산보연 Data forecasting 07월 22일 Version

작성자 : 이은경

< What to do >

: Data level 별(level4 - UP1, UP2, SEX로 grouping , level5 - UP1, UP2, SEX, CAL2로 grouping) 설명변수를 YEAR, 반응변수를 질병 통합 누적 발생률로 하여 단순선형회귀모형, 1차 spline 두 가지 모형 적합

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

: train data로 각각의 단순선형회귀모형, 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 $\left(\frac{1}{4}\sum_{t=2015}^{2018}|y_{it}-\widehat{y_{it}}|/y_{it}\right)\times 100$ (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)

: Data level별 적합한 단순선형회귀모형, spline model의 performance를 시각적으로 표현하기 위해 plots 생성

[Result 제시]

1) level 4 Data

: level 4 data의 경우, 이전과 다르게 SEX를 기준으로 한 grouping이 추가되었다. 따라서, 결과를 SEX에 따라 나누어 제시하려고 한다. (구분은 level 4 data의 "NAME" 변수 이용)

① Leukemia Data

: 각 "NAME" 별 단순선형회귀모형, 1차 spline function을 적용한 후에 MAPE를 계산한 뒤, 두 모형 중 더 작은 MAPE를 가지는 값을 따로 저장하였다. 결과는 아래 표와 같다. (더 작은 MAPE 값을 주는 모형 이름은 제시하지 않고, 뒤 그래프로 제시)

NAME	Female	Male
XA1	13.1	0.74
XA2	-99	77.9
XA3	8.59	28.4
XB5	30.63	7.67
XB6	-99	-99
XB7	-99	0.81
XB8	-99	-99
XC10	15.14	8.8
XC11	4.1	2.43
XC12	11.24	20.01
XC13	11.4	8.4
XC14	5.65	10.83

NAME	Female	Male
XC15	4.83	11.56
XC16	38.77	6.12
XC17	11.37	2.48
XC18	41.25	2.28
XC19	32.27	9.36
XC20	9.13	3
XC21	5.71	6.17
XC22	17.36	5.34
XC23	6.76	9.67
XC24	5.7	4.75
XC25	5.47	2.01
XC26	2.13	2.59

NAME	Female	Male
XC27	8.88	8.22
XC28	3.43	5
XC29	2.28	1.74
XC30	4.27	2.08
XC31	3.04	4.65
XC32	26.98	2.93
XC33	4	10.72
XC34	-99	6.52
XD35	77.55	15.49
XD36	-99	31.33
XE37	-99	8.61
XE38	12.98	8.98

NAME	Female	Male
XE39	21,94	-99
XF41	5.24	1
XF42	20.43	22.08
XG45	15.36	10.92
XG46	2.98	2.13
XG47	2.54	2.57
XH49	3.97	3.48
XH50	58.26	10.62
XH51	9.16	4.02
XH52	10.73	7.13
XI55	5.76	3.41
XI56	10.48	15.97

NAME	Female	Male
XJ58	20.03	6.92
XJ59	-99	13.8
XJ60	23.99	14.97
XJ61	13.8	9.73
XJ62	21.85	34.31
XJ63	11.91	6.22
XK64	2.61	2.05
XK65	4.99	3.48
XK66	25.22	1.75
XL68	4.35	3.2
XM70	15.49	4.96
XM71	4.52	5.91

NAME	Female	Male
XM72	7.66	4.09
XM73	8.61	29.39
XN74	7.92	1.81
XN75	3.34	1.03
XN76	13.83	10.11
XO84	5.06	4.21
XP85	6.12	8.3
XQ86	7	5.43
XQ87	9.35	16.14
XR90	10.03	7.91
XR91	36.26	6.24
XS94	10.7	4.31

NAME	Female	Male
XS95	9.6	6.59
XS96	8.73	5.86
XT97	-99	100
XT98	-99	-99
XU99	-99	4.31

---

1) **파란색으로 표시한 부분**은 UP1, UP2, 동일한 사업장을 SEX로 나누었을 때, 한 성별의 MAPE 값만 "NaN" 값을 보이는 경우를 의미한다. 이에 대한 정보는 다음과 같다.

NAME	사업장명	통합 누적 발생률에 0이 있는 성별
A2	임업	여성
B7	비금속광물 광업; 연료용 제외	여성
C34	산업용 기계 및 장비 수리업	여성
D36	수도업	여성
E37	하수, 폐수 및 분뇨 처리업	여성
E39	환경 정화 및 복원업	남성
J59	영상, 오디오 기록물 제작 및 배급업	여성
T97	가구 내 고용 활동	여성
U99	국제 및 외국기관	여성

: 동일한 사업장에 대해 성별로 나누었을 경우 여성의 통합 누적 발생률이 0인 경우가 대부분이다.

2) **보라색으로 표시한 부분**은 UP1, UP2, 동일한 사업장을 SEX로 나누었을 때, 성별 간 MAPE의 차이가 낮은 사업장을 의미한다. 이에 대한 정보는 다음과 같다.

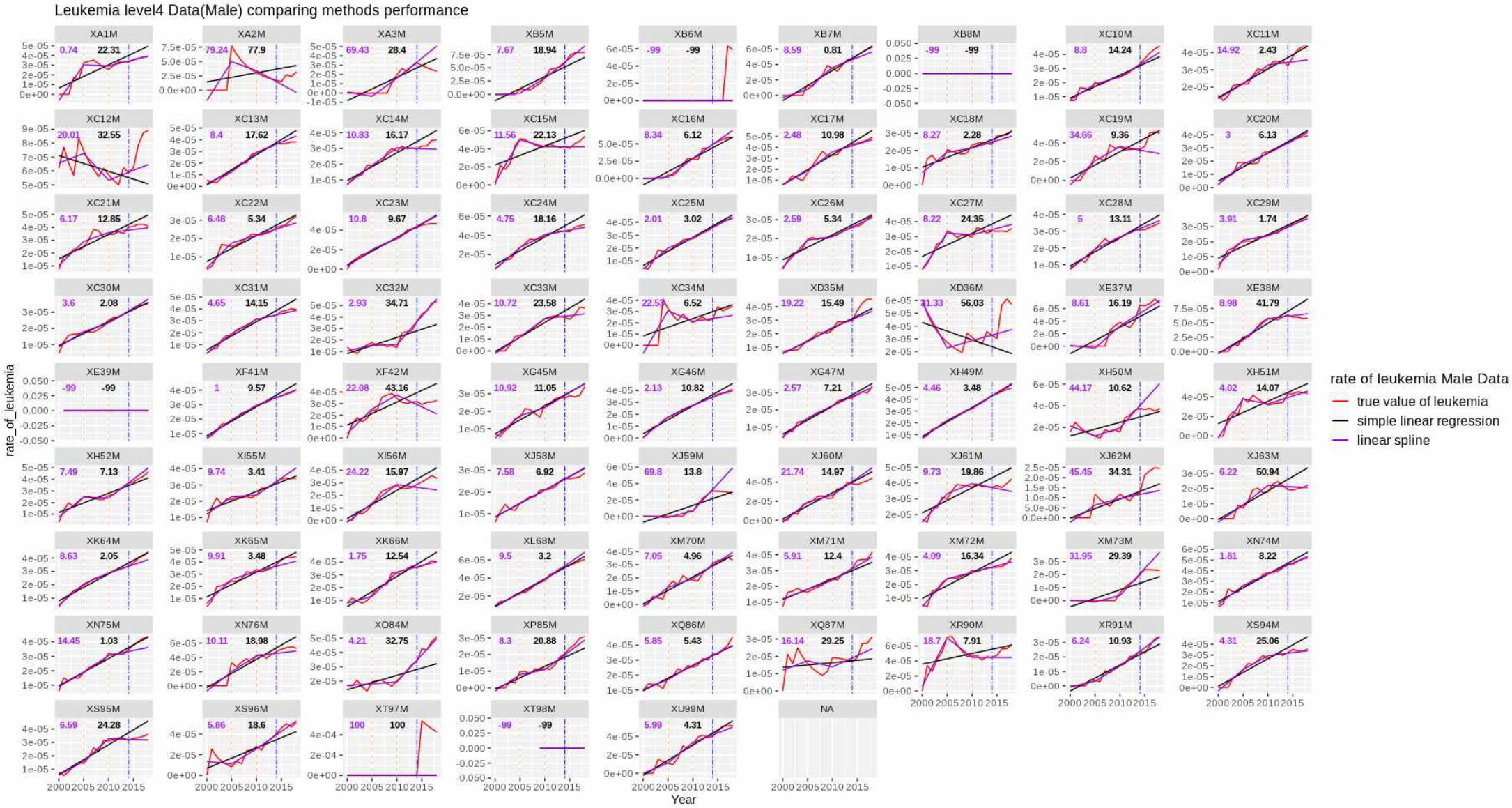
NAME	사업장명	성별 간 MAPE 차이의 절댓값
G47	소매업 ; 자동차 제외	0.03
C26	전자부품, 컴퓨터, 영상, 음향 및 통신장비 제조업	0.46
C21	의료용 물질 및 의약품 제조업	0.46

: 해당 사업장들은 연도가 변함에 따라 성별 구성원 비율의 차이가 크지 않고 근로자들의 성 비율이 비슷해 성별 간 MAPE 차이가 적은 것으로 추측된다. (Data check 7월 4일 추가사항 version 참고)

3) **빨간색으로 밑줄을 그어 놓은 부분**은 UP1, UP2, 동일한 사업장을 SEX로 나누었을 때, 성별 간 MAPE의 차이가 큰 사업장을 의미한다.

NAME	사업장명	성별 간 MAPE 차이의 절댓값
B7	비금속광물 광업; 연료용 제외	98.19
U99	국제 및 외국기관	94.69
C34	산업용 기계 및 장비 수리업	92.48

--- Level 4 Male Leukemia data plot --- (그래프 순서는 2018년 기준 추적 인년 합계 순위)



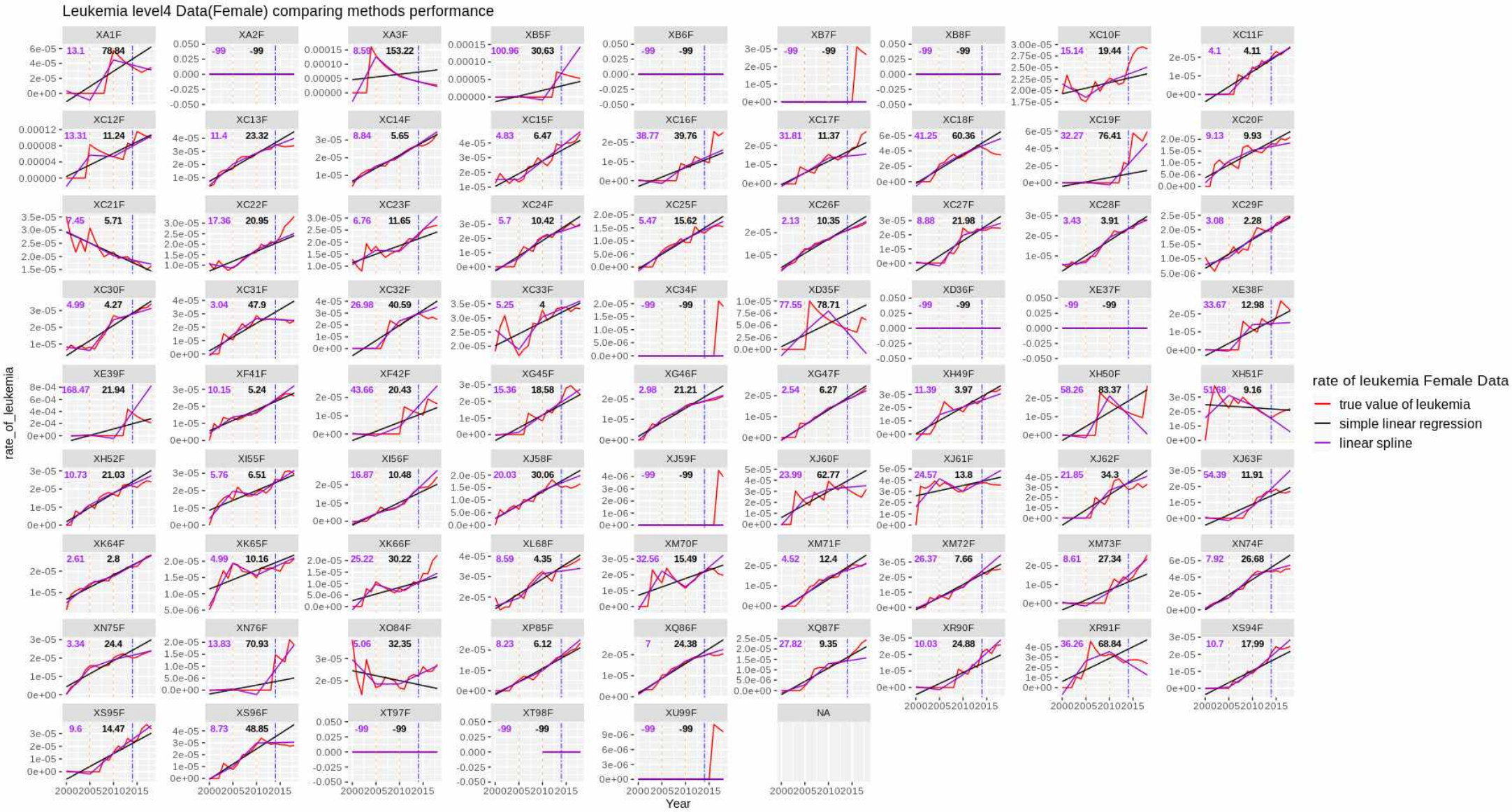
: **빨간색 실선**이 **실제 통합 누적 발생률 값**이고, **검은색 점선**은 적합한 **단순선형회귀모형**으로 **예측한 값**이며, **보라색 실선**은 **1차 spline function 적합 통해 예측한 값**을 의미한다.

: **파란색 수직**은 **YEAR=2014**를 의미하는 선이며, **주황색 수직선**은 **YEAR=2005, 2010년** 1차 spline function의 "knots"를 의미한다.

: 그래프의 순서는 2018년 기준 UP1 기준으로 grouping 한 후 계산한 추적 인년 합계의 순위이다. (높은 값에서 낮은 값 순으로 정렬)

: 그래프 안, **보라색 글씨**는 **1차 spline function 적합 통해 얻은 MAPE 값**, **검은색 글씨**는 **단순선형회귀모형 적합 통해 얻은 MAPE 값**을 의미한다.

--- Level 4 Female Leukemia data plot --- (그래프 순서는 2018년 기준 추적 인년 합계 순위)



: Leukemia level 4 data를 성별에 따라 나누어 제시한 plot을 보면 주로 Male의 추세 변화가 좀 더 안정적인 것을 알 수 있다.

② Lung cancer Data

: 각 "NAME" 별 단순선형회귀모형, 1차 spline function을 적용한 후에 MAPE를 계산한 뒤, 두 모형 중 더 작은 MAPE를 가지는 값을 따로 저장하였다. 결과는 아래 표와 같다. (더 작은 MAPE 값을 주는 모형 이름은 제시하지 않고, 뒤 그래프로 제시)

NAME	Female	Male
XB6	-99	11.91
XE39	-99	16
XC34	-99	27.8
XD36	100	23.12
XF42	59.84	0.76
XA2	-99	48.25
XB5	40.38	2.45
XA3	38.6	5.91
XU99	29.99	3.63
XM70	27.52	2.77
XQ87	25.83	2.65
XC19	29.74	7.24
XK66	23.7	1.75
XM73	24.97	4.66
XA1	24.17	5.31
XC31	19.21	0.59
XE38	16.38	1.96
XJ60	20.45	6.6
XC16	15.77	2.35
XC17	14.65	2.21
XH51	20.39	8.45

NAME	Female	Male
XC25	12.22	1.28
XD35	17.12	6.51
XP85	11.39	0.96
XC11	13.19	2.85
XB7	11.57	1.76
XC21	21.65	11.96
XH52	11.85	2.4
XM71	11.86	2.89
XN76	18.53	9.72
XJ58	8.82	1
XE37	6.87	14.47
XG45	12.36	19.9
XJ59	13.41	6.34
XC26	10.57	3.63
XS94	7.41	0.69
XR91	3.77	9.96
XM72	6.5	0.87
XL68	6.41	1.39
XJ63	4.4	8.97
XC10	7	2.57
XC32	6.73	2.32

NAME	Female	Male
XC24	5.58	1.47
XC29	5.8	1.85
XI56	3.26	7.2
XJ62	7.77	3.89
XR90	17.5	13.71
XO84	4.12	0.35
XK64	4.54	1.03
XI55	5.87	2.64
XC27	7.06	4.17
XC30	2.33	4.89
XH50	10.28	7.77
XH49	5.91	3.41
XS95	4.62	2.33
XN75	5.25	3.36
XC18	7.34	5.67
XK65	3.67	5.31
XN74	1.54	3.13
XC13	3.67	2.4
XQ86	3.88	2.75
XG46	2.51	3.58
XC22	1.96	1.04

NAME	Female	Male
XS96	2.31	3.22
XC23	2.31	3.16
XC33	3.25	3.99
XC28	2.04	2.63
XC20	1.57	2.13
XC15	3.06	3.53
XJ61	2.93	3.36
XG47	2.7	2.34
XF41	2.75	2.72
XB8	-99	-99
XT97	-99	-99
XT98	-99	-99

---

1) **파란색으로 표시한 부분**은 UP1, UP2, 동일한 사업장을 SEX로 나누었을 때, 한 성별의 MAPE 값만 "NaN" 값을 보이는 경우를 의미한다. 이에 대한 정보는 다음과 같다.

NAME	사업장명	통합 누적 발생률에 0이 있는 성별
B6	금속광업	여성
E39	환경 정화 및 복원업	여성
C34	산업용 기계 및 장비 수리업	여성
A2	임업	여성

: **초록색으로 강조한 사업장**은 백혈병 통합 누적 발생률 파악한 결과에서도 "여성"의 MAPE가 "NaN"값을 보였다. (즉, 2015년과 2018년 사이에 해당 사업장의 여성 근로자들의 통합 누적 발생률이 0인 경우가 있음을 의미)

: 주로 여성 근로자들의 통합 누적 발생률 값에 0이 포함되어 있다.

2) **보라색으로 표시한 부분**은 UP1, UP2, 동일한 사업장을 SEX로 나누었을 때, 성별 간 MAPE의 차이가 낮은 사업장을 의미한다. (MAPE 값이 모두 "NaN"인 사업장은 제외) 이에 대한 정보는 다음과 같다.

NAME	사업장명	성별 간 MAPE 차이의 절댓값
J61	우편 및 통신업	0.43
G47	소매업 ; 자동차 제외	0.36
F41	종합 건설업	0.03

: **주황색으로 강조한 사업장**은 관심 질병을 백혈병으로 두어 파악했을 때에도 성별 간 MAPE의 차이가 작았다.
: 성별 간 MAPE의 차이가 적은 이유를 추측해보자면, 사업장 "47", "61"의 경우에는 연도가 변함에 따라 성별 비율의 변화가 비슷하고, 성별 비율이 거의 비슷한 점인 것 같고, 사업장 "41"의 경우에는 남성 근로자의 비율이 압도적이나, 연도가 변함에 따라 성별 비율의 변화가 크지 않고 추적 인년 합계가 큰 사업장에 해당하기 때문에 추세의 변화가 안정적이라 예상된다.

3) **빨간색으로 밑줄을 그어 놓은 부분**은 UP1, UP2, 동일한 사업장을 SEX로 나누었을 때, 성별 간 MAPE의 차이가 큰 사업장을 의미한다.

NAME	사업장명	성별 간 MAPE 차이의 절댓값
B6	금속 광업	87.09
E39	환경 정화 및 복원업	83
C34	산업용 기계 및 장비 수리업	92.48

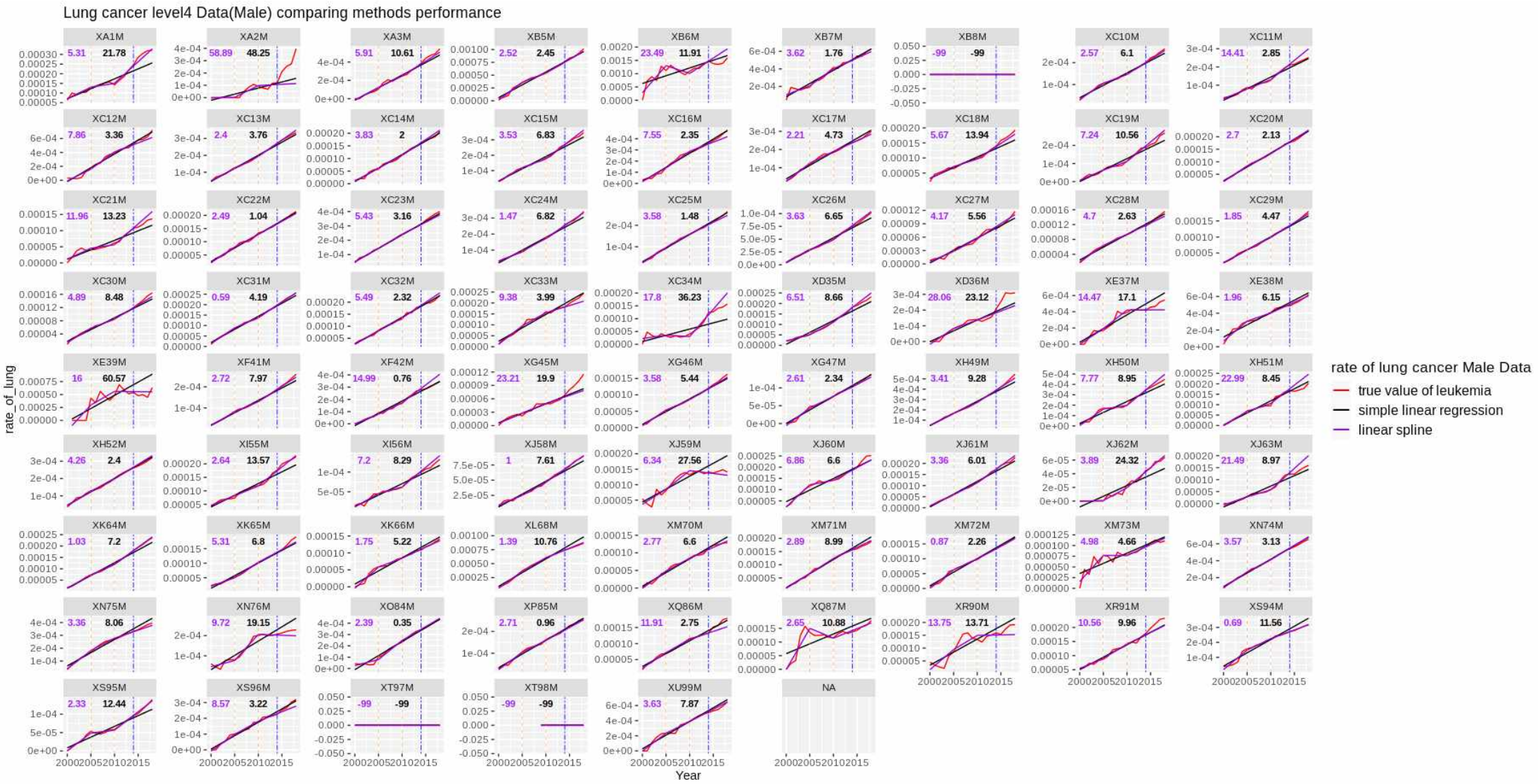
: **주황색으로 강조한 사업장**은 관심 질병을 백혈병으로 두어 파악했을 때에도 성별 간 MAPE의 차이가 큰 값을 가졌다.

참고사항

: Leukemia data와 달리 Lung cancer data는 UP1, UP2가 동일한 사업장을 SEX로 나누었을 때, 주로 남성 근로자들의 MAPE 값이 여성 근로자들의 MAPE 보다 낮았다.

---------- 다음 페이지로 ----------

--- Level 4 Male Lung cancer data plot --- (그래프 순서는 2018년 기준 추적 인년 합계 순위)



: 앞의 "Level 4 Male Leukemia Data"와 비교해보면, 통합 누적 발생률 추세가 더 안정적인 것을 알 수 있다.

---------- 다음 페이지로 ----------

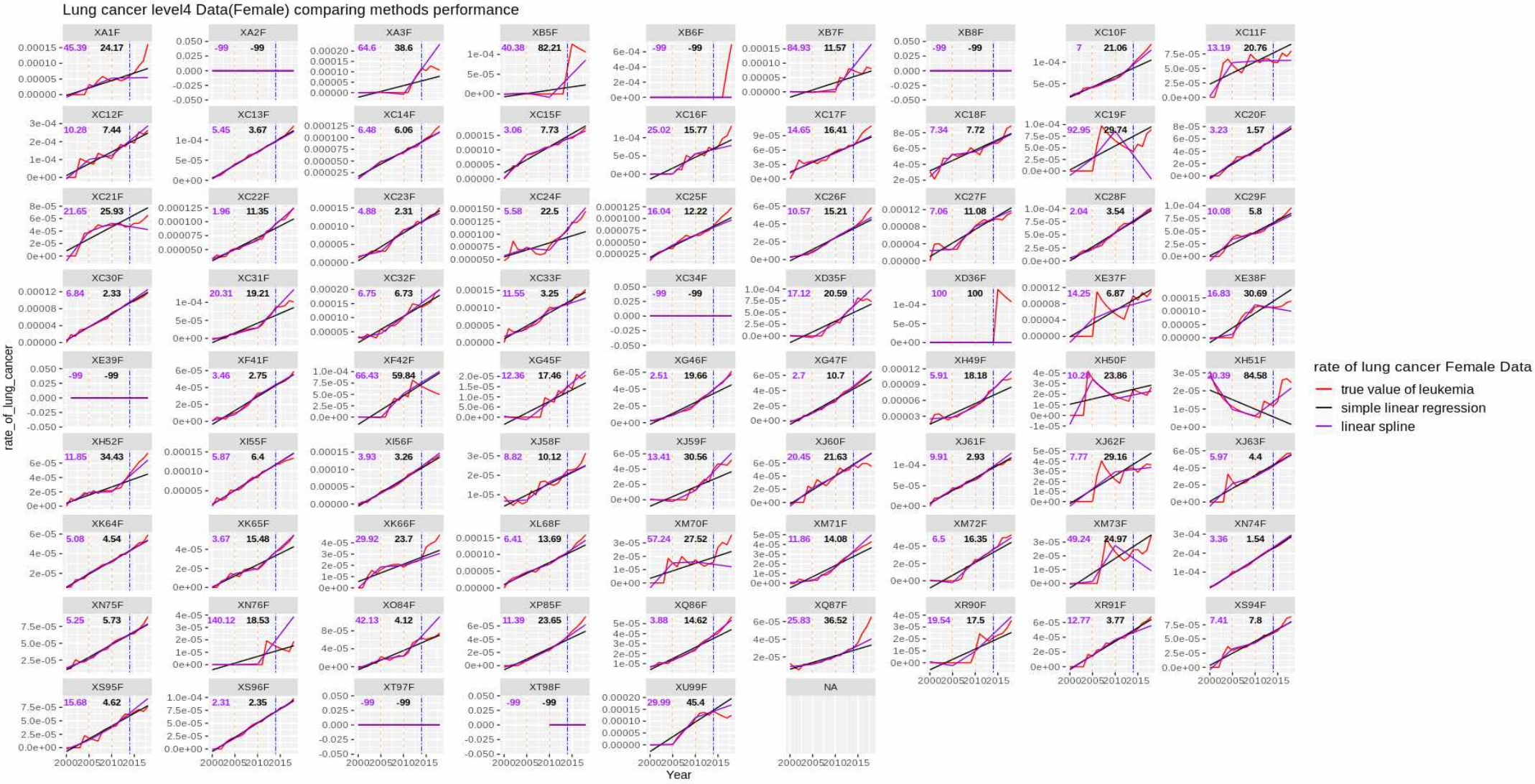
: **빨간색 실선**이 **실제 통합 누적 발생률 값**이고, **검은색 점선**은 적합한 **단순선형회귀모형**으로 **예측한 값**이며, **보라색 실선**은 **1차 spline function 적합 통해 예측한 값**을 의미한다.

: **파란색 수직**은 **YEAR=2014**를 의미하는 선이며, **주황색 수직선**은 **YEAR=2005, 2010년** 1차 spline function의 "knots"를 의미한다.

: 그래프의 순서는 2018년 기준 UP1 기준으로 grouping 한 후 계산한 추적 인년 합계의 순위이다. (높은 값에서 낮은 값 순으로 정렬)

: 그래프 안, **보라색 글씨**는 **1차 spline function 적합 통해 얻은 MAPE 값**, **검은색 글씨**는 **단순선형회귀모형 적합 통해 얻은 MAPE 값**을 의미한다.

--- Level 4 Female Lung cancer data plot --- (그래프 순서는 2018년 기준 추적 인년 합계 순위)



: "Level 4 Male Lung cancer data"와 비교해보면, 남성 근로자들에 비해 여성 근로자들의 통합 누적 발생률 추세가 더 불안정한 것을 확인할 수 있다.

1) level 5 Data

: level 5 data의 경우, SEX와 CAL(입사 시기 : 이때, CAL의 범주가 5개이므로 2000년을 기준으로 2가지 범주로 좁힘 - CAL2라는 새로운 변수 생성)을 기준으로 한 grouping이 추가되었다. 따라서, 결과를 SEX와 CAL2에 따라 나누어 제시하려고 한다. (구분은 level 5 data의 "NAME" 변수 이용)

① Leukemia Data

: 각 "NAME" 별 단순선형회귀모형, 1차 spline function을 적용한 후에 MAPE를 계산한 뒤, 두 모형 중 더 작은 MAPE를 가지는 값을 따로 저장하였다. 결과는 아래 표와 같다. (더 작은 MAPE 값을 주는 모형 이름은 제시하지 않고, 뒤 그래프로 제시)

: 표에 제시되는 "Female01"은 여성이면서 2000년 전에 입사한 근로자, "Female234"는 여성이면서 2000년 이후에 입사한 근로자, "Male01"은 남성이면서 2000년 전에 입사한 근로자, "Male234"는 남성이면서 2000년 이후에 입사한 근로자를 지칭한다.

NAME	Female01	Female234	Male01	Male234
XA1	13.51	-99	18.12	14.49
XA2	-99	-99	31.14	-99
XA3	5.8	-99	22.05	20.22
XB5	32.82	-99	6.89	31.39
XB6	-99	-99	-99	-99
XB7	-99	-99	4.35	6.35
XB8	-99	-99	-99	-99
XC10	2.33	12.78	5.1	5.68
XC11	22.24	100	6.86	47.25
XC12	12.06	-99	22.37	53.42
XC13	5.12	17.89	4.68	16.93
XC14	11.52	6.07	6.05	18.64
XC15	2.39	33.29	7.03	31.47
XC16	43.83	34.15	9.66	4.77
XC17	11.54	25.21	1.14	15.48
XC18	29.66	12.65	7.42	10.02
XC19	16.61	100	6.28	-99
XC20	5.35	24.86	10.14	15.68
XC21	3.53	35.83	2.23	5.55
XC22	16.77	9.02	1.75	23.01
XC23	13.38	42.93	3.73	18.28
XC24	5.74	80.37	2.71	21
XC25	21.31	6.73	1.28	4.84
XC26	5.77	6.87	1.28	4.84
XC27	19.16	22.23	1.94	18.21
XC28	14.92	7.29	10.71	6.22
XC29	6.2	2.39	3.6	3.28
XC30	3.09	6.9	2.35	3.32

NAME	Female01	Female234	Male01	Male234
XC31	30.26	60.43	4.38	13.34
XC32	37.68	15.33	16.06	18.83
XC33	6.14	16.17	13.51	7.12
XC34	-99	-99	22.31	37.81
XD35	54.8	-99	14.45	7.99
XD36	-99	-99	16.59	-99
XE37	-99	-99	7.97	13.71
XE38	39.64	23.62	2.88	31.27
XE39	-99	27.27	-99	-99
XF41	5.11	10.75	1.11	1.99
XF42	28.31	-99	14.74	51.49
XG45	10.98	32.22	13.34	10.64
XG46	3.73	6.08	1.76	4.3
XG47	6.32	8.49	4.55	16.46
XH49	9.76	8.19	3.87	8.86
XH50	-99	148.38	12.45	62.37
XH51	18.94	100	6.57	30.87
XH52	6.66	11.46	1.74	1.96
XI55	3.57	18.17	5.01	3.96
XI56	6.79	17.26	16.71	6.8
XJ58	7.78	35.95	3.09	3.75
XJ59	-99	-99	21.14	40.3
XJ60	26.82	47.86	16.19	14.43
XJ61	18.17	19.38	9.9	7.93
XJ62	-99	16.24	15.95	45.77
XJ63	38.71	35.85	5.56	22.5
XK64	3.31	3.86	1.36	13.33
XK65	6.36	12.27	3.03	8.33

NAME	Female01	Female234	Male01	Male234
XK66	32.36	11.58	9.64	18.86
XL68	3.6	10.45	2.75	5.22
XM70	14.72	17.9	4.67	12.84
XM71	10.76	7.43	3.53	10.59
XM72	16.15	5.16	9.15	18.91
XM73	35.95	24.92	24.55	12.22
XN74	5.83	7.01	0.88	6.99
XN75	16.68	13.56	3.2	9.53
XN76	-99	15.87	5.81	60.26
XO84	6.91	29.84	9.35	14.42

NAME	Female01	Female234	Male01	Male234
XP85	4.2	4.25	12.89	6.13
XQ86	3.55	22.03	4.6	4.55
XQ87	4.73	25.16	13.48	25.35
XR90	12.12	34.2	10.73	25.52
XR91	10.86	18.87	9.42	26.11
XS94	12.58	19.53	12.08	17.38
XS95	10.52	9.61	8.85	4.71
XS96	14.87	9.15	10.32	6.42
XT97	-99	-99	-99	100
XT98	<NA>	-99	<NA>	-99
XU99	-99	-99	4.01	92.68

: "XT98"(가구 내 고용활동) 사업장을 보면, "Female01", "Male01" 범주의 경우 값이 <NA>값을 보인다. 이는 2000년 전에 이 사업장에 고용된 사람이 없으므로 추적 인년 합계가 집계되지 않았고 이 부분이 연결되어 통합 누적 발생률이 <NA>로 처리된 것으로 추측된다.

---

1) **파란색으로 표시한 부분**은 UP1, UP2, 동일한 사업장을 SEX와 CAL2로 나누었을 때, 한 범주의 MAPE 값만 "NaN" 값을 보이는 경우를 의미한다. 이에 대한 정보는 다음과 같다.

NAME	사업장명	2015년에서 2018년 사이 통합 누적 발생률 중 0이 있는 범주
XA1	농업	여성 근로자 + 입사 시기가 2000년 이후
XA3	어업	여성 근로자 + 입사 시기가 2000년 이후
XB5	석탄, 원유 및 천연가스 광업	여성 근로자 + 입사 시기가 2000년 이후
XC12	담배 제조업	여성 근로자 + 입사 시기가 2000년 이후
XC19	코크스, 연탄 및 석유 정제품 제조업	남성 근로자 + 입사 시기가 2000년 이후
XD35	전기, 가스, 증기 및 공기 조절 공급업	여성 근로자 + 입사 시기가 2000년 이후
XF42	전문직별 공사업	여성 근로자 + 입사 시기가 2000년 이후
XH50	수상 운송업	여성 근로자 + 입사 시기가 2000년 이전
XJ62	컴퓨터 프로그래밍, 시스템 통합 및 관리업	여성 근로자 + 입사 시기가 2000년 이전
XN76	임대업; 부동산업 제외	여성 근로자 + 입사 시기가 2000년 이전

: MAPE 값이 "NaN"값을 보이는 범주 대부분이 여성 근로자이면서 입사 시기가 2000년 이후이다.

2) **보라색으로 표시한 부분**은 UP1, UP2, 동일한 사업장을 SEX와 CAL2로 나누었을 때, 같은 성별 내에서 CAL2간 MAPE의 차이가 낮은 사업장을 의미한다. (MAPE 값이 모두 "NaN"인 사업장은 제외) 이에 대한 정보는 다음과 같다.

① 성별이 남성인 경우

NAME	사업장명	CAL2 간 MAPE 차이의 절댓값
XC29	기타 기계 및 장비 제조업	0.32
XH52	창고 및 운송 관련 서비스업	0.22
XQ86	보건업	0.05

② 성별이 여성인 경우

NAME	사업장명	CAL2 간 MAPE 차이의 절댓값
XS95	개인 및 소비용품 수리업	0.91
XK64	금융업	0.55
XP85	교육 서비스업	0.05

3) **빨간색 밑줄로 표시한 부분**은 UP1, UP2, 동일한 사업장을 SEX와 CAL2로 나누었을 때, 같은 성별 내에서 CAL2간 MAPE의 차이가 큰 사업장을 의미한다.

① 성별이 남성인 경우

NAME	사업장명	CAL2 간 MAPE 차이의 절댓값
XC19	코크스, 연탄 및 석유 정제품 제조업	92.72
XU99	국제 및 외국기관	88.67
XD36	수도업	82.41

② 성별이 여성인 경우

NAME	사업장명	CAL2 간 MAPE 차이의 절댓값
XA3	어업	93.2
XC12	담배 제조업	86.94
XA1	농업	85.49

----------그래프 참고사항 ----------

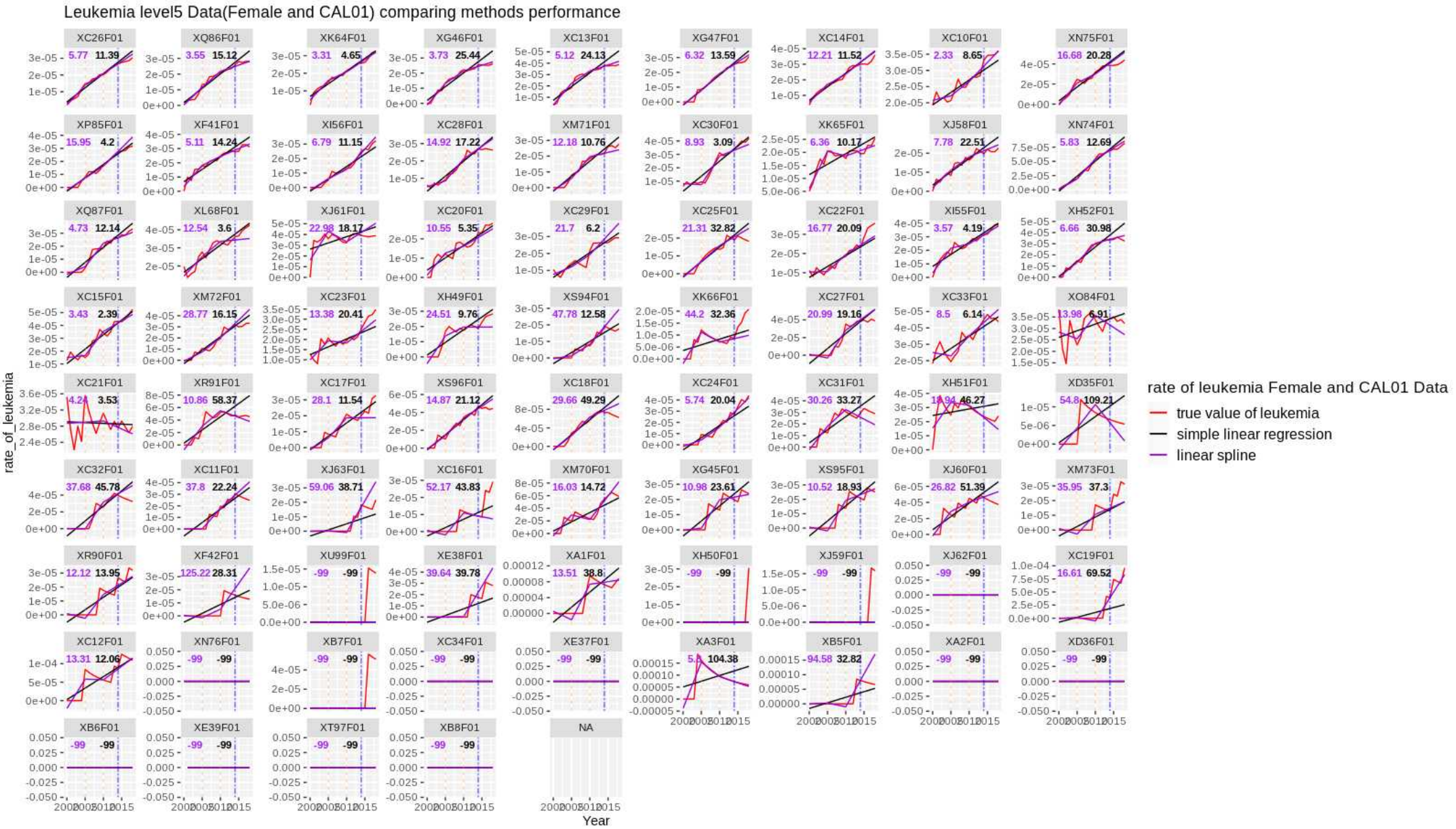
: **빨간색 실선**이 **실제 통합 누적 발생률 값**이고, **검은색 점선**은 적합한 **단순선형회귀모형**으로 **예측한 값**이며, **보라색 실선**은 **1차 spline function 적합 통해 예측한 값**을 의미한다.

: **파란색 수직**은 **YEAR=2014**를 의미하는 선이며, **주황색 수직선**은 **YEAR=2005, 2010년** 1차 spline function의 "knots"를 의미한다.

: 그래프의 순서는 2018년 기준 UP1 기준으로 grouping 한 후 계산한 추적 인년 합계의 순위이다. (높은 값에서 낮은 값 순으로 정렬)

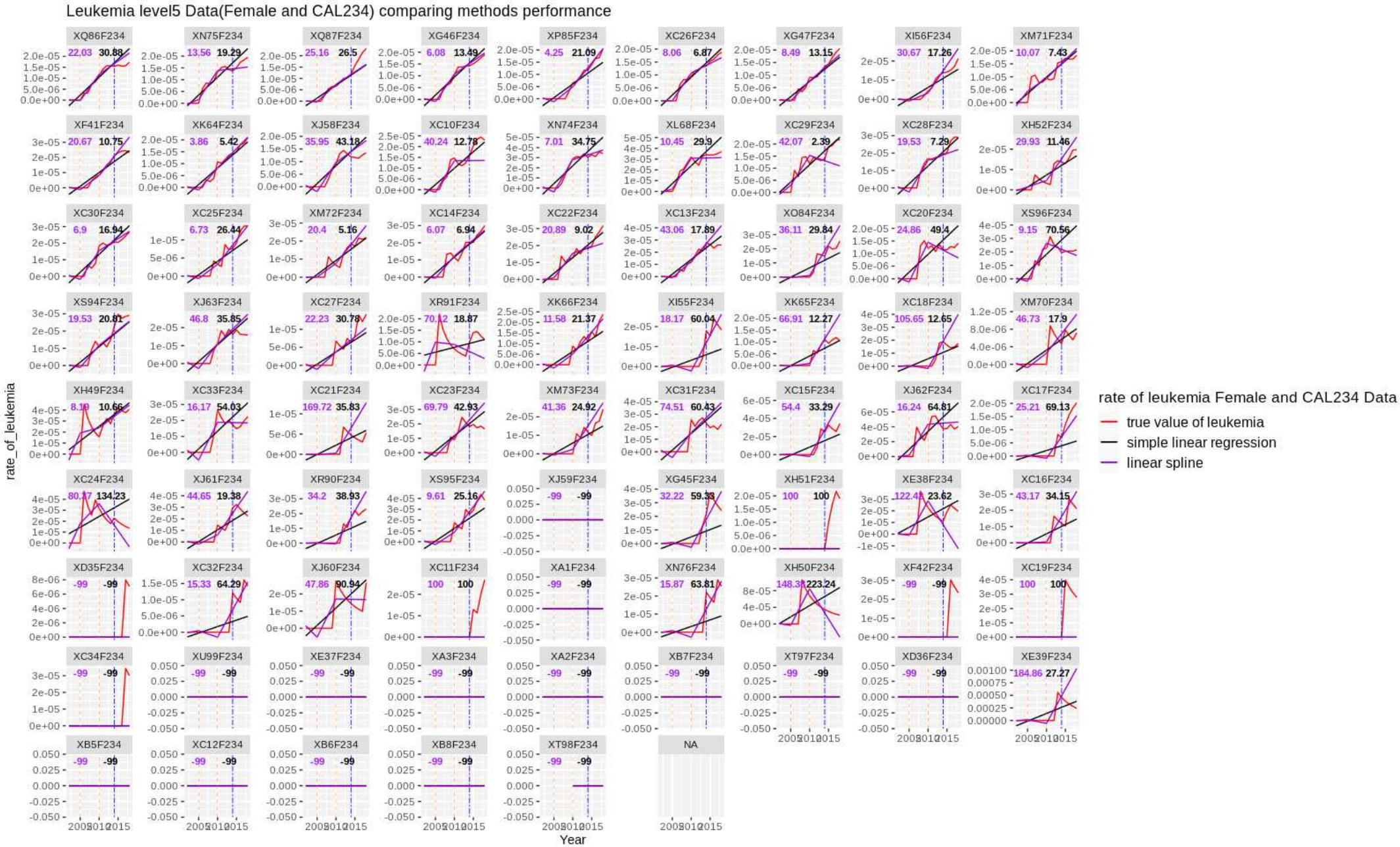
: 그래프 안, **보라색 글씨**는 **1차 spline function 적합 통해 얻은 MAPE 값**, **검은색 글씨**는 **단순선형회귀모형 적합 통해 얻은 MAPE 값**을 의미한다.

---- Level 5 Female + CAL2 = “01” Leukemia data ---- (그래프 순서는 2018년 기준 추적 인년 합계 순위)



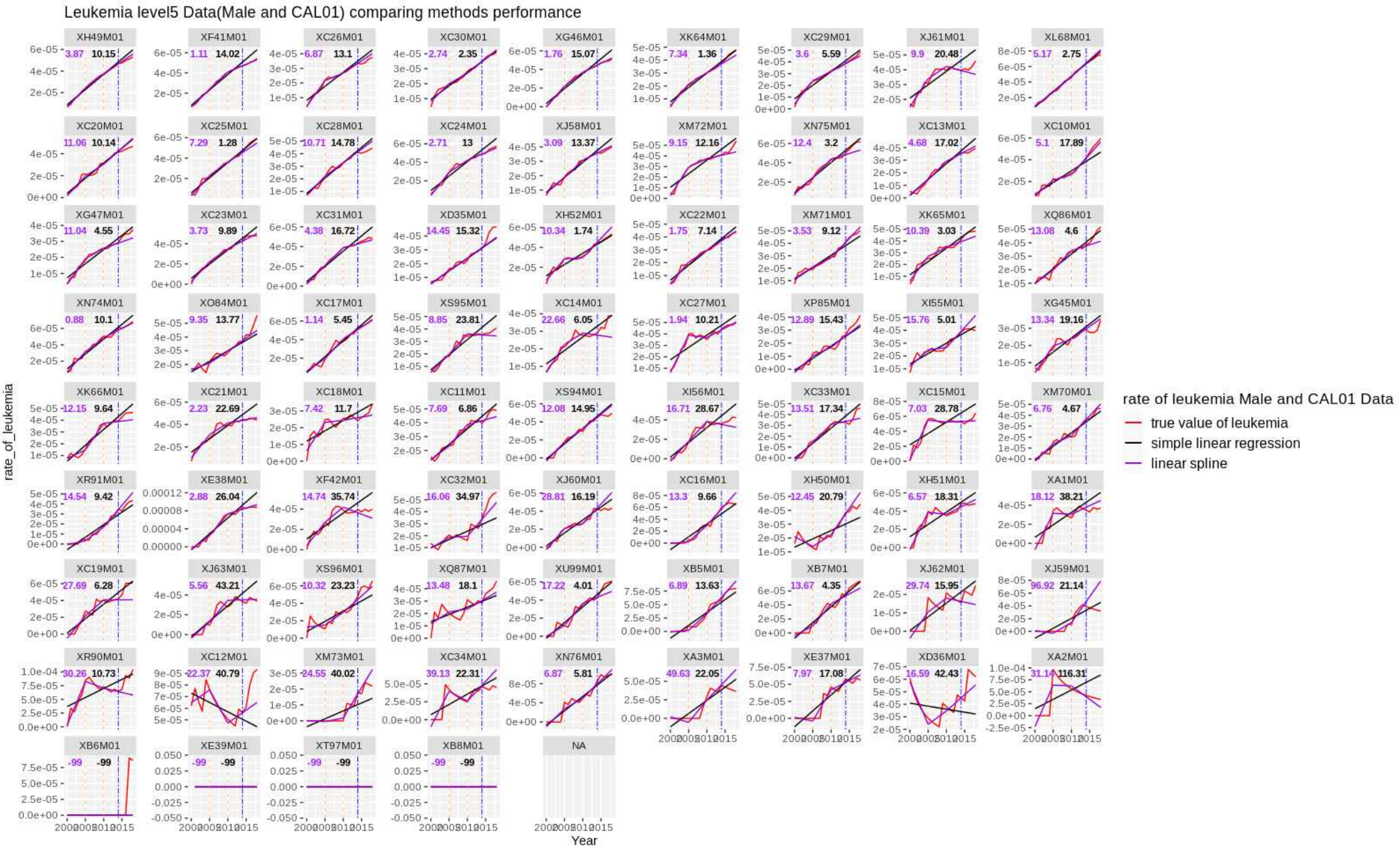
: 여성이면서 입사 시기가 2000년 이전인 근로자들의 경우, 사업장의 규모가 적을수록 질병 통합 누적 발생률이 “0”이 포함된 경우가 많다.

---- Level 5 Female + CAL2 = “234” Leukemia data ---- (그래프 순서는 2018년 기준 추적 인년 합계 순위)



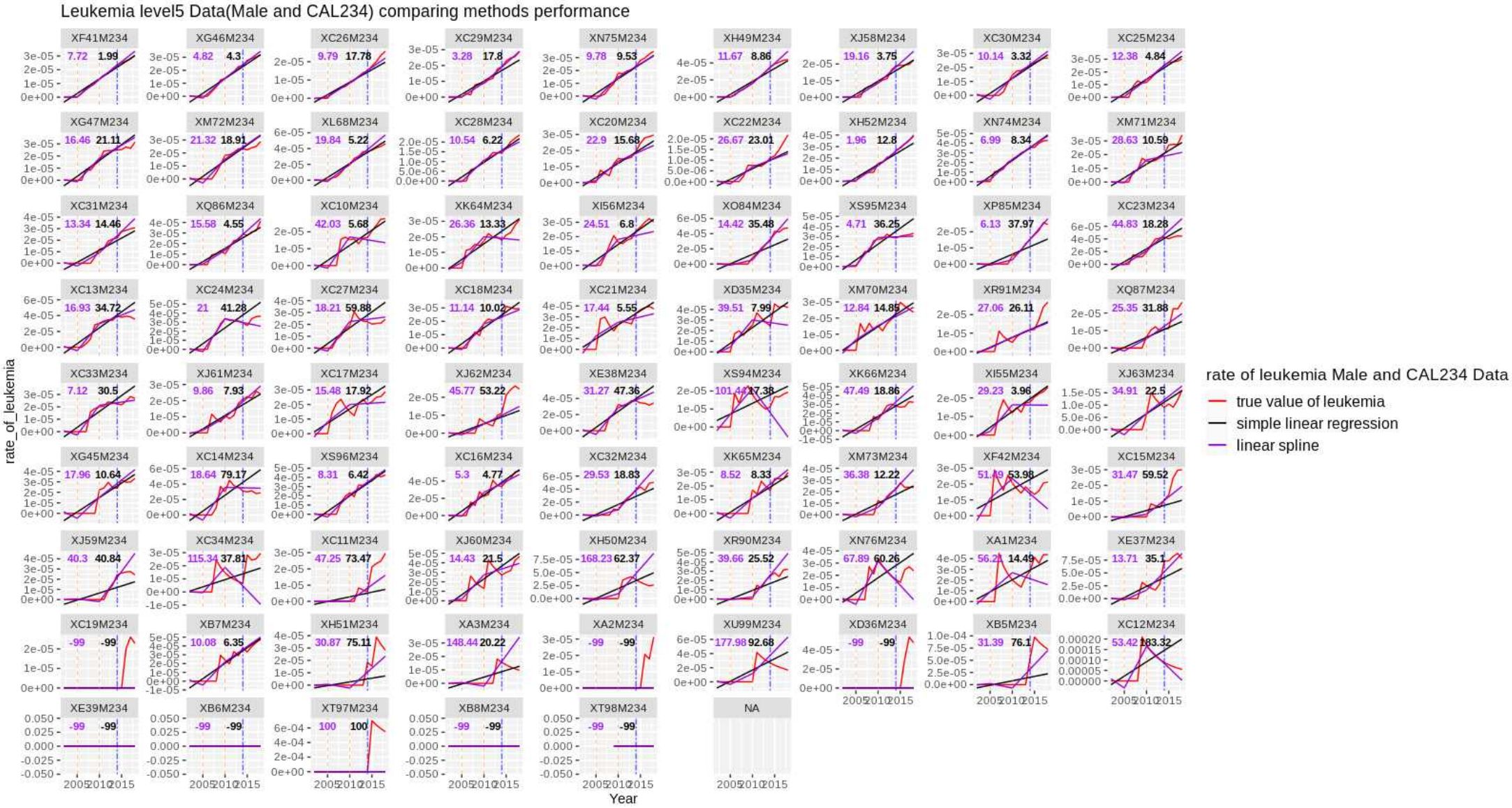
: 여성이면서 입사 시기가 2000년 이후인 근로자들의 경우, 사업장의 규모가 커도 통합 누적 발생률 변화 추세가 불안정한 경우가 대부분이다. / MAPE가 “NaN”인 사업장들이 좀 많다.

---- Level 5 Male + CAL2 = “01” Leukemia data ---- (그래프 순서는 2018년 기준 추적 인년 합계 순위)



: 남성이면서 입사 시기가 2000년 이전인 근로자 집단의 경우, 입사 시기는 동일하지만 여성인 근로자 집단에 비해 통합 누적 발생률 추세 변화는 안정적이지만, 여전히 사업장 규모가 적을수록 불안정한 추세 변화를 보인다.

---- Level 5 Male + CAL2 = “234” Leukemia data ---- (그래프 순서는 2018년 기준 추적 인년 합계 순위)



: 남성이면서 입사 시기가 2000년 이후인 근로자 집단의 경우, 입사 시기는 동일하지만 여성인 근로자 집단과 비슷하게 통합 누적 발생률의 추세가 불안정한 경우가 대부분이다. 이를 통해 추측하자면, 입사 시기가 2000년 이후인 근로자들이 속한 집단의 통합 누적 발생률 추세는 성별에 상관없이 많이 불안정하다고 말할 수 있다.

② Lung cancer Data

: 각 "NAME" 별 단순선형회귀모형, 1차 spline function을 적용한 후에 MAPE를 계산한 뒤, 두 모형 중 더 작은 MAPE를 가지는 값을 따로 저장하였다. 결과는 아래 표와 같다. (더 작은 MAPE 값을 주는 모형 이름은 제시하지 않고, 뒤 그래프로 제시)

: 표에 제시되는 "Female01"은 여성이면서 2000년 전에 입사한 근로자, "Female234"는 여성이면서 2000년 이후에 입사한 근로자, "Male01"은 남성이면서 2000년 전에 입사한 근로자, "Male234"는 남성이면서 2000년 이후에 입사한 근로자를 지칭한다.

NAME	Female01	Female234	Male01	Male234
XA1	22.42	20.39	3.92	15.01
XA2	-99	-99	56.66	18.49
XA3	17.92	53.6	3.74	12.09
XB5	42.02	-99	2.52	21.46
XB6	-99	-99	7.31	28.6
XB7	37.51	95.39	6.53	22.74
XB8	-99	-99	-99	-99
XC10	4.07	11.3	4.2	4.55
XC11	8.17	55.79	4.47	10.74
XC12	5.45	-99	3.26	-99
XC13	4.41	4.79	1.43	5.55
XC14	5.38	3.3	1.6	8.29
XC15	3.52	19.94	1.55	9.15
XC16	23.58	31.19	2.29	2.53
XC17	6.64	4.83	2.5	5.01
XC18	5.3	21.06	6.14	6.52
XC19	55.56	100	8.85	22.74
XC20	2.49	5.34	2.9	5.86
XC21	10.66	17.38	7.04	20.51
XC22	1.3	4.24	1.18	6.35
XC23	5.26	4.54	3.84	2.59
XC24	3.01	26.32	1.08	7.33
XC25	10.9	12.07	2.55	2.9
XC26	6.17	16.7	2.15	12.09
XC27	9	25.99	5.51	8.05
XC28	4.44	11.7	3.3	7.98
XC29	1.53	21.89	0.88	7.94
XC30	1.83	5	8.07	4.38
XC31	19.06	14.45	4.53	15.25
XC32	5.78	13.12	1.34	5.22
XC33	4.78	4.46	2.43	16.4
XC34	-99	-99	22.43	12.4
XD35	9.34	58.5	3.35	5.25

NAME	Female01	Female234	Male01	Male234
XD36	100	-99	23.54	62.09
XE37	63.44	43.61	9.05	7.9
XE38	3.3	22.4	1.24	3.73
XE39	-99	-99	25.29	-99
XF41	6.19	3.55	2.35	3.91
XF42	43.42	103.5	1.98	1.99
XG45	12.13	-99	17.98	24.92
XG46	1.34	4.25	4.57	3.7
XG47	1.46	3.23	5.05	1.64
XH49	4.39	4.51	2.47	5.14
XH50	10.14	-99	6.46	7.46
XH51	24.86	-99	2.49	89.57
XH52	12.21	13.58	0.88	7.09
XI55	2.69	25.12	6.5	21.56
XI56	1.42	2.76	8.27	1.05
XJ58	6.46	10.14	0.31	4.87
XJ59	52.47	10.59	10.85	15.24
XJ60	27.86	21.85	4.43	43.75
XJ61	2.34	22.14	1.45	23.51
XJ62	64.69	18.35	6.27	2.61
XJ63	6.64	9.25	12.72	8.61
XK64	7.32	6.33	0.37	7.39
XK65	5.71	23.9	8.45	6.2
XK66	17.93	13.94	2.5	9.54
XL68	3.73	19.54	0.5	2.68
XM70	7.88	59.19	1.56	6.27
XM71	8.73	10.81	1.05	3.58
XM72	2	30.27	2.33	3.97
XM73	38.87	32.42	3.41	9.62
XN74	1.44	5.41	1.43	2.19
XN75	1.96	6.24	2.22	1.18
XN76	-99	31.39	3.88	16.41
XO84	4.66	16.93	2.61	5.96

NAME	Female01	Female234	Male01	Male234
XP85	14.57	6.23	3.15	6.66
XQ86	1.06	6.66	8.18	19.87
XQ87	14.28	21.89	2.06	11.65
XR90	18.46	22.45	10.74	29.11
XR91	8.87	4.06	12.75	3.64
XS94	2.35	20.21	1.48	2.27
XS95	3.13	34.11	4.37	11.18
XS96	2.52	3.84	4.57	8.83
XT97	-99	-99	-99	-99
XT98	<NA>	-99	<NA>	-99
XU99	35.74	51.68	3.36	12.11

---

1) **파란색으로 표시한 부분**은 UP1, UP2, 동일한 사업장을 SEX와 CAL2로 나누었을 때, 한 범주의 MAPE 값만 "NaN" 값을 보이는 경우를 의미한다. 이에 대한 정보는 다음과 같다.

NAME	사업장명	2015년에서 2018년 사이 통합 누적 발생률 중 0이 있는 범주
XB5	석탄, 원유 및 천연가스 광업	여성 근로자 + 입사 시기가 2000년 이후
XD36	수도업	여성 근로자 + 입사 시기가 2000년 이후
XG45	자동차 및 부품 판매업	여성 근로자 + 입사 시기가 2000년 이후
XH50	수상 운송업	여성 근로자 + 입사 시기가 2000년 이후
XH51	항공 운송업	여성 근로자 + 입사 시기가 2000년 이후
XN76	임대업; 부동산업 제외	여성 근로자 + 입사 시기가 2000년 이전

: MAPE 값이 "NaN"값을 보이는 범주 대부분이 백혈병 데이터와 마찬가지로 여성 근로자이면서 입사 시기가 2000년 이후이다.

2) **보라색으로 표시한 부분**은 UP1, UP2, 동일한 사업장을 SEX와 CAL2로 나누었을 때, 같은 성별 내에서 CAL2간 MAPE의 차이가 낮은 사업장을 의미한다. (MAPE 값이 모두 "NaN"인 사업장은 제외) 이에 대한 정보는 다음과 같다.

① 성별이 남성인 경우

NAME	사업장명	CAL2 간 MAPE 차이의 절댓값
XC10	식료품 제조업	0.35
XC16	목재 및 나무제품 제조업; 가구제외	0.24
XF42	전문직별 공사업	0.01

② 성별이 여성인 경우

NAME	사업장명	CAL2 간 MAPE 차이의 절댓값
XC13	석유제품 제조업; 의복 제외	0.38
XC33	기타제품 제조업	0.32
XH49	육상 운송 및 파이프라인 운송업	0.12

: 같은 성별 내에서 CAL2간 MAPE의 차이가 낮은 사업장들의 대분류를 보면 대부분 "C"(제조업)인 부분이 눈에 띈다.

3) **빨간색 밑줄로 표시한 부분**은 UP1, UP2, 동일한 사업장을 SEX와 CAL2로 나누었을 때, 같은 성별 내에서 CAL2간 MAPE의 차이가 큰 사업장을 의미한다.
① 성별이 남성인 경우

NAME	사업장명	CAL2 간 MAPE 차이의 절댓값
XC12	담배 제조업	95.74
XH51	항공 운송업	87.08
XE39	환경 정화 및 복원업	73.71

② 성별이 여성인 경우

NAME	사업장명	CAL2 간 MAPE 차이의 절댓값
XC12	**담배 제조업**	93.55
XH50	수상 운송업	88.86
XG45	자동차 및 부품 판매업	86.87

: **초록색으로 강조한 사업장**은 이전 백혈병 데이터에서도 여성 근로자 중에서 CAL2간 MAPE의 차이가 큰 사업장을 의미한다.
: 성별에 상관없이 CAL2간 MAPE의 차이가 큰 사업장은 사업장 "C12"(담배 제조업)인 것을 알 수 있다.

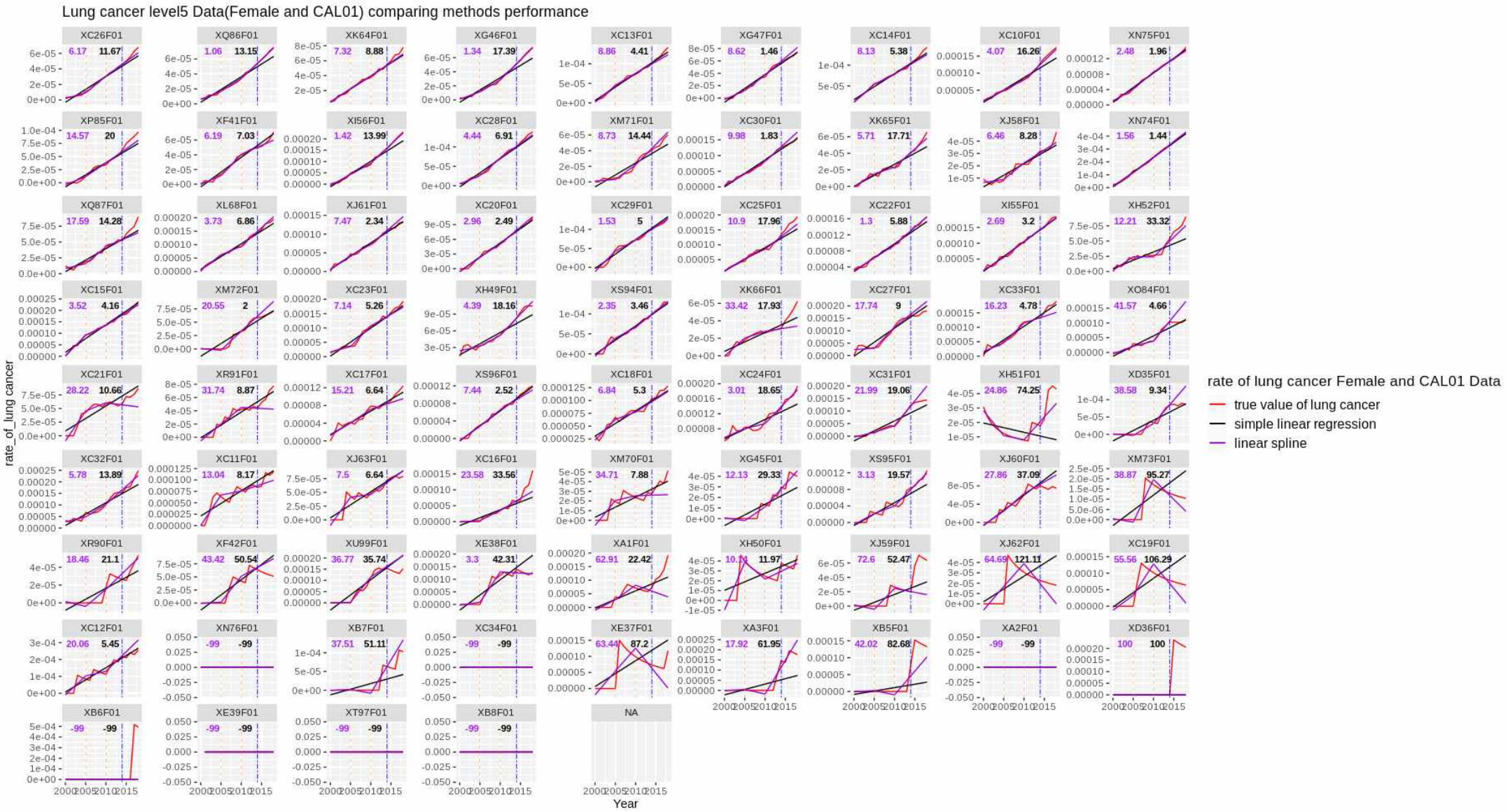
참고사항

: Leukemia data와 Lung cancer data 모두 UP1, UP2가 동일한 사업장을 SEX와 CAL2로 나누었을 때, 여성이면서 입사 시기가 2000년대 이후인 근로자 집단의 추세 변화가 좀 더 불안정함을 보인다.

----------그래프 참고사항 ----------

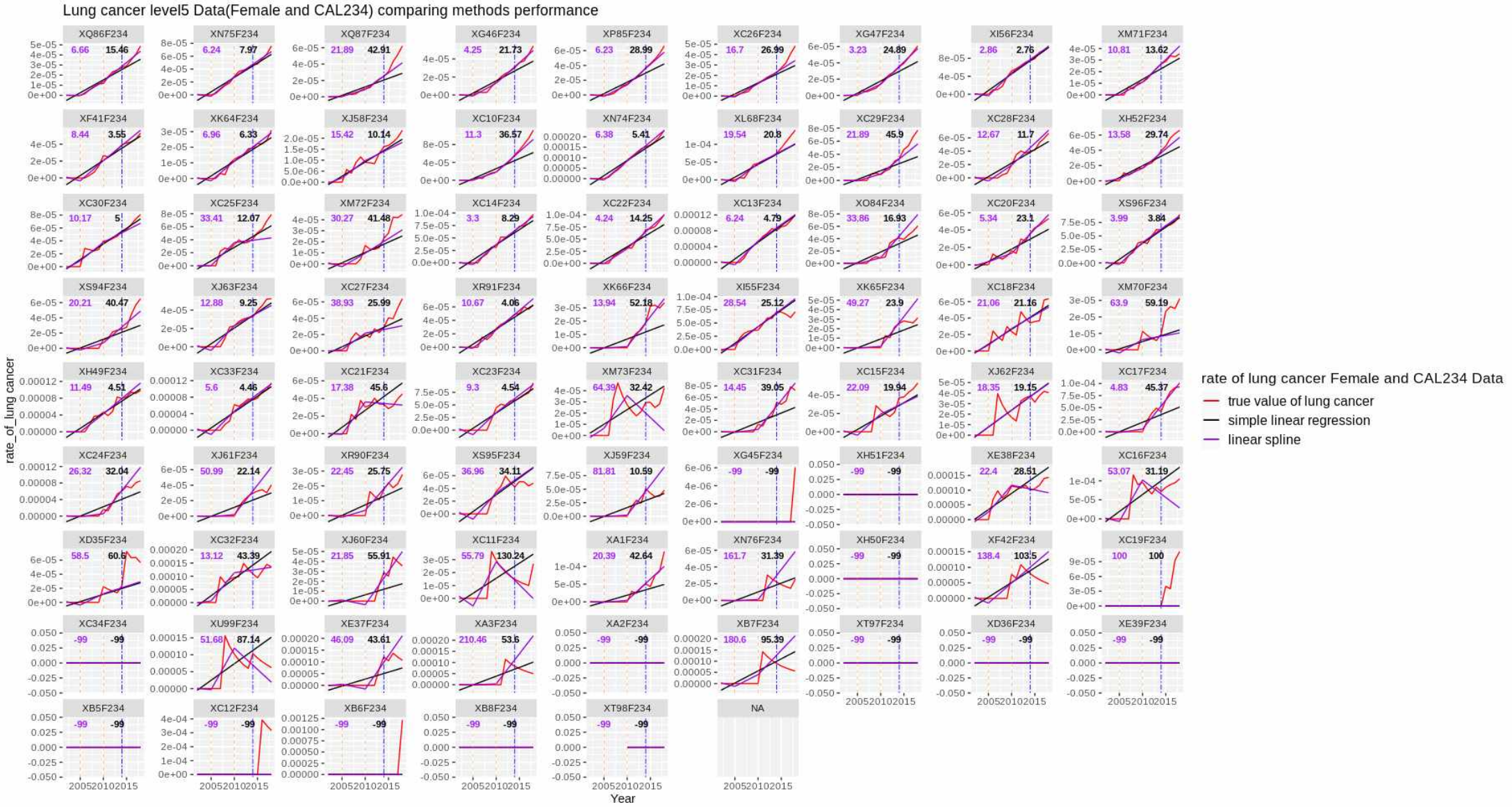
: **빨간색 실선**이 **실제 통합 누적 발생률 값**이고, **검은색 점선**은 적합한 **단순선형회귀모형**으로 **예측한 값**이며, **보라색 실선**은 **1차 spline function 적합 통해 예측한 값**을 의미한다.
: **파란색 수직**은 YEAR=2014를 의미하는 선이며, **주황색 수직선**은 YEAR=2005, 2010년 1차 spline function의 "knots"를 의미한다.
: 그래프의 순서는 2018년 기준 UP1 기준으로 grouping 한 후 계산한 추적 인년 합계의 순위이다. (높은 값에서 낮은 값 순으로 정렬)
: 그래프 안, **보라색 글씨**는 **1차 spline function 적합 통해 얻은 MAPE 값**, **검은색 글씨**는 **단순선형회귀모형 적합 통해 얻은 MAPE 값**을 의미한다

---- Level 5 Female + CAL2 = “01” Lung cancer data ---- (그래프 순서는 2018년 기준 추적 인년 합계 순위)



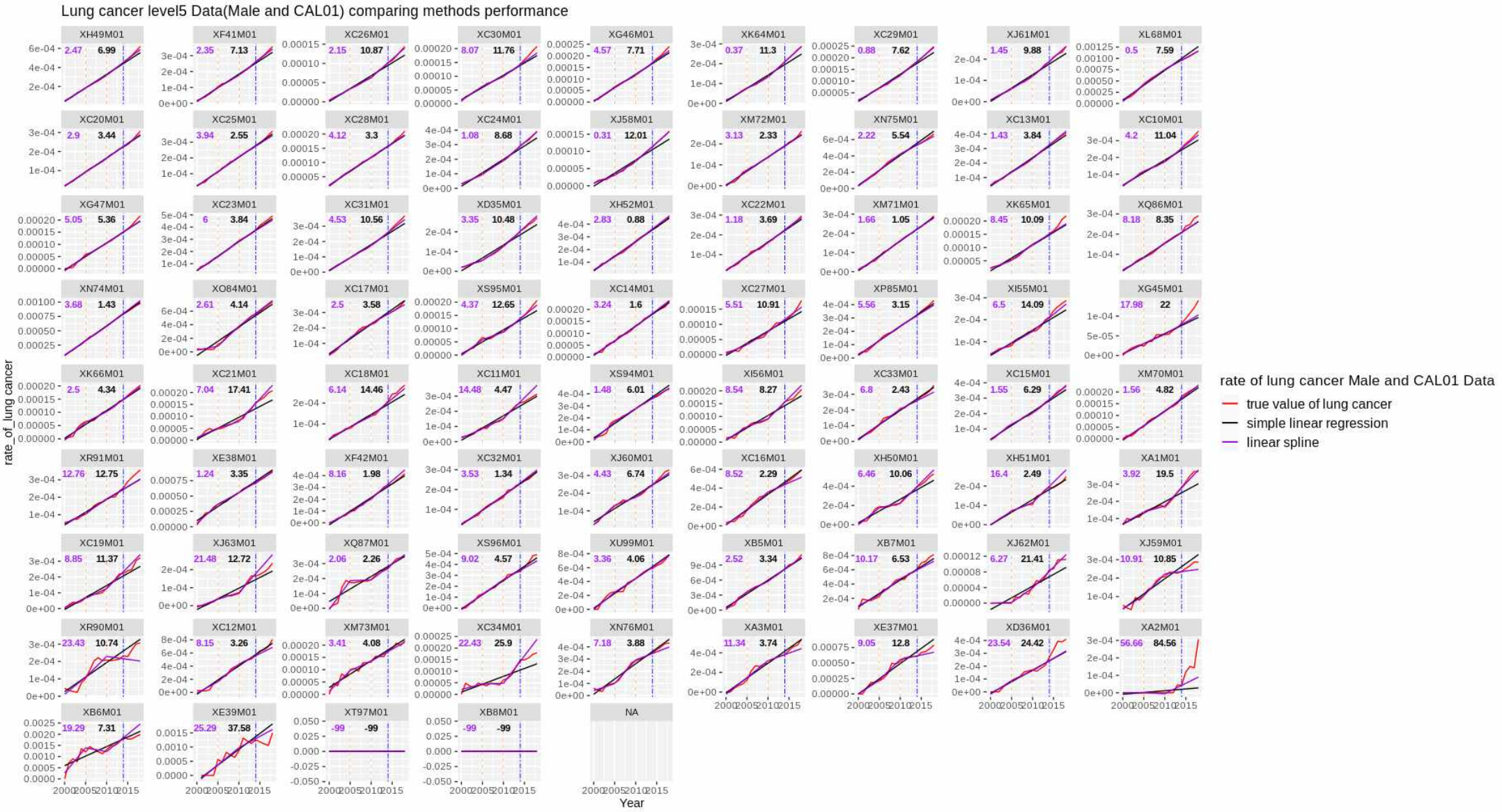
: 그래프를 보면, 같은 성별과 같은 입사 시기를 가지는 근로자 집단이지만 백혈병의 통합 누적 발생률 추세 변화보다는 좀 더 안정적인 추세를 보이는 것을 알 수 있고, 사업장의 규모가 적을수록 추세의 불안정함이 커진다는 점을 볼 수 있다.

---- Level 5 Female + CAL2 = “234” Lung cancer data ---- (그래프 순서는 2018년 기준 추적 인년 합계 순위)



: 같은 성별임에도 불구하고 입사 시기가 2000년 이후인 여성 근로자 집단의 통합 누적 발생률 추세가 더 불안정함을 알 수 있다. 또한, 백혈병 데이터와 마찬가지로, MAPE가 “NaN”인 사업장이 좀 더 많음을 볼 수 있다.

---- Level 5 Male + CAL2 = “01” Lung cancer data ---- (그래프 순서는 2018년 기준 추적 인년 합계 순위)



: 같은 성별, 입사 시기를 가지는 근로자 집단임에도 불구하고 백혈병의 통합 누적 발생률 추세 변화보다 훨씬 더 안정적임을 알 수 있다.

: 사업장의 규모가 적어도 추세의 불안정함이 많이 크지 않다는 점도 눈에 띈다.

: 같은 입사 시기이더라도 여성 근로자 집단보다 추세 변화가 훨씬 더 안정적이다.

---- Level 5 Male + CAL2 = “01” Lung cancer data ---- (그래프 순서는 2018년 기준 추적 인년 합계 순위)

: 같은 성별이더라도 입사 시기가 2000년 이전인 근로자 집단보다는 통합 누적 발생률 추세가 좀 더 불안정함을 알 수 있다.

: 그러나, 여성이면서 입사 시기가 2000년 이후인 근로자 집단보다는 훨씬 더 안정적인 추세를 보이고, MAPE가 “NaN”값인 사업장이 드물다.