The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.3.549
29(3), 549–557

A comparison of imputation methods for the consecutive missing temperature data

Hee-Kyung Kim[a] · In-Kyeong Kang[a] · Jae-Won Lee[b] · Yung-Seop Lee[a,1]

[a]Department of Statistics, Dongguk University; [b]KMA National Climate Data Center

(Received March 23, 2016; Revised March 30, 2016; Accepted March 31, 2016)

Abstract

Consecutive missing values are likely to occur in long climate data due to system error or defective equipment. Furthermore, it is difficult to impute missing values. However, these complicated problems can be overcame by imputing missing values with reference time series. Reference time series must be composed of similar time series to time series that include missing values. We performed a simulation to compare three missing imputation methods (the adjusted normal ratio method, the regression method and the IDW method) to complete the missing values of time series. A comparison of the three missing imputation methods for the daily mean temperatures at 14 climatological stations indicated that the IDW method was better thanx others at south seaside stations. We also found the regression method was better than others at most stations (except south seaside stations).

Keywords: consecutive missing value, missing value imputation, adjusted normal ratio methods, regression method, IDW method

1. 서론

우리나라는 1904년 근대기상관측을 시작으로 현재 100여년이 넘는 장기간의 기후자료를 축적해오고 있다. 보다 정확한 기후예보를 위해서는 이러한 장기간의 기후자료에 대한 정확한 분석과 연구가 필수적이다. 그러나 장기간의 기후 자료가 누적되다 보면 자료의 수집과정에서 자료 수집자의 실수나 시스템적 오류, 측정 장비의 고장 등의 원인으로 결측이 발생하여 완전한 시계열관측 자료를 얻어내기 어려운 경우가 종종 발생하게 된다. 특히 관측 장비의 고장과 같은 원인은 연속적인 결측을 발생시키게 된다. 일반적으로 결측값은 원래의 자료 분포를 왜곡시킬 가능성이 있으며, 많은 결측을 포함한 자료에 대한 분석은 분석 결과에 대한 신뢰성을 잃을 수 있다. 이러한 결측 자료에 대해서는 효율적이고 과학적인 방법을 사용하여 결측값을 새로운 값으로 대체(imputation)함으로써 정보 손실을 막을 수 있고, 정보의 손실을 최소화 하는 범위 내에서 결측 자료를 대체 한 이후 향후 분석을 수행하면 분석 결과의 예측력을 높일 수 있다 (Lee, 2003).

This work was funded by the Korea Meteorological Administration Research and Development Program under Grant KMIPA 2015-1020.

[1]Corresponding author: Department of Statistics, Dongguk University, 30 Pildong-ro 1-gil, Jung-gu, Seoul 04620, Korea. E-mail: yung@dongguk.edu

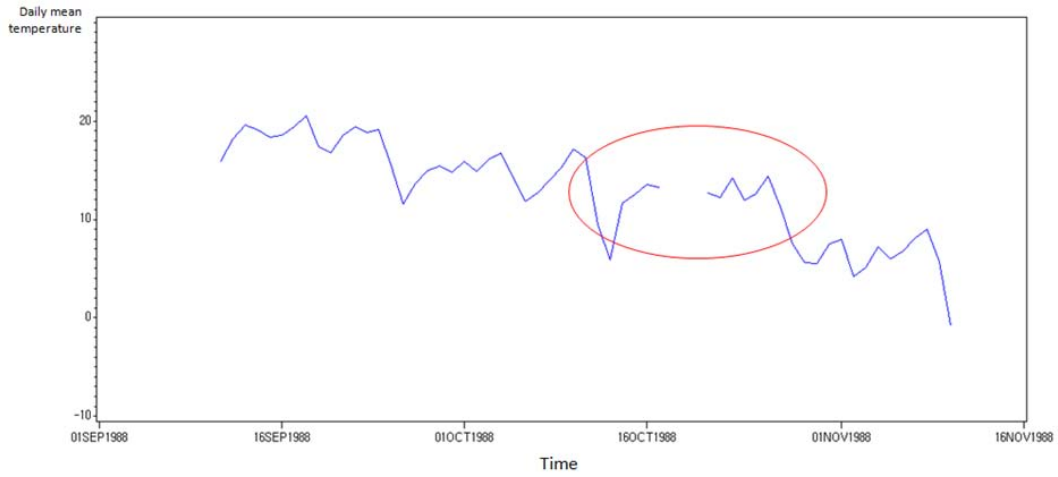


Figure 1.1. Daily mean temperatures with consecutive 3 missings at Gawnaksan(116).

기후관측 자료의 경우 관측 장비의 결함이나 일정시기의 자료 손실로 인해 발생한 연속적인 결측 형태를 갖는 시계열 결측 자료를 대체하는 것에 어려움이 따른다. Figure 1.1은 이러한 연속적 결측을 보여주는 실례로, 우리나라 기후관측소 중 관악산(116) 지점의 일평균기온을 나타낸 것이다. 그림을 보면 1988년 10월 16일부터 18일까지 연속적으로 3일의 일평균기온값이 결측인 것을 알 수 있다. 이와 같이 연속적인 시계열 결측 자료를 대체하기 위해서는 같은 시점에 관측된 다른 주변 지점(surrounding stations)의 시계열 자료로 결측 자료를 대체하는 것이 좋다. 이때 주변 지점은 결측이 발생한 지점의 시계열 자료와 관련성이 높은 지점의 자료일수록 복원의 정확도가 높을 것으로 예상해 볼 수 있다. 이에 대한 선행 연구로 Paulhus와 Kohler (1952)는 산술평균법(simple arithmetic average method)과 정규화비율 방법(normal ratio method)을 이용하여 결측된 시계열 관측 자료를 대체하기 위해 같은 시점에 관측된 다른 시계열 자료를 이용하였다. Young (1992)도 마찬가지로 같은 시점에 관측된 다른 자료를 이용하여 다중판별 분석과 회귀 방법(regression method), 수정된 정규화비율 방법(adjusted normal ratio method)으로 시계열 결측 자료를 대체하였다. 해외에서는 다양한 기법들을 비교분석하면서 기후자료 결측에 대한 복원 문제를 오랫동안 연구하였으며 현재에도 활발히 진행되고 있다 (Azman 등, 2015). 국내에서는 최근 우리나라 기후관측소 부산(159) 지점의 기온자료를 대상으로 데이터마이닝 기법인 support vector machine(SVM) 방법을 적용하고 다른 결측복원 기법들과 이를 비교분석한 Jung (2014)의 연구가 있었다. 하지만 이는 부산(159) 지점에 대한 제한적 결과일 뿐, 우리나라를 대표할 수 있는 기후관측소들의 자료를 모두 이용하여 우리나라의 기후자료에 적합한 결측복원 기법을 개발한 연구는 아직 없었다.

본 연구에서는 우리나라를 대표할 수 있는 14개 기후관측소의 일평균기온 자료에 대한 시뮬레이션을 통해 수정된 정규화비율 방법, 회귀 방법, IDW 방법 등을 적용하여 결측값들에 대한 복원의 정확도를 비교분석하여 우리나라 기온자료에 적합한 결측복원 기법을 찾고자 한다.

2. 연구 방법

자료의 결측이 발생한 경우 주변지점자료를 이용하여 복원값을 산출하게 되는데, 이때 결측이 발생한 지점과 관련성이 높은 주변지점의 연속적인 기후자료를 참조시계열(reference series)이라 한다. 본 연구

에서는 결측복원 기법들간의 비교분석을 위해 1985년 1월 1일부터 2014년 12월 31일까지 30년간의 일평균기온자료를 이용하였다. 결측이 연속적으로 3개 발생한 경우, 4개 발생한 경우, 5개 발생한 경우에 대해 각각 시뮬레이션을 실시하였다. 자료는 우리나라를 대표할 수 있는 춘천(101), 강릉(105), 서울(108), 인천(112), 대전(133), 포항(138), 부산(159), 목포(165), 여수(168), 제주(184), 대구(143), 광주(156), 전주(146), 울산(152) 등의 14개 기후관측소의 일평균기온자료를 이용하여 시뮬레이션을 실시하였다. Figure 2.1에 분석대상인 14개 지점의 기후관측소의 분포가 나타나있다.

임의로 연속적인 결측을 발생시킨 후 결측 발생 시점을 기준으로 전후 각각 60일 간의 시계열 자료를 사용하여 결측이 발생한 지점과 가장 유사한 주변지점을 선정한다. Durre 등 (2010)에서는 대상지점을 중심으로 75km 이내의 주변지점 중 기온자료와 상관성이 가정 높은 지점을 선정하기 위한 측도로써 Legates와 McCabe (1999)가 제시한 d-index를 이용하였다. 본 연구에서도 이와 비슷하게 검정대상 지점과의 거리가 반경 70km 이내인 주변지점 중 d-index 값이 가장 높은 3개 지점의 기온자료를 참조시계열로 선택하였다. Legates와 McCabe (1999)가 제시한 d-index는 식 (2.1)과 같다.

$$d = 1.0 - \frac{\sum_{i=1}^{n} |X_i - Y_i|}{\sum_{i=1}^{n} \left[\left| Y_i - \bar{X} \right| + \left| X_i - \bar{X} \right| \right]}. \tag{2.1}$$

여기서 Y_i 결측이 발생한 지점의 시계열이고 X_i는 참조시계열, $\bar{X}$는 참조시계열의 평균이다. 또한 n은 분석에 사용되는 관측치 수를 의미한다.

d-index를 기준으로 관련성이 높은 주변지점을 선정하여 참조시계열을 선택한 후, 대표적인 결측복원 기법이라고 할 수 있는 수정된 정규화비율 방법, 회귀 방법, IDW 방법을 적용하여 결측값을 대체하였다. 각 방법에 대한 간략한 내용은 다음과 같다.

2.1. 수정된 정규화비율 방법(adjusted Normal ratio method)

정규화비율 방법은 Paulhus와 Kohler (1952)가 제안한 방법이다. Young (1992)은 결측이 존재하는 시계열과 참조시계열 간의 상관계수를 이용하여 새로운 가중치 w_i을 만들고 이를 이용하여 가중 평균하는 방법으로 수정된 형태의 정규화비율 방법을 제안하였다. Young (1992)이 제안한 이 방법을 수정된 정규화비율 방법이라 부르겠다. r_i를 결측이 존재하는 시계열과 i번째 참조시계열 사이의 상관계수라 하면 i번째 참조시계열에 대한 새로운 가중치 w_i는 식 (2.2)와 같이 정의할 수 있다.

$$w_i = \frac{r_i^2 (n_i - 2)}{1 - r_i^2}, \quad i = 1, 2, \ldots, K. \tag{2.2}$$

결측값은 식 (2.3)과 같이 수정된 정규화비율 방법에 의하여 추정된다. 결측 대체값 $\hat{Y}_t$은 t시점의 결측에 대한 추정값으로 K개의 참조시계열의 관측값을 새로운 가중치 w_i를 이용하여 가중 평균함으로써 계산된다.

$$\hat{Y}_t = \frac{1}{\sum_{i=1}^{K} w_i} (w_1 X_{t1} + w_2 X_{t2} + \cdots + w_K X_{tK}). \tag{2.3}$$

2.2. 회귀 방법(regression method)

회귀 방법은 오차들의 제곱합을 최소로 하는 최소제곱법을 이용하는 방법으로 오차들이 서로 독립이고

정규분포를 따를 때 효과적인 방법이다. 결측값 대체를 위해 다음과 같은 회귀모형을 가정한다.

$$Y_t = \beta_0 + \beta_1 X_{t1} + \cdots + \beta_k X_{tK} + \epsilon_t, \quad \epsilon_t \sim N\left(0, \sigma^2\right),$$

여기서 $t = 1, \ldots, T$는 자료가 관측된 시점으로 이 기간의 관측 자료는 결측이 존재하지 않는 완전한 자료라고 가정한다. K개의 참조시계열 $X_{t1}, X_{t2}, \ldots, X_{tK}$를 독립변수로 사용하고, 결측값을 포함하는 시계열 Y_t를 종속변수로 한다. 이와 같은 모형에서 최소제곱법에 의해 예측값과 실제값의 제곱편차합을 최소로 하는 모수 $\beta_0, \beta_1, \ldots, \beta_k$를 추정한다. 추정된 회귀계수 $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$을 이용하여 구축된 회귀모형에 결측이 존재하는 시점과 동일한 시점의 참조시계열 관측값을 대입하면 분석대상 시계열의 결측값을 추정할 수 있다. 즉, t시점의 결측값은 아래의 식 (2.4)에 의해 추정된다.

$$\hat{Y}_t = \beta_0 + \hat{\beta}_1 X_{t1} + \hat{\beta}_2 X_{t2} + \cdots + \hat{\beta}_k X_{tk}. \tag{2.4}$$

2.3. IDW 방법(inverse distance weighting method)

IDW 방법은 일반적인 결측자료 보다 기후관측소의 기후자료에 결측이 발생한 경우 이를 복원하는데 많이 이용되는 방법으로 결측값을 복원하고자 하는 지점과 가까운 지점들의 거리를 가중치로 사용하여 결측값을 복원하는 방법이다 (Di Piazza 등, 2011). 기본적으로 거리가 가까운 주변 지점에 더 큰 가중치를 부여하게 되고, 거리가 먼 주변 지점에 대해서는 작은 가중치를 부여하게 된다. 결측이 존재하는 지점과 주변 지점(참조 시계열) 사이의 거리의 역수를 가중치 w_i^*로 정의하고, 다음의 식 (2.5)에 의해 t시점의 결측값을 추정할 수 있다.

$$\hat{Y}_t = \frac{\sum_{i=1}^{K} [X_{ti} w_i^*]}{\sum_{i=1}^{K} w_i^*}, \tag{2.5}$$

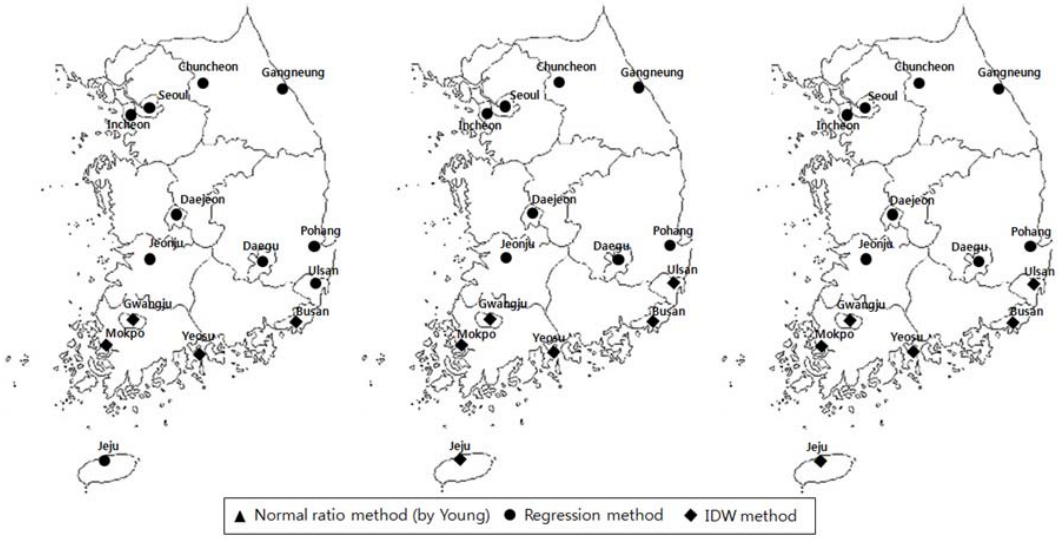
여기서 $\hat{Y}_t$는 t시점에 대한 예측값이고 X_{ti}는 i번째 주변지점의 시계열값을 의미한다. IDW 방법은 기후관측소들의 밀도가 높은 경우 결측복원의 정확도가 높은 것으로 알려져 있다. 현재 IDW 방법은 가중치 산정에 있어 다양한 방법들을 적용시켜 IDW 방법에 대한 다양한 변형을 다룬 연구들이 있다 (Teegavarapu와 Chandramouli, 2005; You 등, 2008).

결측에 대한 추정값이 실제값과 얼마나 가까운지를 평가함으로써 앞에서 제시한 결측복원 기법들의 정확성을 비교해 볼 수 있다. 실제값과 추정된 결측값과의 차이를 평가하는 정확도 측도로서 식 (2.6)의 평균 제곱근 오차(root mean square error; RMSE)와 식 (2.7)의 평균절대오차(mean absolute error; MAE)를 사용하였다.

$$\mathrm{RMSE} = \sqrt{\sum_{t'=1}^{M} \frac{\left(Y_{t'} - \hat{Y}_{t'}\right)^2}{M}}, \tag{2.6}$$

$$\mathrm{MAE} = \frac{1}{M} \sum_{t'=1}^{M} \left| Y_{t'} - \hat{Y}_{t'} \right|, \tag{2.7}$$

여기서 t'은 전체 자료 기간 중 결측이 존재하는 시점으로 전체 자료에 M개의 결측($t' = 1, 2, \ldots, M$)이 존재한다. $Y_{t'}$는 t'시점에 결측인 시계열의 실제값이고 $\hat{Y}_{t'}$는 t'시점의 결측복원 기법에 의한 추정한 값이다. RMSE 값과 MAE 값이 작을수록 결측 추정값이 실제값과 유사하다고 말할 수 있으며, 결측복원 기법의 정확성이 높다고 할 수 있다.



(a) 3 Missings (b) 4 Missings (c) 5 Missings

Figure 3.1. Best imputation method at 14 stations for each missings case.

3. 연구결과

14개 기후관측소의 일평균기온에 대해 임의로 연속적인 결측을 발생시키고 수정된 정규화비율 방법, 회귀 방법, IDW 방법을 각각 적용시켜 결측값을 대체한 결과의 정확도로 RMSE 값이 Table 3.1에 나타나있다. Table 3.1의 각 지점에 대한 RMSE 값은 계절적 요인을 고려하기 위해 자료를 봄(3월–5월), 여름(6월–8월), 가을(9월–11월), 겨울(12월–2월)의 4계절로 구분하고 각 계절마다 자료에서 임의의 연속적 결측을 발생시켜 시뮬레이션 한 결과의 평균이다. 즉, 각 지점별로 4계절에 대한 RMSE 값을 평균한 결과이고, 가장 정확한 것으로 나타난 복원 기법을 진하게 표시하였다. Table 3.1의 RMSE 값을 살펴보면 연속적으로 3일의 결측이 발생한 경우 부산(159), 목포(165), 여수(168), 광주(156) 지점의 일평균기온에 대해서는 IDW 방법이 가장 정확한 것으로 나타났으며, 그 외 지점에서는 대부분 회귀 방법이 가장 정확한 것으로 나타났다. 연속적 결측이 4일, 5일인 경우도 부산(159), 목포(165), 여수(168), 제주(184), 광주(156), 울산(152) 지점의 일평균기온에 대해서는 IDW 방법이 가장 정확한 것으로 나타났으며, 그 외 지점에 대해서는 대부분 회귀 방법이 가장 정확한 것으로 나타났다. Table 3.1의 가장 아래에 나타나있는 값은 14개 지점의 분석결과(RMSE)를 모두 평균한 것으로 회귀 방법이 가장 정확한 것을 알 수 있다. 이러한 결과는 Table 3.2의 MAE 값을 살펴보아도 동일한 결론을 얻을 수 있다. Table 3.2의 MAE 값은 Table 3.1의 RMSE 값과 마찬가지로 각 지점별로 4계절에 대한 결과를 평균한 값으로 MAE 값을 기준으로 부산(159), 목포(165), 여수(168), 제주(184), 광주(156), 울산(152) 지점 등에서는 IDW 방법이 가장 정확한 것으로 나타났고, 그 외 지점에서는 대부분 회귀 방법이 가장 정확한 것으로 나타났다.

시뮬레이션 결과 RMSE 값과 MAE 값을 기준으로 가장 정확한 것으로 나타난 결측복원 기법이 분석대상의 지형적인 특성에 따라 달라지는지를 살펴보기 위해 기후관측소별로 가장 정확한 것으로 나타났던 결측복원 기법을 지도상에 표시해 본 결과 Figure 3.1과 같았다. 14개 지점 중 부산(159), 목포(165), 여수(168), 제주(184), 광주(156), 울산(152) 지점 등과 같이 남쪽 해안가에 위치한 기후관측소의 경우

Table 3.1. Missing imputation simulation results for daily mean temperatures (RMSE)

Station	Imputation method	RMSE		
		3 Missings	4 Missings	5 Missings
Chuncheon (101)	Adjusted normal ratio method	0.4909	0.5495	0.5166
	Regression method	**0.3069**	**0.4822**	**0.4124**
	IDW method	0.4497	0.5144	0.5054
Gangneung (105)	Adjusted normal ratio method	0.8976	0.6144	0.7650
	Regression method	**0.3635**	**0.3717**	**0.2724**
	IDW method	0.6843	0.5326	0.4379
Seoul(108)	Adjusted normal ratio method	1.1135	0.5441	0.8759
	Regression method	**0.9051**	**0.3772**	**0.7005**
	IDW method	1.1177	0.5052	0.8498
Incheon(112)	Adjusted normal ratio method	0.7631	0.9590	0.8964
	Regression method	**0.3170**	**0.6719**	**0.5082**
	IDW method	0.7843	0.9434	0.9124
Daejeon(133)	Adjusted normal ratio method	0.6877	0.5074	0.4275
	Regression method	**0.6333**	**0.3707**	**0.3038**
	IDW method	0.6390	0.5044	0.4484
Pohang(138)	Adjusted normal ratio method	1.1112	0.9433	1.2909
	Regression method	**0.4314**	**0.4008**	**0.6119**
	IDW method	1.0577	0.8268	1.2607
Busan(159)	Adjusted normal ratio method	0.6346	0.5666	0.7604
	Regression method	0.6334	0.7146	1.0449
	IDW method	**0.6191**	**0.5557**	**0.6746**
Mokpo(165)	Adjusted normal ratio method	0.8065	0.8854	0.5314
	Regression method	1.0638	1.0275	0.4866
	IDW method	**0.7244**	**0.8374**	**0.4459**
Yeosu(168)	Adjusted normal ratio method	0.4456	0.4617	0.5110
	Regression method	0.5214	0.4602	0.5124
	IDW method	**0.4174**	**0.4588**	**0.4671**
Jeju(184)	Adjusted normal ratio method	0.5185	0.9900	0.5690
	Regression method	**0.4719**	1.1026	0.6714
	IDW method	0.5308	**0.9290**	**0.5620**
Daegu(143)	Adjusted normal ratio method	1.7226	1.9529	1.8331
	Regression method	**0.5243**	**0.9203**	**0.9945**
	IDW method	1.6168	1.7777	1.7827
Gwangju(156)	Adjusted normal ratio method	0.5754	0.6829	0.6550
	Regression method	0.6018	0.7251	0.8050
	IDW method	**0.5558**	**0.6508**	**0.6434**
Jeonju(146)	Adjusted normal ratio method	0.6321	0.7052	0.8800
	Regression method	**0.4262**	**0.4220**	**0.6655**
	IDW method	0.7247	0.8089	0.9347
Ulsan(152)	Adjusted normal ratio method	0.4821	0.5374	0.4834
	Regression method	**0.4304**	0.5416	0.5287
	IDW method	0.5410	**0.4852**	**0.4519**
Mean	Adjusted normal ratio method	0.7772	0.7786	0.7854
	Regression method	**0.5450**	**0.6135**	**0.6085**
	IDW method	0.7473	0.7379	0.7412

RMSE = root mean square error.

Table 3.2. Missing imputation simulation results for daily mean temperatures (MAE)

Station	Imputation method	MAE		
		3 Missings	4 Missings	5 Missings
Chuncheon (101)	Adjusted normal ratio method	0.4649	0.4599	0.4194
	Regression method	**0.2837**	**0.4115**	**0.3527**
	IDW method	0.4302	0.4390	0.4127
Gangneung (105)	Adjusted normal ratio method	0.8665	0.5211	0.7223
	Regression method	**0.2861**	**0.3489**	**0.2264**
	IDW method	0.6369	0.4991	0.3700
Seoul(108)	Adjusted normal ratio method	1.0232	0.4910	0.7423
	Regression method	**0.8075**	**0.2864**	**0.4573**
	IDW method	1.0425	0.4598	0.7238
Incheon(112)	Adjusted normal ratio method	0.7412	0.8855	0.8387
	Regression method	**0.2888**	**0.6014**	**0.4359**
	IDW method	0.7525	0.8400	0.8524
Daejeon(133)	Adjusted normal ratio method	0.6522	0.4576	0.3690
	Regression method	**0.6039**	**0.3066**	**0.2502**
	IDW method	0.6127	0.4542	0.3846
Pohang(138)	Adjusted normal ratio method	1.0740	0.8816	1.1680
	Regression method	**0.3388**	**0.3223**	**0.5481**
	IDW method	1.0277	0.7711	1.1523
Busan(159)	Adjusted normal ratio method	0.5294	0.4815	0.6514
	Regression method	0.5459	0.6611	0.9061
	IDW method	**0.4979**	**0.4668**	**0.5682**
Mokpo(165)	Adjusted normal ratio method	0.7204	0.8055	0.4496
	Regression method	0.9626	0.9145	0.4106
	IDW method	**0.6331**	**0.7567**	**0.3944**
Yeosu(168)	Adjusted normal ratio method	0.3573	0.3697	0.4062
	Regression method	0.4711	0.4014	0.4092
	IDW method	**0.3509**	**0.3687**	**0.3728**
Jeju(184)	Adjusted normal ratio method	0.4956	0.8587	0.4653
	Regression method	**0.4015**	0.9888	0.5431
	IDW method	0.4817	**0.8033**	**0.4571**
Daegu(143)	Adjusted normal ratio method	1.6832	1.8465	1.7049
	Regression method	**0.4689**	**0.6126**	**0.8660**
	IDW method	1.5685	1.5133	1.6460
Gwangju(156)	Adjusted normal ratio method	0.4581	0.5808	0.5840
	Regression method	0.5138	0.6202	0.7306
	IDW method	**0.4468**	**0.5499**	**0.5715**
Jeonju(146)	Adjusted normal ratio method	0.5718	0.6295	0.7792
	Regression method	**0.3467**	**0.3594**	**0.5253**
	IDW method	0.6561	0.7351	0.8564
Ulsan(152)	Adjusted normal ratio method	0.3947	0.4701	0.3938
	Regression method	**0.3466**	0.4877	0.4520
	IDW method	0.4669	**0.4319**	**0.3799**
Mean	Adjusted normal ratio method	0.7166	0.6956	0.6924
	Regression method	**0.4761**	**0.5348**	**0.5081**
	IDW method	0.6860	0.6603	0.6530

MAE = mean absolute error.

IDW 방법이 가장 정확한 것으로 나타났고, 남쪽 해안가를 제외한 대부분의 기후관측소에서는 대체로 회귀 방법이 가장 정확한 것으로 나타났다. 이러한 특징은 결측이 3일인 경우, 4일인 경우, 5일인 경우에서 모두 유사하게 나타나고 있다.

4. 결론 및 향후과제

장기간의 기후 자료가 누적되다 보면 자료의 수집과정에서 자료 수집자의 실수나 시스템적 오류, 측정장비의 고장 등과 같은 원인으로 결측이 발생하여 완전한 시계열관측 자료를 얻어내기 어려운 경우가 종종 발생하게 된다. 특히 관측 장비의 고장과 같은 원인은 연속적인 결측을 발생시키게 된다. 이러한 경우 적합한 방법을 통해 결측을 복원시키면 정보의 손실을 막을 수 있고, 나아가 분석의 신뢰성을 높일 수 있다. 본 연구에서는 우리나라 기후자료에 이러한 연속적 결측이 발생했을 때 가장 적합한 결측복원 기법을 찾고자 하였다. 따라서 우리나라를 대표할 수 있는 기후관측소 14개 지점의 일평균기온을 대상으로 연속적 결측을 임의로 발생시키고, 이를 수정된 정규화비율 방법, 회귀 방법, IDW 방법을 적용하여 복원시키는 시뮬레이션을 수행하였다. 그 결과 남쪽 해안가에 위치한 지점에 대해서는 IDW 방법이 비교적 정확한 것으로 나타났고, 남쪽 해안가를 제외한 대부분의 지점에서 회귀 방법이 다른 방법들에 비해 정확한 것으로 나타났다. 이는 일평균기온에 대한 분석 결과이고, 향후 강수량과 같은 다른 기후요소들에 대해서도 시뮬레이션을 적용시켜 어떤 복원 기법이 우리나라 기후자료에 대해 정확한 복원을 가능하게 하는지 확인해 볼 필요가 있다. 또한 본 연구에서는 참조시계열을 이용하여 결측을 복원하는 방법들에 대해 다루었지만, 울릉도(132) 지점과 같이 주변지점의 정보를 활용할 수 없는 경우가 있다. 이렇게 참조시계열을 사용할 수 없는 경우에 대해서도 적합한 결측복원 기법의 연구가 향후 이루어져야 할 것이다.

References

Azman, M. A., Zakaria, R., and Radi, N. F. A. (2015). Estimation of missing rainfall data in Pahang using modified spatial interpolation weighting methods, *The 2nd ISM International Statistical Conferencd 2014 (ISM-II): Empowering the Applications of Statistical and Mathematical Sciences*, **1643**, 65–72

Di Piazza, A., Lo Conti, F., Noto, L. V., Viola, F., and La Loggia, G. (2011). Comparative analysis of different techniques for spatial interpolation of rainfall data to create a serially complete monthly time series of precipitation for Sicily, Italy, *International Journal of Applied Earth Observation and Geoinformation*, **13**, 396–408.

Durre, I., Menne, M. J., Gleason, B. E., Houston, T. G., and Vose, R. S. (2010). Comprehensive automated quality assurance of daily surface observations, *National Climatic Data Center*, **49**, 1615–1633.

Jung, S.-Y. (2014). *A study of consecutive missing value imputation method using reference series in time series*, M.S. Thesis, Department of Statistics, Graduate School of Dongguk University, Seoul, Korea.

Lee, Y.-S. (2003). *Data Mining Cookbook by Olivia Parr Rud*, Kyowoo Publishing Company, Seoul.

Legates, D. R., and McCabe Jr., G. J. (1999). Evaluating the use of "goodness-of -fit" measures in hydrologic and hydroclimatic model evaluation, *Water Resources Research*, **35**, 233–241.

Paulhus, J. L. H. and Kohler, M. A. (1952). Interpolation of missing precipitation records, *Monthly Weather Review*, **80**, 129–133.

Teegavarapu, R. S. V. and Chandramouli, V. (2005). Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records, *Journal of Hydrology*, **312**, 191–206.

You, J. S., Hubbard, K. G., and Goddard, S. (2008). Comparison of methods for spatially estimating station temperatures in a quality control system, *International Journal of Climatology*, **28**, 777–787.

Young, K. (1992). A three-way model for interpolating for monthly precipitation values, *Monthly Weather Review*, **120**, 2561–2569.

연속적 결측이 존재하는 기온 자료에 대한 결측복원 기법의 비교

김희경[a] · 강인경[a] · 이재원[b] · 이영섭[a,1]

[a]동국대학교 통계학과, [b]기상청 국가기후데이터센터

(2016년 3월 23일 접수, 2016년 3월 30일 수정, 2016년 3월 31일 채택)

요 약

장기간의 기후 자료가 누적되다 보면 자료의 수집과정에서 시스템적 오류나 측정 장비의 고장 등으로 인하여 연속적 결측이 종종 발생하게 된다. 연속적인 결측 형태를 갖는 경우 시계열 결측 자료를 대체하는 것에 어려움이 따른다. 이러한 경우 참조시계열을 이용하여 결측값을 대체할 수 있다. 참조시계열은 결측이 발생한 시계열과 관련성이 높은 주변지점의 시계열로 구성할 수 있다. 본 연구에서는 결측값을 대체시킬 수 있는 3가지 결측복원 기법-수정된 정규화비율 방법, 회귀 방법, IDW 방법-을 비교하는 시뮬레이션을 수행하였다. 우리나라 14개 지점의 기후관측소의 일평균기온값을 대상으로 비교한 결과 남쪽 해안가에 위치한 기후관측소의 자료에 대해서는 IDW 방법이 가장 정확한 것으로 나타났으며, 그 외 지역의 기후관측소 자료에 대해서는 회귀 방법이 가장 정확한 것으로 나타났다.

주요용어: 연속적 결측, 결측값 대체, 수정된 정규화비율 방법, 회귀 방법, IDW 방법

이 연구는 기상청 “기상기술개발사업”(KMIPA 2015-1020)의 지원으로 수행되었습니다.

[1]교신저자: (04620) 서울특별시 중구 필동로 1길 30, 동국대학교 통계학과. E-mail: yung@dongguk.edu