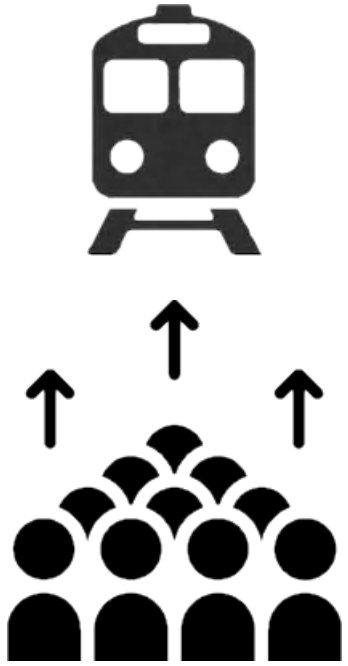
교통수요 예측 4단계 모형

- 통행 발생
- 통행 분포
- 교통수단 선택
- 통행 배정

본 프로젝트 4단계 모형

- 통행 발생 데이터 수집
- 도로명에 따른 데이터로 통합
- 예측 데이터 설정
- 데이터 예측 수행

문제 정의



문제 정의

대중교통 이용자 수를 통한 상습 정체구간 예측

데이터 수집 및 분석
데이터 전처리

<Raw Data>

- **지하철 승/하차 데이터(일별/ 2014.01~2020.03)**
  - Column 수 : 26개
  - Class 수(단위 : 년) : 200805개
- 지하철 역 주소 데이터

- **버스 승/하차 데이터(월별 / 2014.01~2020.03)**
  - Column 수 : 55개
  - Class 수(단위 : 월) : 39080개
- 버스 정류장 주소 데이터

- **공공자전거 이용 데이터(일별 / 2017.07~2019.11)**
  - Column 수 : 11개
  - Class 수(단위 : 월) : 520651개
- 공공자전거 대여소 주소 데이터

- **도로명별 차량속도 데이터(일별/ 2014.01~2020.03)**
  - Column 수 : 36개
  - Class 수(단위 : 월) : 149076개

<1차 가공 – 도로명별 통합>

- **지하철 승/하차 데이터**
 -> 지하철 역 주소 데이터를 이용하여 역명에 따른 도로명 부여

- 버스 승/하차 데이터 & 공공자전거 이용 데이터
 -> 버스 정류장과 공공자전거 대여소 데이터는 위도 경도만 나와있어 주소 좌표 변환 툴인, GeoCoder-Xr을 사용하여 도로명 부여

위도	경도
37.51425	127.0611
37.5233	127.0385
37.52512	127.0525
37.51864	127.0354
37.51764	127.0225
37.5293	127.0356
37.50155	127.0386

도로명주소		
서울특별시	강남구	봉은사로
서울특별시	강남구	도산대로
서울특별시	강남구	도산대로
서울특별시	강남구	언주로
서울특별시	강남구	도산대로
서울특별시	강남구	압구정로
서울특별시	강남구	테헤란로

데이터 수집 : 서울 열린 데이터 광장 & 서울시 교통정보시스템 활용

데이터 수집 및 분석
데이터 전처리

<2차 가공 – 월별 통합>

- **지하철 승/하차 데이터 & 공공자전거 이용 데이터**
  : 일별 -> 월별

- **도로명별 차량속도 데이터**
  : 일별 -> 월별

※ 주말을 제외한 일별 데이터를 더하여 계산

ROADNAME	YYMM
강남대로	1707
강변역로	1707
강서로	1707
경인로	1707
고덕로	1707
공항대로	1707
과천대로	1707
광평로	1707
금호로	1707
남부순환로	1707
녹사평대로	1707

- 모든 데이터가 존재하는 2017.07 ~ 2019.11까지 데이터 사용.

<3차 가공 – 데이터 컬럼 추가 >

- **버스 승/하차 데이터**
  : 평일과 주말의 비율을 따져 MODBus_aa 컬럼 추가 (aa = 시간)

사용일자	승차총승객수	하차총승객수
20190101	26	5
20190101	4	0
20190101	0	3
20190101	25	10
20190101	0	32
20190101	30	0
20190101	2	91
20190101	101	3
20190101	0	27
20190101	34	0
20190101	24	18
20190101	50	7
20190101	15	11

<일자별 버스 승/하차 데이터>

데이터 전처리

상습 정체 구간 설정

SPD_07	SPD_08	SPD_09	SPD_17	SPD_18	SPD_19	IsJAM
34.772	30.1683	25.4168	22.8437	22.4455	20.9403	0
23.1459	22.8353	20.2642	19.7973	20.1834	20.2486	0
27.6199	25.5341	22.0218	20.7565	20.0964	19.7146	0
31.1919	25.8264	21.728	23.3681	21.3067	20.4009	0
29.686	26.8102	24.8669	25.6361	25.2954	24.886	0
28.9541	25.7288	23.0892	23.2019	22.2689	20.896	0
41.5817	28.4957	24.5733	25.1752	23.8781	24.14	0
31.7183	26.5859	23.9696	24.8189	24.6982	24.0327	0
24.1604	21.133	18.7933	17.8384	17.612	17.644	1
35.5289	30.498	26.8483	27.3675	26.1979	24.8325	0
43.6826	38.9406	35.314	32.15	31.5961	29.2553	0
25.6552	22.7042	19.0846	17.9481	17.3981	17.2022	1
22.3946	20.8201	18.8036	15.5074	15.3422	15.0201	1
25.5353	24.2546	20.988	19.1415	18.8138	17.6519	1

SPD_07, SPD_08, SPD_09 : 출근 시간대의 도로 속도
SPD_17, SPD_18, SPD_19 : 퇴근 시간대의 도로 속도

<IsJAM>
속도가 20km/h 이하인 시간대가 3개 이상이면
상습 정체 구간으로 설정

*** 참고 : 경기도 교통 DB 센터 –**
교통혼잡구간 정체관리 전략 (2017)

데이터 전처리

데이터 정규화

SubIN_07	SubIN_08	SubIN_09	SubIN_17	SubIN_18	SubIN_19	SubOUT_07	SubOUT_08
2047	3194	2699	9703	16692	11878	9331	17079
4477	6266	4165	2721	3134	2444	1392	2254
4599	4803	2038	1075	1122	746	656	925
957	1073	406	302	432	211	358	618
2475	1868	793	800	784	343	513	695
1383	1261	573	952	1007	572	470	1011
356	558	242	200	333	141	183	259
1685	1856	1201	1362	2099	1036	1291	2264
880	1068	500	466	452	300	391	541
3298	4342	2411	1895	2953	1807	1388	2967
251	426	254	584	770	496	546	867
1146	1641	892	756	880	573	454	1122
603	847	507	492	767	492	349	950
1696	1813	1042	1386	3010	1402	1254	3857
4451	5854	3187	1313	1848	1253	845	1644

SubIN_07	SubIN_08	SubIN_09	SubIN_17	SubIN_18	SubIN_19	SubOUT_07	SubOUT_08
0.379035	0.889274	1.955559	5.547606	5.028813	6.429606	6.61484524	4.52242102
2.274309	3.017266	3.757649	0.787186	0.240238	0.648299	0.27218978	-0.1221673
2.369463	2.003838	1.143021	-0.33508	-0.47038	-0.39226	-0.3158181	-0.5385354
-0.47111	-0.57995	-0.86313	-0.86212	-0.71409	-0.72012	-0.5538974	-0.6347168
0.712853	-0.02925	-0.3874	-0.52258	-0.58976	-0.63923	-0.4300642	-0.6105931
-0.13885	-0.44973	-0.65784	-0.41894	-0.511	-0.49889	-0.4644179	-0.5115921
-0.53117	-0.58342	-0.74758	-0.7503	-0.70702	-0.66558	-0.5275329	-0.6588404
1.35475	1.6845	1.601533	0.224008	0.176311	0.257935	0.26899408	0.10121157
-1.02175	-1.02813	-1.04997	-0.66985	-0.59471	-0.54547	-0.4036997	-0.5567065
-0.3237	-0.1865	-0.26571	-0.55258	-0.55586	-0.49828	-0.4772007	-0.4768164
-0.74721	-0.73651	-0.73897	-0.73258	-0.59577	-0.54792	-0.5610877	-0.5307031
0.105273	-0.06735	-0.08132	-0.12303	0.196442	0.009744	0.1619383	0.38004352
2.254031	2.731871	2.555437	-0.17281	-0.21397	-0.08157	-0.1648215	-0.3132768
-0.60058	-0.58757	-0.03829	1.570589	0.904592	0.996378	0.9496771	0.96465072

- 버스 승/하차 MAX값 : 938431명
- 버스 승/하차 MIN값 : 80명
- 공공자전거 MAX값 : 5739명
- 공공자전거 MIN 값 : 0명

- 버스 승/하차 데이터와 공공자전거(따릉이) 이용객 데이터의 편차가 크기 때문에 정규화 실시

데이터 수집 및 분석
데이터 전처리

최종 데이터

지하철 승/하차 승객수	+	버스 승/하차 승객수	+	공공자전거 이용객수	상습 정체구간

SubIN_07	SubIN_08
0.367879	0.877273
2.241339	2.989043
2.335397	1.98334
-0.47248	-0.58076
0.697855	-0.03425
-0.14405	-0.45152
-0.93583	-0.93478
0.088788	-0.0425
-0.53184	-0.58419
1.332364	1.666437
-1.01678	-1.02552
-0.32677	-0.1903

●●●

BusOUT_19	MODBusIN_07
2.854708726	1.998596231
-0.75083976	-0.677739229
0.526946189	0.407792137
1.322420698	1.141173272
-0.149033129	0.223036929
-0.518164293	-0.481610108
0.394254244	0.445978241
-0.605990424	-0.610283414
-0.388574688	-0.364972355
-0.568573374	-0.473008782
3.568719263	3.366553532
-0.425172846	-0.448906167
-0.019963205	-0.321682168
-0.395062829	-0.406498168

●●●

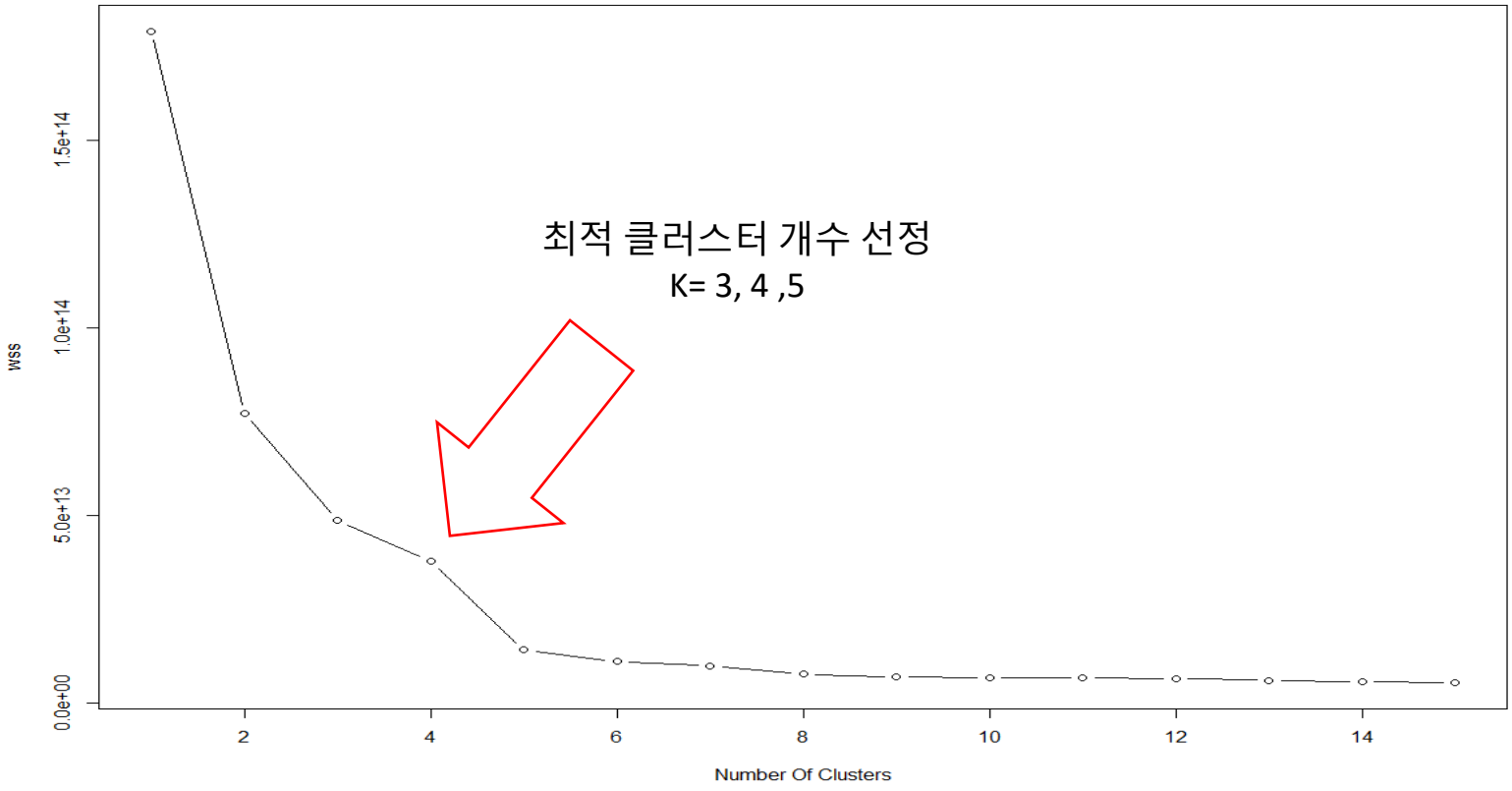
Byc_17	Byc_18	Byc_19	IsJAM
-0.21161	0.197438	0.257427	0
-0.58556	-0.58282	-0.50297	0
-0.57457	-0.6102	-0.44961	0
-0.09612	0.149528	0.808829	0
-0.62956	-0.63757	-0.62748	0
-0.23911	-0.01816	-0.06274	0
NA	NA	NA	0
NA	NA	NA	0
-0.62956	-0.61704	-0.59636	1
0.316325	0.385658	1.000041	0
-0.47008	-0.5178	-0.46295	0
0.5198	0.522545	0.977807	1
-0.16212	-0.17558	0.048428	1
0.27233	0.519123	0.875531	1

- 총 Column 수 : 지하철(12개) + 버스(24개) + 공공자전거(6개) = 42개
- 총 Class 수 : 2785개

데이터 수집 및 분석

K-means clustering

비지도 학습 – 클러스터링(K-means분석)



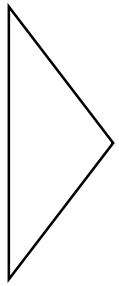
데이터 수집 및 분석

K-means clustering

비지도 학습 – 클러스터링(K-means분석)

클러스터	K=3	K=4	K=5
1	1/4	1/4	19/54
2	33/77	6/11	12/15
3	9/15	24/60	1/4
4		12/21	6/11
5			5/12

K 값에 따른 클러스터 구성
(상습정체구간/ 클러스터 노드 수)



클러스터	K=3	K=4	K=5
1	0.25	0.25	0.35
2	0.43	0.55	0.8
3	0.6	0.4	0.25
4		0.57	0.55
5			0.42

상습 정체 도로 예측 적중률

- 분석 1. K 값이 5일 때 가장 큰 적중률을 가진 클러스터가 생성됨
- 분석 2. 사용한 변수들은 도로들의 특징을 분석하는데 도움을 줄 수 있음을 확인

데이터 수집 및 분석

K-means clustering

비지도 학습 – 클러스터링(K-means분석)

클러스터	K=3	K=4	K=5
1	0.25	0.25	0.35
2	0.43	0.55	이태원로, 세종대로, 대학로 서초대로,서초중앙로, 신반포로, 신촌로,압구정로, 여의나루로, 율곡로,을지로, 테헤란로, 퇴계로,학동로
3	신촌로, 왕산로, 왕십리로, 종로, 한강대로 등	0.4	0.25
4		서초중앙로, 신반포로, 압구정로, 테헤란로 등	0.55
5			0.42

일반적으로 유동인구가 많은 서울의 번화가

서울의 중심부에 위치한 도로명 위주

서울의 강남 부분에 위치한 도로명 위주

지하철 승/하차, 버스 승/하차, 서울시 공공자전거 이용객 수는
상습 정체구간 예측에 영향을 주는 변수임을 시사

데이터 분석

모델 선택 – 지도 학습

모델	정확도
Decision Tree	85.70%
Random Forest	**93.50%**
Logistic Regression	88.46%
SVM	84.69%

샘플링 10회 진행 후, 분석 -> 가장 정확도가 높은 Random Forest 선정

데이터 분석

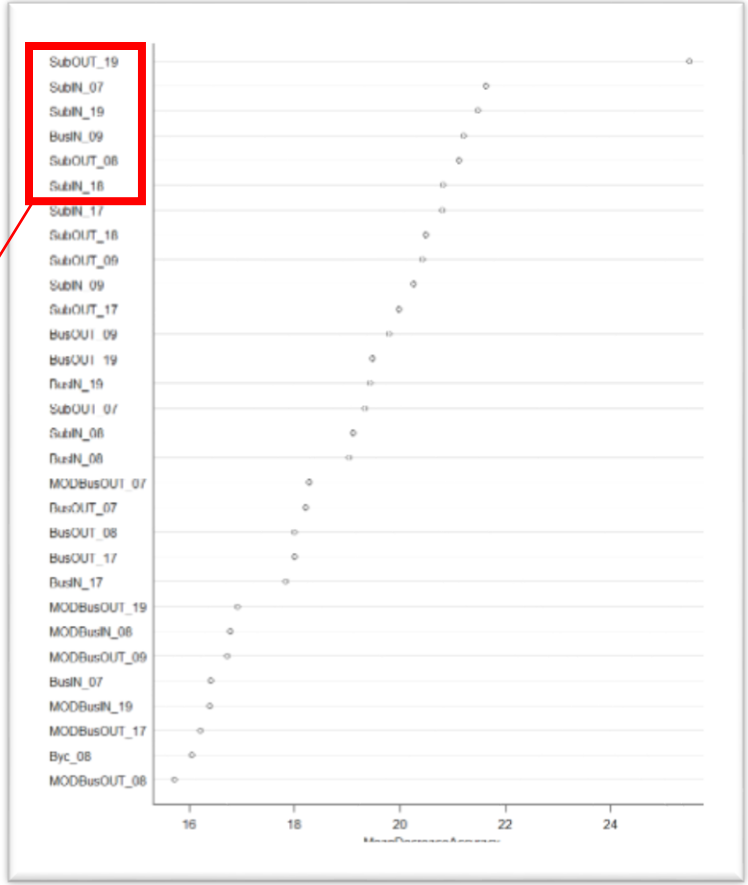
Random Forest

```
> model

Call:
 randomForest(formula = IsJAM ~ ., data = train, ntree = which.min(model$er
r.rate[,      1]), proximity = TRUE, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 383
No. of variables tried at each split: 6

        OOB estimate of  error rate: 6.67%
```

<과적합 방지>

- ntree 개수 : 383개
- mtry : 6개

주요 변수
- SubOUT_19
- SubIN_07
- SubIN_19
- BusIN_09
- SubOUT_08
- SubIN_18



데이터 분석

Random Forest

<Test1>

실제 \ 예측	0	1
0	50	**3**
1	2	41

정확도 : 94.79%

강남대로, 보문로, 세종대로

<Test4>

실제 \ 예측	0	1
0	50	**3**
1	3	40

정확도 : 93.75%

과천대로, 백범로, 보문로

<Test6>

실제 \ 예측	0	1
0	51	**2**
1	2	41

정확도 : 95.83%

보문로, 사직로

실제는 상습정체구간이 아니지만 Random Forest
예측결과 "**보문로**"가 상습정체구간으로 나옴

결론
활용 방안

교통 혼잡 개선 가능성

원인	원인 유형	개선 방안
도로 기하구조 측면, 도로 운영 측면 및 교통 수요와 공급의 불균형에 의해 발생	시간에 따른 원인	교통 신호 체계 효율화 대중 교통 시설 개선
	공간에 따른 원인	국지적 개선 사업 우회 도로 신설 지하 차도 활용
	교통량에 따른 원인	도로 신설 및 확장 교통 운영 개선 교통량 분산 및 우회 유도

결론
활용 방안

실제 변화에 빠르게 대응할 수 있는 교통 정책 수립 가능

투자 비용 대비 효과 높은 교통 수요 예측 방법
(이미 존재하는 데이터 활용)

기존의 교통 수요 추정 방법론 보완, 관련 연구 범위 확장
(학술적 기대효과)

결론
한계 및 애로사항

데이터 통합 기준의 회색 지대

도로명 추출 방식의 차이 (위도, 경도 이용 / 도로명 주소 이용)
지하철의 공간적 특성 (여러 도로가 모인 지점에 위치한 지하철 역)

이용 불가 데이터

택시 데이터, 카쉐어링 데이터 접근 불가능
범위 확장 제한 (대중 교통 -> 교통 수단)

최신 데이터 활용 불가능

따릉이 이용객 수 데이터 범위의 한계 (2017 ~ 2019.11)

감사합니다