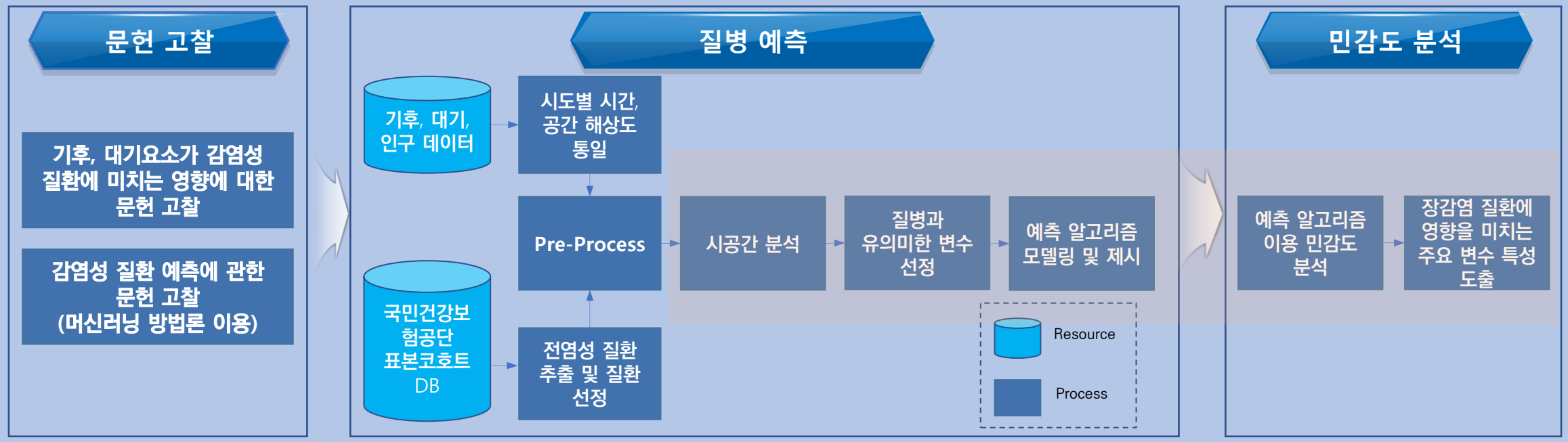
기후변화에 따른 감염성 질환 예측

환경정책평가연구원
강선아

기후변화에 따른 감염성 질환 예측

- 연구내용
  - 2009년~2013년(5개년)동안 발생하는 장감염 질환의 시공간 분석 및 예측 알고리즘 구축, 민감도 분석을 통해 장감염 질환에 영향을 미치는 주요 변수 파악
- 연구대상
  - 2009년부터 2013년까지 연속적으로 발생한 장감염 질환(국민건강보험공단 표본 코호트 DB 기준)



<그림 2-1> 연구 프로세스

데이터 전처리: 질병 선정 및 건수 도출

설명변수 데이터 전처리

- Step 1. 공간 해상도
  - 측정소 데이터는 공간적으로 점(point)데이터이고, 시군구/시도의 경우 면(polygon)데이터임
  - 공간해상도를 맞추기 위하여 같은 시군구/시도에 위치한 측정소의 데이터를 평균 내어 매칭
- Step 2. 시간 해상도
  - 시간해상도는 년, 월, 일 모두 다르며, 분석을 수행할 시간해상도는 월 단위
  - 월 단위보다 시간해상도가 낮은 경우: 시간, 일 단위 데이터의 평균을 월 단위 데이터로 사용
  - 월 단위보다 시간해상도가 높은 경우: 연 데이터를 12로 나누어 사용하고, 농도, 밀도(인구)와 같은 경우 연 데이터를 그대로 사용

질병 건수 산정 및 질병 선정

- Step 1. 자격 DB와 진료 DB 연계: 무진료 기간을 0으로 처리하여 질병 건수 산정
- Step 2. 진료 DB 주상병명과 부상병명이 다른 경우: 다른 케이스로 간주
- Step 3. 2009~2013년 연속 발생 질병만 분석 대상으로 고려
- Step 4. 질병코드(한국질병표준사인분류 기준) 소수점 그룹화 : 질병 레벨을 높여 건수를 산출

질환의 발생건수 및 월별 시계열 분석을 통해 장감염 질환을 분석대상 질병으로 선정

장감염 질환 위험지역 분석

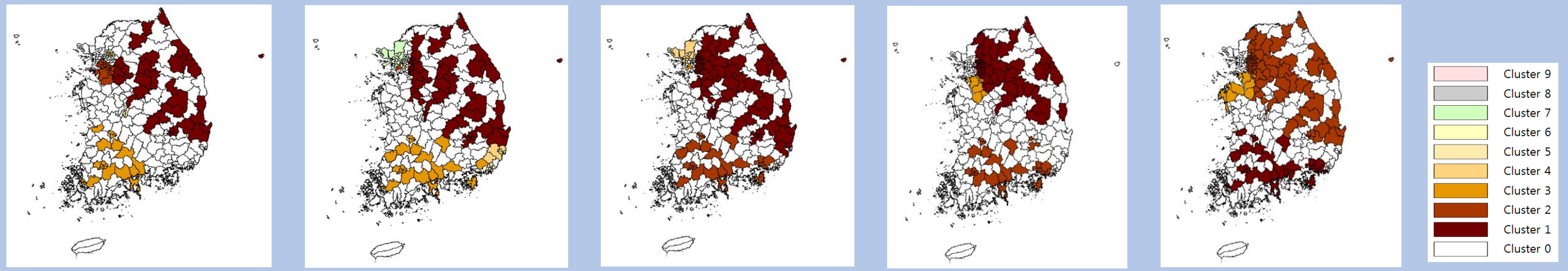
- 장감염 질환 위험지역 분석(2009년~2013년)

- SaTScan 프로그램을 이용하여 장감염 질환의 발생위험 지역을 분석
- 위험 지역 분석 결과, 2009년부터 2012년까지 고위험 지역은 강원도, 경상도 일부이며 2013년에는 경상도, 전라도 일부임
- 고위험 지역에 속하는 시군구는 2009년 35개, 2010년 57개, 2011년 87개, 2012년 62개, 2013년 35개로 나타남(<그림 2-2 참조>)

<표 2-1> SaTSCan 분석 결과

연도	Cluster	No. of case	Relative risk	Log likelihood ratio	P-value
2009	cluster 1(35)	11,276	0.75	427.31	<0.000
2010	cluster 1(57)	30,890	0.85	294.80	<0.000
2011	cluster 1(87)	57,205	0.86	343.94	<0.000
2012	cluster 1(62)	45,835	0.82	640.30	<0.000
2013	cluster 1(35)	29,574	1.28	692.26	<0.000

자료: 저자 작성.



2009년 2010년 2011년 2012년 2013년

<그림 2-2> 연도별 장감염 발생 위험 지역 클러스터링 결과

자료: 저자 작성.

장감염 질환 예측

- 장감염 질환 예측 알고리즘 구축

- 기상인자, 대기인자, 인구통계적 데이터 및 지리적 특성을 나타내는 위, 경도 데이터를 이용하여 장감염 질환 예측 알고리즘 (OLS 회귀분석, LASSO 회귀분석, 심층신경망)을 구축

- 알고리즘 학습을 위해 데이터를 학습 데이터와 테스트 데이터로 구분하였으며, 이 때 층화추출법을 사용함 (연도별, 월별, 시군구별로 데이터를 구분한 후 학습 데이터 70%, 테스트 데이터 30% 추출)

- 데이터의 범위를 통일시키기 위해 min-max 표준화를 수행

<표 2-2> 시군구별 장감염 질환 발생 예측 알고리즘에 사용된 설명변수와 목표변수

구분	변수명
설명변수(41개)	월(month), 위도, 경도, SO2_mean, CO_mean, O3_mean, NO2_mean, PM10_mean, SO2_max, CO_max, O3_max, NO2_max, PM10_max, 평균기온, 평균최고기온, 평균최저기온, 최고기온, 최저기온, 평균현지기압, 평균해면기압, 최고해면기압, 최저해면기압, 평균수증기압, 최고수증기압, 최저수증기압, 평균이슬점온도, 평균상대습도, 최소상대습도, 월합강수량, 일최다강수량, 평균풍속, 최대풍속, 최대순간풍속, 일조시간합, 일조율, 월적설량합, 평균.최저초상온도, 최저초상온도, 평균지면온도, 총인구수, 인구밀도
목표변수	시군구별 월별 장감염 질병 발생 건수

자료: 저자 작성.

장감염 질환 예측

- 장감염 질환 예측 알고리즘 구축(OLS 회귀분석)

- R package "lm"으로 분석 수행
- 단계적 선택법을 이용하여 AIC가 작은 변수의 조합을 추출, 장감염 질환 예측에 사용된 변수는 총 31개임
- 장감염 질환 발생 =

$$\beta_0 + \beta_1 x_{1ijt} + \beta_2 x_{2ijt} + \cdots + \beta_{31} x_{31ijt} + \epsilon$$

- 모형의 R^2값은 0.7856으로 해당 알고리즘은 높은 설명력을 가짐
- 장감염 질환 예측결과 RMSE가 20.040으로 나타났으며, 질병을 세분화하여 모델링을 구축한 결과는 기타 세균성 장감염과 바이러스성 및 기타 감염성 장감염 각각의 RMSE가 4.342, 5.440으로 나타남

<표 2-3> OLS 회귀분석 결과(장감염 전체)

모델							
변수	Coef.	s.e	P>t	변수	Coef.	s.e	P>t
month2	-5.429***	-3.728	4.53E-14	평균기온	-230.900***	-3.992	6.62E-05
month3	-4.723*	-2.296	0.000195	평균최고기온	96.200***	3.435	0.000596
month4	-11.560***	-3.787	0.021679	평균최저기온	121.300**	3.208	0.001343
month5	-13.580**	-3.138	0.000154	최저기온	-18.430	-1.503	0.13287
month6	-13.780*	-2.44	0.001709	평균해면기압	-16.210*	-2.432	0.015024
month7	-8.164	-1.186	0.014695	최고해면기압	17.460***	3.566	0.000365
month8	0.008	0.001	0.235735	최저해면기압	15.240***	3.895	9.90E-05
month9	-6.712	-1.341	0.999016	평균수증기압	31.190*	2.576	0.010023
month10	-20.950***	-5.916	0.179914	최고수증기압	-28.660***	-5.288	1.27E-07
month11	-8.935***	-3.973	3.46E-09	최저수증기압	13.150*	2.284	0.02239
month12	7.711***	5.662	7.18E-05	평균 이슬점온도	33.700**	2.869	0.004128
경도	-2.497***	-7.45	1.56E-08	월합강수량	7.902*	2.448	0.014391
위도	-2.523***	-7.091	1.04E-13	평균풍속	-9.940*	-2.562	0.010429
CO_mean	7.208*	1.978	1.46E-12	최대풍속	20.830***	5.181	2.27E-07
O3_mean	25.030***	6.409	0.047966	일조시간합	28.320***	11.038	<2e-16
NO2_mean	9.968**	3.197	1.56E-10	평균.최저 초상온도	52.920**	3.05	0.002293
PM10_ mean	-14.210***	-4.386	0.001394	최저초상온도	-21.590	-1.764	0.077806
SO2_max	50.540***	10.227	1.17E-05	평균지면온도	-22.960*	-2.31	0.020905
CO_max	-27.910***	-6.664	<2e-16	총인구수	180.900***	131.59	<2e-16
O3_max	-7.837*	-2.403	2.86E-11	인구밀도	-19.570***	-12.481	<2e-16
PM10_max	-13.360**	-3.091	0.016268				

주: Coef: Coefficient,s.e: standard error, Significance level: * p<.05, ** p<.01, *** p<.001

자료: 저자작성.

장감염 질환 예측

- 장감염 질환 예측 알고리즘 구축(LASSO 회귀분석)

- R package "lars"로 분석 수행
- LASSO 회귀분석 결과 에러가 최소가 되도록 하는 lamda 값은 0.0021임
- 모형의 회귀계수를 보면 평균이슬점온도는 -0.075로 모형에 거의 영향을 미치지 않는 변수로 나타남
- 대기인자 중 O3평균, SO2 최대값, CO 최대값의 회귀계수가 각각 22.891, 49.804, -32.465로 모형 구축에 중요한 변수로 나타남
- 기상인자 중 평균 수증기압, 평균상대습도의 회귀계수는 각각 29.306, 12.423으로 장감염 질환과 양의 상관관계가 있었으며, 평균해면기압, 최고수증기압의 회귀계수는 각각 -11.933, -28.642로 음의 상관관계가 있었음
- 장감염 질환 예측결과 RMSE가 20.033으로 나타났으며, 질병을 세분화하여 모델링을 구축한 결과는 기타 세균성 장감염과 바이러스성 및 기타 감염성 장감염 각각의 RMSE가 4.342, 5.431로 나타남

<표 2-4> LASSO회귀분석 결과(장감염 전체)

lasso regression model					
변수	month2	month3	month4	month5	month6
회귀계수	-4.574	-1.098	-6.669	-6.478	-5.072
변수	month7	month8	month9	month10	month11
회귀계수	-0.717	6.865	-0.339	-17.080	-7.091
변수	month12	경도	위도	SO2_mean	CO_mean
회귀계수	7.067	-2.123	-2.297	1.302	7.298
변수	O3_mean	NO2_mean	PM10_mean	SO2_max	CO_max
회귀계수	22.891	8.179	-12.011	49.804	-32.465
변수	O3_max	NO2_max	PM10_max	평균기온	평균최고기온
회귀계수	-7.067	4.018	-19.040	-3.544	1.561
변수	평균최저기온	최고기온	최저기온	평균현지기압	평균해면기압
회귀계수	2.011	-1.625	-8.937	-0.398	-11.933
변수	최고해면기압	최저해면기압	평균수증기압	최고수증기압	최저수증기압
회귀계수	16.040	15.205	29.306	-28.642	14.522
변수	평균이슬점온도	평균상대습도	최소상대습도	월합강수량	일최다강수량
회귀계수	-0.075	12.423	-2.742	9.443	-1.391
변수	평균풍속	최대풍속	최대순간풍속	일조시간합	일조율
회귀계수	-8.178	17.110	3.076	7.496	20.037
변수	월적설량합	평균. 최저초상온도	최저초상온도	평균지면온도	총인구수
회귀계수	8.687	51.530	-21.042	-20.576	182.701
변수	인구밀도				
회귀계수	-21.193				

자료: 저자 작성.

장감염 질환 예측

- 장감염 질환 예측 알고리즘 구축(심층 신경망)

- R package "h2o"를 이용하여 분석 수행
- 각각의 모델마다 최적의 파라미터를 찾기 위해 hidden layer의 개수 및 epoch를 변화시켜 모델링
- 장감염 질환 예측결과 RMSE가 15.656으로 나타났으며, 질병을 세분화하여 모델링을 구축한 결과는 기타 세균성 장감염과 바이러스성 및 기타 감염성 장감염 각각의 RMSE가 4.071, 5.042로 나타남

<표 2-5> 심층신경망 분석결과

model	parameters	RMSE
장감염 전체	Hidden layer :3 Hidden node: 500 Epoch: 30 Activation function: ReLU	15.656
기타 세균성 장감염	Hidden layer :3 Hidden node: 500 Epoch: 7 Activation function: ReLU	4.071
바이러스성 및 기타 감염성 장감염	Hidden layer :3 Hidden node: 500 Epoch: 70 Activation function: ReLU	5.042

자료: 저자 작성.

결과: 심층신경망 평균제곱근오차 10~25% 개선

- 장감염 질환 전체 예측: OLS/ LASSO 회귀분석보다 심층신경망의 성능이 대략 25% 향상
- 기타 세균성 장감염 및 바이러스성 및 기타 감염성 장감염 예측: OLS/ LASSO 회귀분석보다 심층신경망의 성능이 대략 10% 향상

장감염 질환 예측 모델 성능 비교

추정 대상	OLS	LASSO	심층신경망
장감염 질환 전체	20.040	20.033	15.656
기타 세균성 장감염	4.342	4.342	4.071
바이러스성 및 기타 감염성 장감염	5.440	5.431	5.042

민감도 분석: 주요 변수 양적 영향 파악

- 장감염 질환 발생에 영향을 미치는 주요 변수를 파악하기 위해 민감도 분석 수행
  - 민감도 분석은 다음과 같은 방법으로 수행됨

테스트 데이터 셋의 설명변수가 i개 있다면, i개 설명변수에 대해 각각 10%씩 값을 상승시킴

$$x_i = (1 + 0.1)x_i$$

구축된 모델링(심층 신경망)에 X_i를 적용하여 장감염 발생빈도 변화율을 계산

$$yhat{\cdot}0 = DNN(x)$$

$$yhat{\cdot}1 = DNN(1.1x)$$

$$dy/y = \frac{1}{N}\sum\frac{yhat{\cdot}1 - yhat{\cdot}0}{yhat{\cdot}0}$$

민감도 분석 결과: 기후변수 양적 영향 파악

- 지역, 인구의 영향이 크게 나타남
  - 위도, 경도 값을 변화시켰을 때 장감염 예측 건수가 가장 크게 변동하는 것으로 나타남
  - 총 인구 수가 증가하는 비율만큼 장감염 발생 증가
- 전체 장감염: [촉진요인] 현지기압, 일조율, 평균최저초상온도 [억제요인] 평균지면온도, 최저기온
- 세균성 장감염: [촉진요인] 평균현지기압, 평균최저초상온도, 최저초상온도, 오존, 일산화탄소
- 바이러스성 및 기타 감염성 장감염:[촉진요인] 일조율, 일조시간 [억제요인] 평균현지기압, 최고/최저/평균 해면기압, 평균/최소 상대습도

민감도 분석 결과

변인	장감염 전체	변인	기타세균성장감염	변인	바이러스 및 기타 감염성 장감염
평균현지기압	5.1	평균현지기압	2.78	일조율	13.59
일조율	3.05	평균최저초상온도	1.67	일조시간합	8.99
평균최저초상온도	1.54	최저초상온도	1.63	최소상대습도	-2.58
평균최고기온	1.51	O3_max	1.45	평균상대습도	-4.32
평균기온	0.18	CO_mean	1.17	평균해면기압	-8.36
월적설량합	0	평균기온	-0.37	최고해면기압	-8.76
최저초상온도	-0.04	평균최고기온	-0.57	최저해면기압	-35.19
평균최저기온	-0.63	평균최저기온	-0.85	평균현지기압	-61.34
최고기온	-0.92				
최저기온	-1.2				
평균지면온도	-2.37				

결론

위험지역 분석

- SaTScan을 이용하여 연도별 장감염 발생의 고위험 지역을 분석한 결과 2009년부터 2012년까지 경상도, 강원도가 대체적으로 장감염 발생 위험 지역군에 속하였으며, 2013년에는 경상도와 전라도 부근이 장감염 발생 위험 지역군에 속함
- 장감염 발생 지역에 대해 클러스터링한 결과 2009년에는 고위험 지역에 속하는 시군구가 35개, 2010년에는 57개, 2011년 87개, 2012년 62개, 2013년 35개이며, 각 연도별 고위험 지역의 상대적 위험도는 2009년 0.75, 2010년 0.85, 2011년 0.86, 2012년 0.82, 2013년 1.28로 점점 증가하는 경향을 보임

장감염 예측 결과

- 전체 장감염 질환에 대해서는 OLS 회귀분석과 LASSO 회귀분석의 RMSE가 각각 20.040, 20.033이고 심층신경망의 RMSE는 15.656으로 에러가 대략 25% 감소하여 모델의 성능이 향상됨을 확인
- 기타 세균성 장감염과 OLS 회귀분석과 LASSO 회귀분석의 RMSE가 모두 4.342이고 심층신경망의 RMSE가 4.071로 에러가 대략 10% 감소하였으며, 바이러스 및 기타감염성 장감염의 경우 OLS 회귀분석과 LASSO의 회귀분석의 RMSE가 각각 5.440, 5.431이고 심층신경망의 RMSE가 5.042로 에러가 대략 10% 감소하여 모든 예측에서 심층신경망 모델의 성능이 향상됨을 확인

민감도 분석 결과

- 장감염 질환 전체에 대해 반사실적 실험을 수행한 결과 기후적 요인에서는 평균최고기온, 최저해면기압, 평균수증기압, 일조율, 평균최저초상온도 등 기온과 관련된 변수가 장감염 질환 전체에 대해 영향을 미치는 중요 변수인 것으로 파악됨
- 기타 세균성 장감염의 경우 평균기온, 최고기온, 최저기온 및 강수량, 일조율이 중요 변수인 것으로 나타났으며 기온이 낮아질수록 질환의 발생이 증가할 것이라는 실험결과가 나옴
- 바이러스 및 기타감염성 장감염의 경우 기압과 상대습도가 질환 발생에 많은 영향을 미치는 것으로 나타남