산보연 과제 07월 15일 Version - Forecasting

작성자 : 이은경

< What to do >

: 사업장 중분류(UP2)별 설명변수를 YEAR, 반응변수를 질병 통합 누적 발생률로 하여 1차 spline, natural spline model 두 가지 모형 적합

: train data와 validation data는 YEAR 기준으로 split / train data는 2000년 ~ 2014년, validation data는 2015년 ~ 2018년 자료로 지정

: train data로 각각의 spline model 적합 후, validation data의 통합 누적 발생률 예측

: validation data의 연도에 따른 실제 통합 누적 발생률 알고 있으므로 예측값과 비교

: Performance criteria는 MAPE로 사용 $\left(\frac{1}{4}\sum_{t=2015}^{2018}|y_{it}-\hat{y_{it}}|/y_{it}\right)\times 100$ (이때, i는 각 사업장, t는 연도 의미)

(통합 누적 발생률 true value 중 0이 있어 MAPE값이 "NaN"값이 나오는 사업장이 있는 경우, 이 사업장의 MAPE는 "-99"로 대체)

: 적합한 spline model의 performance를 시각적으로 표현하기 위해 plots 생성

[Result 제시]

1) 1차 spline function 적용 (Knots : 2005년, 2010년)

① Lung Cancer Data

: 적합 후, validation set의 질병 통합 누적 발생률 예측 후, MAPE 계산. / MAPE를 오름차순으로 정렬

UP2	MAPE
2	59.78
36	33.15
45	22.27
6	21.59
51	20.96
34	18.84
87	17.62
39	16.50
37	16.44
42	16.20
63	15.42
11	12.91
91	11.72
90	11.60
33	10.58
16	9.59
76	9.57
73	8.67
50	8.10
1	7.96
12	7.03
35	6.98
66	6.84
96	6.75
85	6.65
86	6.56
84	6.33
59	6.10
99	6.01
21	5.98
18	5.97
56	5.92

UP2	MAPE
23	5.72
25	5.68
19	5.62
32	5.42
70	5.24
75	5.18
60	4.95
61	4.35
27	4.01
71	3.94
7	3.93
28	3.84
46	3.76
10	3.72
49	3.62
3	3.57
74	3.53
65	3.47
55	3.40
30	3.23
52	3.22
41	3.21
62	3.19
5	3.16
17	3.09
20	3.08
47	2.88
29	2.80
15	2.56
14	2.47
24	2.45
72	2.43

UP2	MAPE
95	2.43
22	2.37
38	2.17
94	2.12
13	1.79
26	1.77
31	1.68
58	1.60
64	1.36
68	1.30
8	-99
97	-99
98	-99

: Lung Cancer Data에서 2000년 ~ 2014년 자료를 가지고 통합 누적 발생률에 대해 YEAR을 설명변수로 하여 1차 spline 함수를 적합한 후, 2015년 ~ 2018년 자료의 통합 누적 발생률을 예측하고 MAPE를 계산해 보았다. 그 결과, 사업장 "2"(임업)의 MAPE 값이 59.78로 가장 큰 값을 보였다. 그리고 단순선형회귀모형을 적합했을 때, 가장 큰 MAPE를 보인 사업장 "39"(환경 정화업)이 16.50으로 MAPE 값이 낮아진 것을 확인하였다. 단순선형회귀모형 적합 때와 결과를 비교해보면 MAPE 값이 전체적으로 낮아진 것을 확인할 수 있고, 사업장 "8"(광업 - 광업, 자원, 원유), 사업장 "97", "98"(가구 내 고용활동)은 MAPE가 "NaN" 값을 보인다.

: 1차 spline function 적합, 예측한 결과를 시각화

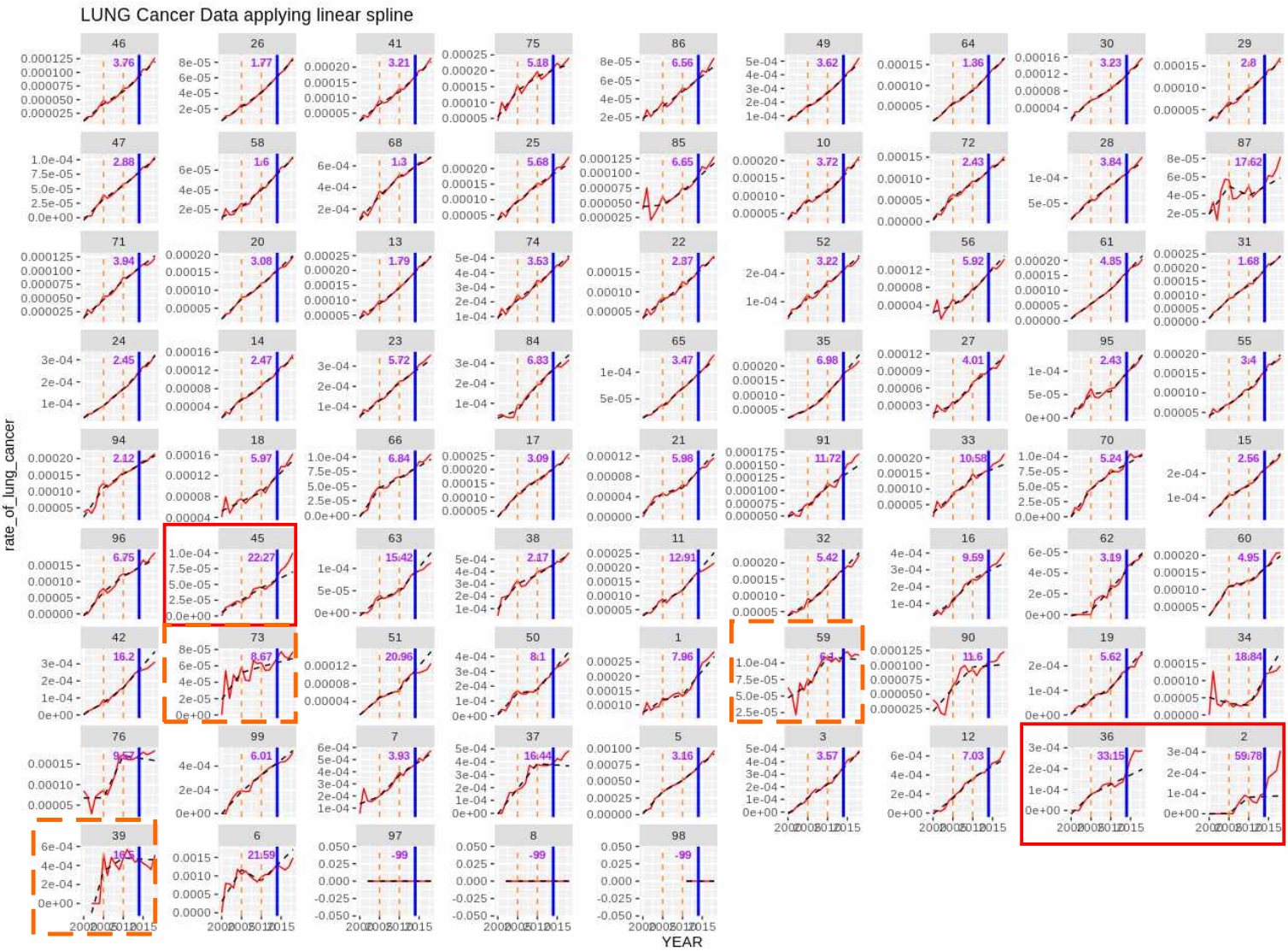
1) Version 1

: **빨간색 실선**이 **실제 통합 누적 발생률 값**이고, **검은색 점선**은 적합한 1차 spline function으로 **예측한 값**이다.

: **파란색 수직선**은 YEAR=2014를 의미하는 선이며, **주황색 수직선**은 YEAR=2005, 2010 "knots"를 의미한다.

: 각 그래프 안 **보라색 글씨는 각 사업장의 MAPE**를 의미한다.

: 그래프의 순서는 2018년 추적 인년 합계 순위이다. (오름차순 정렬)



: Knots를 2005년, 2010년으로 하여 1차 spline function을 적용해보니, 연도별 급격히 변하는 추세를 어느 정도 학습한 것으로 보인다.

: MAPE가 큰 사업장 TOP3에 대해 **빨간색 테두리**를 그려놓았다.

: 해당 사업장에 대한 정보는 아래 표와 같다.

UP2	사업장명	MAPE	MAPE가 큰 이유 추측
2	임업	59.78	통합 누적 발생률 추세가 2014년 이후 급격히 증가
36	용수 공급업	33.15	통합 누적 발생률 추세가 2014년 이후 급격히 증가
45	자동차 및 부품 판매업	22.27	통합 누적 발생률 추세가 2014년 이후 급격히 증가

: 2014년을 기점으로 통합 누적 발생률의 추세가 급격히 변하는 case의 MAPE는 높은 경향을 보인다.

: 추세가 불안정하더라도, 급격히 증가 혹은 감소를 하는 부분이 없으면 MAPE가 높지 않은 값을 보인다.

(예시 : 사업장 "73"(수의업, 디자인업), 사업장 "39"(환경 정화업), 사업장 "59"(영화, 비디오, 오디오물 출판, 상영업)의 경우, 연도별 추세 변화가 많이 불안정하나, 급격히 감소 혹은 증가하는 너비가 크지 않아, MAPE가 예상보다 많이 높지 않다. - 주황색 테두리)

2) Version 2

: Version 1의 그래프에서 validation data 시각화 부분이 눈에 띄지 않아, 2015년 ~ 2018년에 대해 각 사업장(UP2)별 실제 통합 누적 발생률과 적합한 모형 통해 예측한 값 plotting

: **빨간색 실선**이 **실제 통합 누적 발생률 값**이고, **파란색 점선**은 적합한 1차 spline function으로 **예측한 값**이다.

: 각 그래프 안 **보라색 글씨는 각 사업장의 MAPE**를 의미한다.

: 그래프의 순서는 2018년 추적 인년 합계 순위이다. (오름차순 정렬)

: MAPE가 가장 작은 사업장 TOP3는 **파란색** 테두리로, MAPE가 가장 큰 사업장 TOP3는 **빨간색** 테두리로 표시해 놓았다. MAPE가 큰 사업장들의 그래프를 보면, 실제 통합 누적 발생률 변화 추세와 1차 spline function이 예측한 추세와 차이가 많이 나는 것을 볼 수 있다. 해당 사업장에 대한 정보는 다음과 같다.

UP2	사업장명	MAPE
2	임업	59.78
36	용수 공급업	33.15
45	자동차 및 부품 판매업	22.27
58	출판업	1.60
64	금융업	1.36
68	부동산업	1.30

---------- 다음 페이지로 ----------

② Leukemia Data

: 적합 후, validation set의 질병 통합 누적 발생률 예측 후, MAPE 계산. / MAPE를 오름차순으로 정렬

UP2	MAPE
39	152.70
97	100
2	79.96
59	68.26
3	57.82
50	36.66
19	34.45
36	31.15
87	25.85
62	24.03
34	23.79
60	22.47
73	20.35
35	19.59
63	18.92
12	18.87
42	18.24
51	13.91
11	13.72
70	13.65
7	13.48
90	11.65
10	10.97
75	10.51
91	10.45
76	10.34
99	10.27
18	10.18
23	10.17
22	9.84
13	9.72
5	9.46

UP2	MAPE
37	9.14
58	8.88
45	8.59
66	8.55
95	8.55
52	8.45
27	8.01
17	6.71
31	6.70
65	6.14
71	6.11
55	5.78
64	5.73
33	5.46
32	5.46
96	5.38
68	5.37
21	5.35
38	5.16
56	5.12
84	5.06
94	4.94
86	4.53
61	4.36
16	4.06
15	4.03
74	4.03
72	4.01
29	3.94
30	3.78
46	3.76
49	3.73

UP2	MAPE
47	3.65
24	3.60
28	3.29
14	3.20
20	3.08
25	2.97
1	2.25
41	2.00
26	1.45
6	-99
8	-99
98	-99

: Leukemia Data에 대해 2000년 ~ 2014년 자료를 가지고 통합 누적 발생률에 대해 YEAR을 설명변수로 하여 1차 spline function을 적합하고, 2015년 ~ 2018년 자료의 통합 누적 발생률을 예측한 결과, 사업장 “39”(환경 정화업)의 MAPE가 152.7로 가장 크게 나왔다. 이 사업장은 이전 EDA 과정에서 특별히 주목하지 않았던 사업장이었고, 단순선형회귀모형을 적합했을 때는 MAPE가 21.76이었다. 이 사업장에는 단순한 모형의 예측성이 더 나아 보인다. 한편, 사업장 “97”(가구 내 고용활동)은 단순선형회귀모형을 적합했을 때와 1차 spline function을 적용했을 때 MAPE가 모두 100으로 동일하게 나왔다. 사업장 “6”(광업 - 금속, 철, 비금속)은 간접 SIR과 2018년 기준 백혈병 통합 누적 발생률이 2번째로 큰 사업장으로 판단되었었는데, MAPE 값이 “NaN”으로 나와 모형의 적합 정도를 판단하기가 어렵다. 그리고, 사업장 “8”(광업 - 광업, 자원, 원유)과 사업장 “98”(가사 생산 활동)은 폐암 데이터에서와 마찬가지로 MAPE 값이 “NaN”인 것을 확인하였다.

또한, 전체적인 경향을 보았을 때 단순선형회귀모형 적합 후 측정한 MAPE 보다는 낮은 값을 보인다.

---------- 다음 페이지로 ----------

: 1차 spline function 적합, 예측한 결과를 시각화

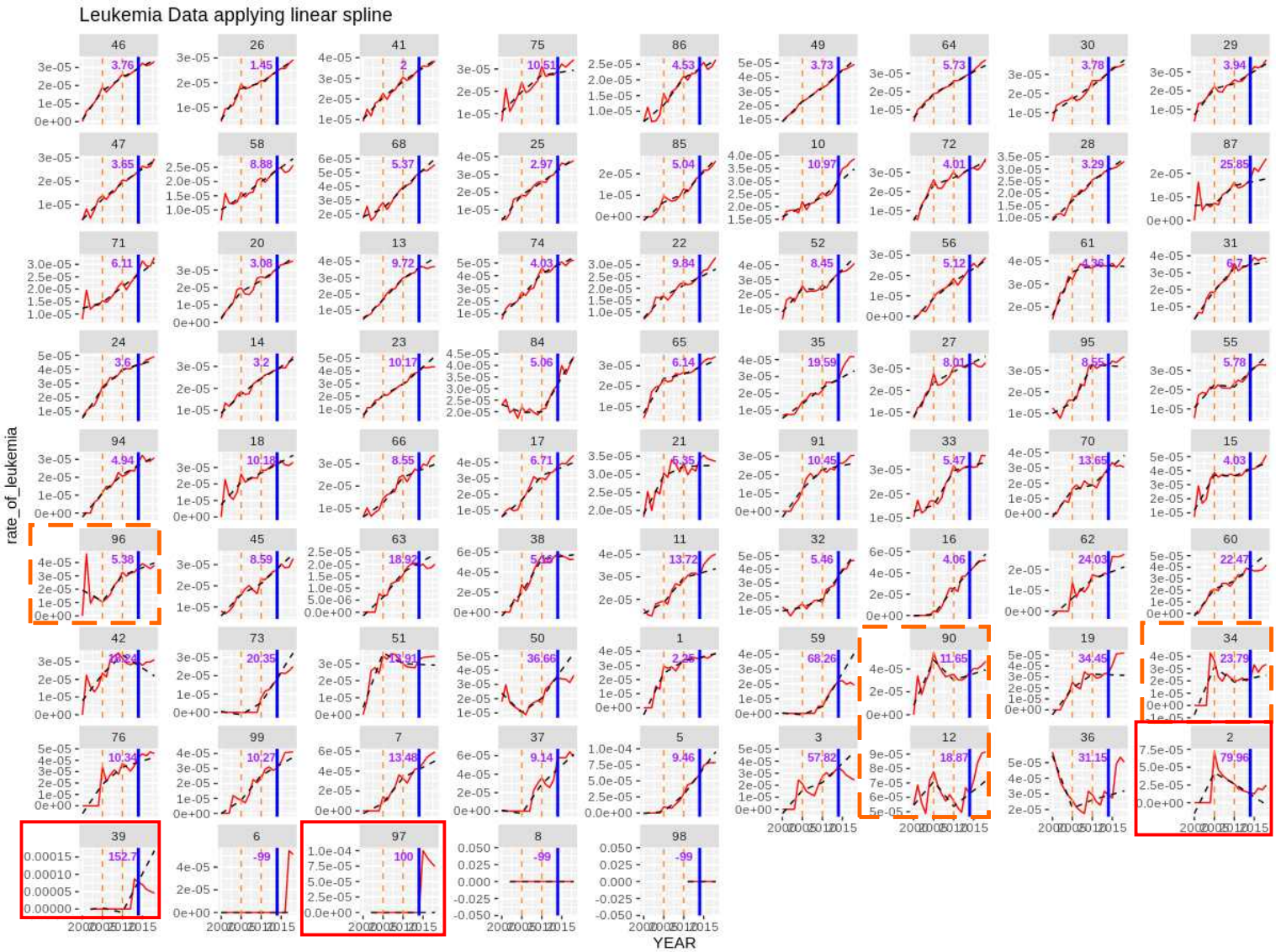
1) Version 1

: **빨간색 실선**이 **실제 통합 누적 발생률 값**이고, **검은색 점선**은 적합한 1차 spline function으로 **예측한 값**이다.

: **파란색 수직선**은 YEAR=2014를 의미하는 선이며, **주황색 수직선**은 YEAR=2005, 2010 "knots"를 의미한다.

: 각 그래프 안 **보라색 글씨는 각 사업장의 MAPE**를 의미한다.

: 그래프의 순서는 2018년 추적 인년 합계 순위이다. (오름차순 정렬)



: Knots를 2005년, 2010년으로 하여 1차 spline function을 적용해보니, 연도별 급격히 변하는 추세를 어느 정도 학습한 것으로 보인다.

: MAPE가 큰 사업장 TOP3에 대해 **빨간색 테두리**를 그려놓았다.

: 해당 사업장에 대한 정보는 아래 표와 같다.

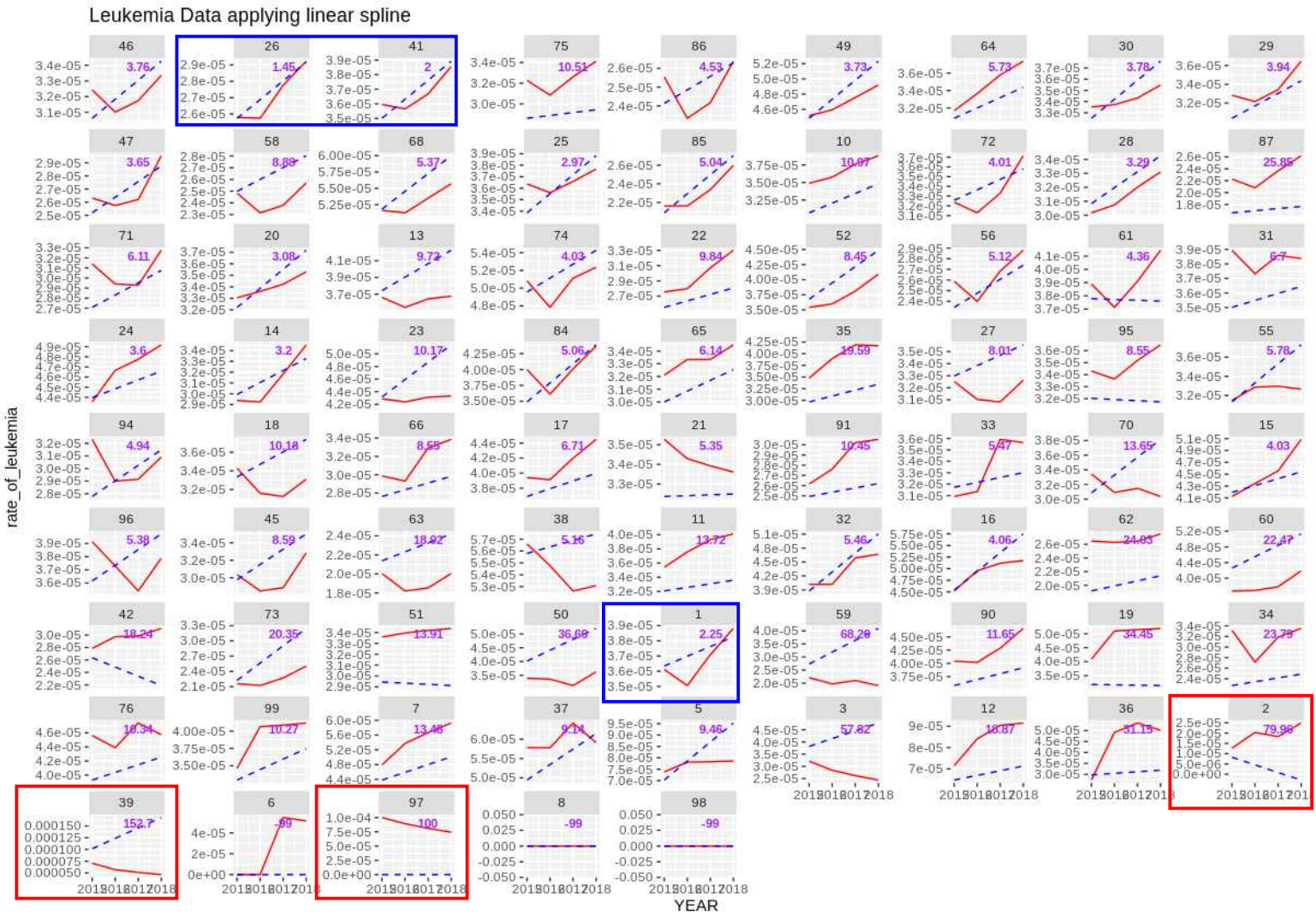
UP2	사업장명	MAPE	MAPE가 큰 이유 추측
39	환경 정화업	152.7	2012년 ~ 2013년 즈음에 통합 누적 발생률이 급격하게 증가하여 2014년 이후에도 추세가 증가할 것이라 예측 하였으나, 실제 통합 누적 발생률은 2014년 이후 감소하는 추세를 보였다.
97	가구 내 고용 활동	100	2014년 이전까지 통합 누적 발생률이 "0"이었기 때문에 2014년 이후에도 동일한 추세 보일 것이라 예상하였으나, 2014년 이후 통합 누적 발생률이 급격히 증가하는 추세를 보였다.
2	임업	79.96	2005년을 기준으로 통합 누적 발생률이 증가하다 감소하는 추세를 보이기 때문에 2014년 이후에도 추세가 감소할 것이라 예상했으나, 실제로는 증가하였다.

: 2014년을 기점으로 통합 누적 발생률의 추세가 급격히 변하는 case의 MAPE는 높은 경향을 보인다.
: 추세가 불안정하더라도, MAPE가 높지 않은 값을 보이는 case도 있다.
(예시 : 사업장 "96"(기타 개인 서비스업), 사업장 "90"(예술, 스포츠업, 건설업 등 각종 사업), 사업장 "12"(제조업 - 담배), 사업장 "34"(산업용 기계 및 장비수리업)의 경우, 연도별 추세 변화가 많이 불안정하나, 2014년을 기준으로 추세 변화의 방향이 정반대이거나 급격히 감소 혹은 증가하는 너비가 크지 않아, MAPE가 예상보다 많이 높지 않다. - 주황색 테두리)

---

2) Version 2

: Version 1의 그래프에서 validation data 시각화 부분이 눈에 띄지 않아, 2015년 ~ 2018년에 대해 각 사업장(UP2)별 실제 통합 누적 발생률과 적합한 모형 통해 예측한 값 plotting
: **빨간색 실선**이 **실제 통합 누적 발생률 값**이고, **파란색 점선**은 적합한 1차 spline function으로 **예측한 값**이다.
: 각 그래프 안 **보라색 글씨는 각 사업장의 MAPE**를 의미한다.
: 그래프의 순서는 2018년 추적 인년 합계 순위이다. (오름차순 정렬)



: MAPE가 가장 작은 사업장 TOP3는 **파란색** 테두리로, MAPE가 가장 큰 사업장 TOP3는 **빨간색** 테두리로 표시해 놓았다. MAPE가 큰 사업장들의 그래프를 보면, 실제 통합 누적 발생률 변화 추세와 1차 spline function이 예측한 추세와 차이가 많이 나는 것을 볼 수 있다. 해당 사업장에 대한 정보는 다음과 같다.

UP2	사업장명	MAPE
39	환경 정화업	152.7
97	가구 내 고용 활동	100
2	임업	2
1	농업	2.25
41	종합 건설업	2.00
26	전자부품, 컴퓨터, 영상, 음향 및 통신장비 제조업	1.45