KIGAS Vol.4, No.4, December, 2000
(Journal of the Korean Institute of Gas)

# 클러스터링 기법을 이용한 공정 데이터의 압축 저장 기법에 관한 연구

김 윤 식 · 모 경 주* · 윤 인 섭
*삼성SDS, 서울대학교 응용화학부
(2000년 12월 7일 접수, 2000년 12월 27일 채택)

# A Study on Process Data Compression Method by Clustering Method

Yoonsik Kim · Kyung Joo Mo* · En Sup Yoon
*Samsung SDS, Korea
School of Chemical Eng., Seoul National University, Korea
(Received 7 December 2000 ; Accepted 27 December 2000)

## 요 약

가스 저장 · 공급 시설을 포함한 화학공정에서 측정된 데이터를 효과적으로 이용하기 위하여 정보의 손실의 최소화하면서 데이터를 압축하여 저장하고 재생할 수 있는 방법에 대한 연구가 진행되어 왔다. 기존에 제안되었던 데이터 압축 저장 방법들의 단점을 극복하기 위하여, 부분 선형화 근사 방법과 $k$-means 클러스터링 알고리즘을 응용한 새로운 공정 데이터의 압축 방법을 제안하였다. 제안된 방법을 실공정 데이터에 적용하여 본 결과, 본 연구에서 제안된 방법이 기존의 방법보다 재현 능력이 우수함을 확인할 수 있었다.

**Abstract** - Data compression and retrieval method are investigated for the effective utilization of measured process data. In this paper, a new data compression method, Clustering Compression(CC), which is based on the $k$-means clustering algorithm and piecewise linear approximation method is suggested. Case studies on industrial data set showed the superior performance of clustering based techniques compared to other conventional methods and showed that CC could handle the compression of multi-dimensional data.

**Key words** : data compression, clustering, multi-dimensional data

## 1. 서 론

최근 안전과 환경에 대한 관심 증가, 그리고 운전 효율 향상에 대한 필요성의 증대로 인하여 공정의 데이터를 수집과 저장에 대한 관심이 증대되었다. 실시간 데이터 베이스(Real-Time Database)와 같은 컴퓨터 기술의 발전으로 인하여 방대한 양의 공정 데이터를 손쉽게 수집하고 저장 할 수 있게 되었다. 최근에는 방대한 데이터를 효과적으로 보관하기 위하여 압축 방법을 이용하여 공정의 모든 데이터를 저장하지 않고 전체를 대표하는 일부분의 데이터만을 저장하는 방법에 대하여 관심이 증대되고 있다.

현재 가스 저장 · 공급 시설을 포함한 대부분의 화학공정에서는 측정된 데이터를 평균값

†주저자 : seeker@pslab.snu.ac.kr

형태로 환산하여 저장하고 있다. 즉, 실시간으로 수집된 공정 데이터들의 값을 모두 저장하지 않고 이들의 1분 평균값, 15분 평균값 혹은 1시간 평균값 등의 형태로 변환하여 저장하는 방법을 사용하고 있다. 이와 같이 공정의 데이터를 시간의 평균값으로 환산하여 저장하는 경우에는 해당 시간 구간 사이에서 발생한 공정의 과도 거동(transient behavior)을 정확하게 묘사하지 못하게 되므로 정보의 손실을 가져오게 된다. 이에, 정보의 손실을 최소화시키면서도 효율적으로 데이터를 압축하여 저장하고 이를 재생하는 방법에 대한 연구가 진행되었다.

현재까지 제안된 데이터를 압축 방법으로는 실시간으로 사용하기에 적합한 부분 선형화 근사 방법[1,2,3,4,5]으로는 Box Car, Backward Slope, Box Car와 Backward Slope의 혼합방법, SDT(Swinging Door Trending), 그리고 PLOT 방법 등이 있다. 그러나 이 방법들은 한번에 하나의 변수만을 다루는 방법으로 상관관계가 커서 함께 감시하여야 할 공정 데이터들의 특성을 고려하지 못한다. 이에, 본 연구에서는 향상된 데이터의 압축 효율을 보이며, 다차원 데이터로 확장할 수 있도록 기존의 부분 선형화 근사 방법을 발전시켜 $k$-means 클러스터링을 이용한 데이터 압축 방법을 제안하고자 한다. 그리고 부록에서는 본 연구에서 제안된 클러스터링을 이용한 압축 후 곡선으로 데이터를 재현하는 방법에 대하여 고찰한다.

## 2. $k$-means 클러스터링 기법

클러스터링이란 주어진 표본 벡터들을 분류하여 각각의 값들을 대표하는 대표값으로 각각의 데이터를 분류하는 방법이다. $k$-means 클러스터링 알고리듬이란 식 (1)과 같이 주어진 $k$개의 대표값을 이용하여 N개의 표본 벡터 $\overline{x_i}$에 대하여 가장 가까운 중심 $\overline{x_a}$와의 유클리디안 거리 E를 최소화시키는 프로세싱 유니트의 중심(prcoessing unit center)을 찾는 방법이다[6].

$$E = \sum_{a=1,k} \sum_{i=1,N} M_{a_i} (\overline{x_a} - \overline{x_i})^2 \qquad (1)$$

여기서, $M_{a_i}$는 클러스터 분할 혹은 멤버쉽 함수로써 0과 1로만 구성된 k×N 행렬이다. 각 열은 데이터를 의미하며 각 행은 클러스터를 의미하는데, 각 열은 하나의 "1"을 가지고 있으며 데이터에 해당하는 클러스터를 식별하게 된다.

가스 및 화학 공정 데이터의 실시간 처리에 사용하기 위해서는 식 (2)와 같은 방법으로 온라인 버전을 사용하여 중심값을 갱신한다.

$$\Delta\overline{x^a} = \eta(\overline{x_i} - \overline{x_a}) \qquad (2)$$

즉, 이전의 중심값을 바탕으로 새롭게 입력된 데이터를 표현하기 위하여, 중심값을 현재값과의 차이를 학습률 $\eta$만큼 보정하여 갱신한다.

## 3. 압축 저장 방법의 구현

CC(Clustering Compression)는 새롭게 입력되는 데이터와 중심값을 비교하여 기록 한계 내에 들어 있는 경우에는 중심값을 갱신하고, 기록 한계를 벗어나는 경우에는 현재의 데이터로 중심값을 새롭게 할당하는 방법을 사용한다. 즉, $k$ 값을 1로 사용하여 클러스터링을 수행하는 것이다. CC 방법에서는 유클리디안 거리를 사용하므로 손쉽게 다차원으로도 확장할 수 있다. 현재 시간 t에 들어온 데이터를 $x_t$, 기록 한계값을 $\sigma$라고 정의할 때 아래와 같은 압축 알고리즘으로 해당 구간을 대표하는 중심값과 시간값을 기록한다(Fig. 1).

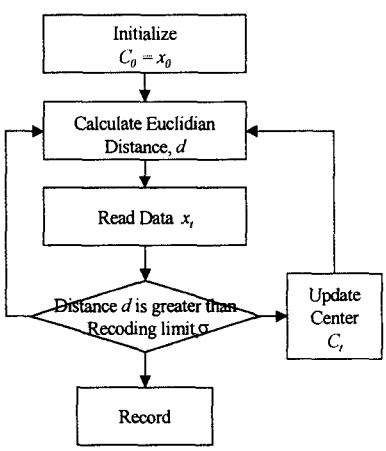


**Fig. 1. Algorithm of clustering compression.**

① 거리의 계산

기존의 중심값과 새롭게 입력된 데이터와의 표준화된 거리 $d$를 계산한다. 표준화된 거리는 중심값과 데이터와의 거리를 기록 한계, $\sigma$로 나눈 값이다.

$$d = \sqrt{(\frac{x_t - c_{t-1}}{\sigma})^2} \qquad (3)$$

② 중심값의 갱신 및 할당

앞서 계산된 거리 $d$가 지정된 기록 한계보다 작은 경우에는 식 (4)과 같이 그 차이를 학습률(learning rate) $\alpha$만큼 보정하여 중심값을 갱신한다. 이때, $\alpha$는 처음에는 큰 값을 사용하다가 0과 1 사이의 값을 가지는 새로운 매개변수 $\beta$를 사용하여 지속적으로 학습률 $\alpha$를 줄여나가면서, 중심값을 갱신하게 된다. 표준화된 거리 d가 기록 한계를 벗어나는 경우에는 대표값을 저장하고 새롭게 입력된 데이터로 중심값을 할당하고 학습률 $\alpha$를 $\alpha_0$로 다시 초기화시킨다. 이와 같은 방법으로 주어진 구간에서의 SSE(Squared Sum Error)를 최소화할 수 있게 된다.

Fig. 2에서는 유클리디안 거리 $d$가 기록 한계보다 작을 때의 각 공정데이터에 따른 클러스터의 중심값의 변화와 학습률의 변화를 보여주고 있다. 즉 공정 데이터가 점차 안정된 경향을 보임에 따라 학습률은 지속적으로 작아져서 클러스터의 중심의 갱신의 정도는 점차 줄어들어 일정한 값에 수렴하는 것을 볼 수 있다.

$$\begin{aligned}
&\text{if} \quad (\ d < 1\ ) \\
&\text{then} \quad c_t = c_{t-1} + \alpha_{t-1} \times (x_t - c_{t-1}) \\
&\qquad\quad \alpha_t = \alpha_{t-1} \times \beta, \qquad 0 < \beta < 1 \\
&\text{else} \quad \text{record} \qquad c_{t-1} \text{ or } x_{t-1} \\
&\qquad\quad \text{update} \qquad c_t = x_t \\
&\qquad\qquad\qquad\quad \alpha_t = \alpha_0
\end{aligned} \qquad (4)$$

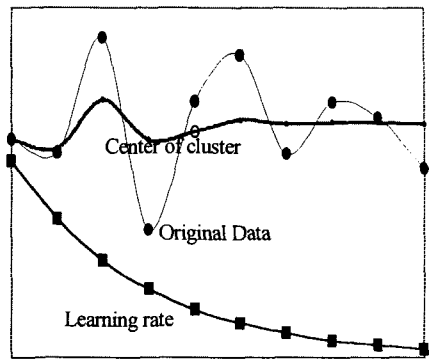


**Fig. 2. Update of trajectory of center value(1-D case).**

③ 대표값의 저장 및 재현

CC에서 저장하는 값으로는 기존의 방법과 동일하게 변화를 일으키기 바로 직전 공정값 $x_{t-1}$을 사용할 수도 있고 해당 구간을 대표하는 클러스터의 중심값, $c_{t-1}$을 사용 할 수도 있다. 저장하는 시간의 경우에도 변화를 일으키기 직전 시간이나 해당 구간의 중앙 시간으로 저장할 수 있다.

그리고 재현 방법에 있어서도 대표값들을 1차 직선으로 재현하는 방법 그리고 상수로 재현하는 방법을 사용할 수 있다.

이를 정리하면, 시간 $t_i \le t \le t_{n-1}$에서의 공정 데이터 $\{x_{t_i}, \cdots, x_{t_{n-1}}\}$의 기록 방법과 재현 방법은 Table 1과 같이 요약될 수 있다.

**Table 1. Cases of clustering compression**

| case | 기록값 | 기록시간 | 재현방법 |
|---|---|---|---|
| CL | 중심값 | $t_{n-1}$ | 1차 |
| CT | 중심값 | $(t_{n-1}-t_i)/2$ | 1차 |
| CV | 중심값 | $t_{n-1}$ | 0차 |
| PV | 공정값 | $t_{n-1}$ | 1차 |

## 4. 다차원 압축 방법

기존의 방법들은 다차원 데이터를 저장할 수 없었으므로 상관관계가 커서 함께 감시해야 할 변수들을 독립적으로 저장해야만 했다. 그러나, 본 연구에서 제안된 클러스터링을 이용한 압축 저장 방법은 다음과 같은 방법으로 다차원 데이터의 압축에 적용할 수 있다.

$$d = \sqrt{\left(\frac{x_1 - c_1}{\sigma_1}\right)^2 + \cdots + \left(\frac{x_i - c_i}{\sigma_i}\right)^2 + \cdots + \left(\frac{x_n - c_n}{\sigma_n}\right)^2} \tag{5}$$

if $d \le 1$

then $c_t = c_{t-1} + \alpha_{t-1}(x_t - c_{t-1})$

$\alpha_t = \alpha_{t-1} \times \beta$

else Record $c_{t-1}$ or $x_{t-1}$

$c_t = x_t$

$\alpha_t = \alpha_0$ (5)

다차원 변수의 압축 저장을 위해서는 먼저 함께 저장할 변수들을 선택한다. 함께 저장할 변수의 선택은 공정의 물질 및 에너지 수지식 등의 해석과 공정 데이터를 통하여 각 변수들간의 통계적인 상관관계 분석을 통하여 찾아낸다.

이렇게 선택된 변수들의 현재 데이터 벡터 $x_t$와 이전 시간까지의 중심값 $c_{t-1}$ 와의 유클리디안 거리 $d$를 식 (5)에 의하여 구한다. 이때, 일부 절대값이 큰 공정 변수에 의해 영향을 받는 것을 방지하기 위하여 각 변수의 정상상태에서의 표준편차를 사용하여 표준화를 시켜서, 해당 공정 변수의 변이 수준을 의미하는 표준 편차에 의해 모든 데이터의 변화가 동일한 수준의 영향을 미치도록 한다. 이렇게 구한 거리 $d$를 기록 한계값과 비교하여 기준 내에 들어 있는 경우에는 중심값 $c_{t-1}$ 를 갱신하고 기준 거리밖에 있는 경우에는 대표값을 기록한 후, 새롭게 중심값을 할당하여 앞서의 과정을 반복하게 된다.

## 5. 사례 연구

### 5.1. 1차원 공정 데이터

각 방법들의 압축 효율을 비교하기 위하여, 실공정 데이터에 대하여 본 연구에서 제안된 클러스터링을 이용한 압축 방법들(CC: CL, CT, CV, PV)과 기존의 Box Car(BC), Backward Slope(BS), 혼합방법(BCBS), 그리고 SDT사용한 압축 방법을 비교해 보았다.

300개의 데이터로 구성된 실제 공정의 데이터에 대하여 동일한 조건에서 각 방법들을 비교하여 보았다. 이때, 기록 한계값 $\sigma$로는 0.09부터 0.9까지 10개의 값을 사용하였다. 이때, 기록 한계가 0.45인 경우에 대하여 원래 공정의 데이터와 각 방법에 의하여 재생된 데이터를 Fig. 3에 나타내었다. 또한, 10단계에 걸친 기록 한계에 대하여 저장되는 데이터의 수 N과 재생 오차 SSE와의 관계를 Fig. 4에 나타내었다. 이때, 압축률을 최대한 높이면서도 재생 오차 SSE를 최소화 할 수 있는 방법이 가장 좋은 방법이 된다.

Fig. 4의 결과에서 보는 바와 같이 기존의 방법들은 대개 비슷한 결과를 나타내지만 그 중에서도 Box Car 알고리듬이 가장 무난하며, Backward Slope 알고리듬과 SDT 방법은 잡음의 영향으로 인하여 압축 효율이 낮거나 재현시의 재생 오차가 커짐을 알 수 있다. 클러스터링을 이용한 방법의 경우에는 직전값을 저장하는 경우(CC-PV)가 기존의 방법들과 비슷한 성능을 나타내며, 중심값으로 시간 구간 전체를 상수로 표현하는 경우(CC-CV)가 가장 뛰어난 성능을 보이는데, 이는 중심값이 주어진 구간에서의 SSE를 최소화시키는 대표값이기 때문이다. 중심값을 직선으로 연결하는 경우(CC-CL)에는 기록 한계값이 작은 경우에는 좋은 성능을 나타내지만, 기록 한계값이 커지는 경우에는 그 성능이 급격히 떨어진다. 이는 공정이 지속적으로 증가하거나 감소하는 경우에 있어서, 주어진 구간에서의 SSE를 최소화하는 중심값은 초기값과 종말값 사이의 중간값을 갖게 되어 실제 공정과 차이가 생겨나기 때문이다. 이 차이가 기록 한계가 클수록 커지기 때문에 급격한 성능의 저하를 가져오게 되는 것이다. 중심값을 주어진 구간의 중심 시간으로 저장하는 경우(CC-CT)에는 기록한계가 작은 경우에는 낮은 효율을 보이는데, 이는 저장 구간이 조밀한 경우 중심 시간의 기록에 있어서의 어려움 때문이다.

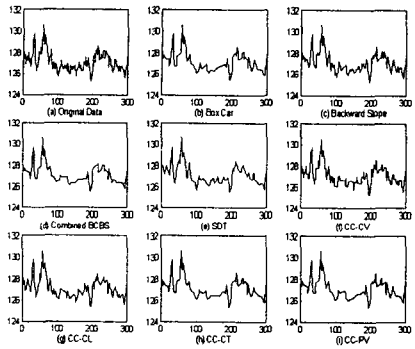

Fig. 3. Comparison of compression by various techniques.

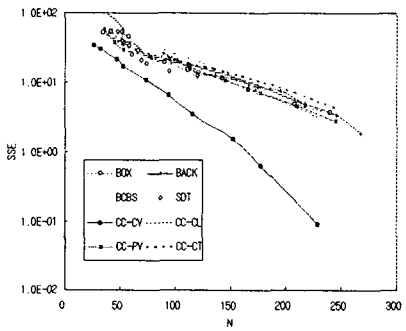


Fig. 4. SSEs vs. Compression Ratio.

### 5.1. 다차원 공정 데이터

앞서 제안된 방법을 보일러 공정의 실제 데이터[7,8]들에 적용하여 그 유용성을 확인하여 보았다. 본 연구에서 사용한 변수는 보일러의 급수량, 증기 생산량, 그리고 드럼 레벨 제어기의 제어 신호 출력의 3가지 변수였다. 이들 변수들은 물질 수지식과 연관이 되며 서로 상관관계가 매우 커서 공정을 감시할 경우에 동시에 고려해야 하는 변수이다.

거리의 계산에 필요한 각 변수의 표준편차의 값은 정상상태의 운전 데이터를 통계적으로 처리하여 얻었다. 300개의 데이터로 이루어진 세 변수에 대하여 각 변수의 표준편차를 기록한계로 사용하여 그 값을 저장하였다. 중심값을 저장하는 방법을 사용한 결과 300개의 데이터를 26개의 데이터로 압축할 수 있었다. 이때 시간과 3개의 변수에 대한 중심값을 저장하므로 실제 저장되는 데이터수는 26×4인 반면, 개개의 변수들을 시간과 변수값으로 저장하는 방법에서는 그보다 더 많은 26×2×3개의 데이터를 저장해야 함을 알 수 있다.

급수량과 스팀 생산량에 대하여 원래의 공정 데이터값과 저장된 데이터값을 Fig. 5로 나타내었다. Fig. 5에서 작은 원으로 표시된 것들이 공정 데이터이며 점으로 표시된 것들이 중심값이다. 저장해야할 데이터 수가 원래의 상관 관계를 유지하면서 상당한 정도로 줄어드는 것을 볼 수 있다.

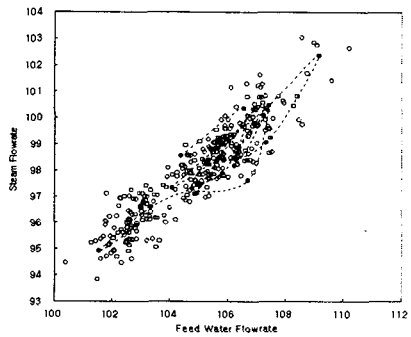


Fig. 5. Process data and recorded value.

압축된 데이터를 3개의 변수에 대하여 다시 재현한 결과를 Fig. 6로 나타내었는데, 각각의 공정 변수들의 경향을 효과적으로 재현함을 확인할 수 있었다. Fig. 6에서 가는 점선으로 표시된 것이 잡음이 포함된 공정 데이터이고 작은 원으로 표시된 점들이 저장된 중심값들이다. 그리고 저장된 중심값들 사이의 재현된 공정값은 선형으로 재현하는 것을 보여 주고 있다. 여기서도 Fig.5에서와 마찬가지로 공정 변수 상호간의 경향성이 압축과 재현 과정에서도 잃어버리지 않고 잘 표현됨을 볼 수 있다.

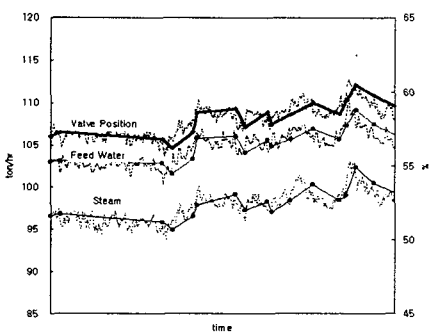


Fig. 6. Reconstruction of data from recorded values.

## 6. 결 론

본 연구에서는 기존에 제안된 데이터의 압축 저장 방법의 단점을 극복하기 위하여, $k$-means 클러스터링 기법을 이용한 압축 저장 방법을 제안하였다. 이를 실제 공정의 데이터에 적용하여 기존의 방법에 비하여 본 연구에서 제안된 해당 시간 구간에서 기록된 중심값을 상수로 재현하는 방법이 기존의 방법에 비하여 우수한 성능을 보임을 확인할 수 있었다. 이는 기존의 방법이 재현 오차에 대한 고려를 하지 않는데 비하여, 클러스터링을 이용한 방법은 주어진 구간에서 SSE를 최소로 하는 중심값을 사용하기 때문으로 생각된다. 또한, 제안된 클러스터링을 이용한 압축 방법을 다차원의 실공정 데이터에 적용하여 본 연구에서 제안된 방법이 상관관계가 깊은 다차원 데이터를 효과적으로 다룰 수 있음을 확인하였다.

## 감사의 글

이 논문은 교육부의 BK21사업과 과학기술부의 국가지정연구실 사업의 지원에 의해 연구되었으므로 이에 감사드립니다.

## 사 용 기 호

$c_t$ : 시간 t에서의 클러스터 중심값
$\boldsymbol{c}_t$ : 시간 t에서의 클러스터 중심값 벡터
E : 유클리디안 거리
$M_{a_i}$ : 클러스터링 분할 함수
N : 표본 벡터의 수
$x_t$ : 시간 t에서의 공정값
$\boldsymbol{x}_t$ : 시간 t에서의 공정값 벡터
$\overline{x_i}$ : 표본 벡터
$\overline{x_a}$ : 표본벡터에 가장 가까운 중심 벡터

## 그리이스 문자

$\alpha$ : 학습률
$\beta$ : 학습률 매개변수
$\eta$ : 학습률
$\sigma$ : 기록한계

## 참 고 문 헌

[1] Bakshi, B. R. and G. Stepanopoulos, "Compression of Chemical Process Data by functional Approximation and Feature Extraction", AIChE Journal, 42 (2), 477 (1996).
[2] Bristol, E. H., "Swing Door Trending : Adaptive Trend Recording", ISA Conf. Proc., 749 (1990).
[3] Hale, J. C. and H. L. Sellas, "Historical Data Recording for Process computers", Chem. Eng. Prog., 38, Nov. (1986).
[4] Gray, M. P., "Vector Quantization", IEEE ASSP MAGAZINE, April, 4 (1984).
[5] Kennedy, J. P., "Data Treatment and Application", Proc. Int. Conf. on Foundations of Computer Aided Process Operations, D. Rippin, J. Hale, and J. Davis, eds., CACHE, Austin, TX (1993).
[6] Moody, J. and C. J. Darken, "Fast Learning in Networks of Locall-Tuned Processing Unit", Neural Computation, 1, 281 (1989).
[7] 이기백, "이상-결과트리모델을 이용한 공정 이상 진단 시스템에 관한 연구", 공학박사학위논문, 서울대학교 화학공학과 (1997).
[8] Mo. K. J., Y. S. Oh, C. W. Jeong and En Sup Yoon, "Development of Operation Aided System for Chemical Process", Expert Systems with Applications, 12(4) (1997).

## 부 록

### A.1 곡선을 이용한 재현 방법에 대한 고찰

앞에서는 기존의 부분 선형 근사 방법이나 본 연구에서 제안된 클러스터링을 이용한 압축 방법의 경우에도 저장된 데이터를 직선이나 상수로 재현하였었다. 여기서는 직선 대신에 곡선을 사용하여 데이터를 재현하는 경우에 대하여 살펴보았다. 본 연구에서는 앞서 살펴본 가상 데이터에 대하여 MATLAB에서 제공하는 내삽(interpolation) 함수인 interp1을 사용하여 그 결과를 비교하여 보았다.

앞서 살펴본 가상 데이터에 대하여 기록 한

계값을 0.6부터 1.0까지 5단계로 변화시켜가면서 동일한 조건에서 직선(linear) 재현, 2차 곡선(spline) 재현 그리고 3차 곡선(cubic) 재현의 3가지를 수행하여 각각의 결과를 비교하여 보았다. 그 결과를 Fig. A-1과 Table A-1에 나타내었는데 2차 곡선으로 재현하는 경우에는 양 끝점에서의 문제로 인하여 오히려 재현 오차가 커지고, 3차 곡선으로 재현한 경우의 결과가 가장 좋음을 알 수 있다. 3차 곡선으로 재현한 경우가 직선으로 재현한 경우와의 재현 오차의 차이는 매우 미세하므로 월등하게 좋은 방법이라고 결론할 수는 없다. 또한 계산 부담을 고려한다면 실용적으로는 직선으로 재현하는 방법이 가장 유용한 방법이 될 것이다.

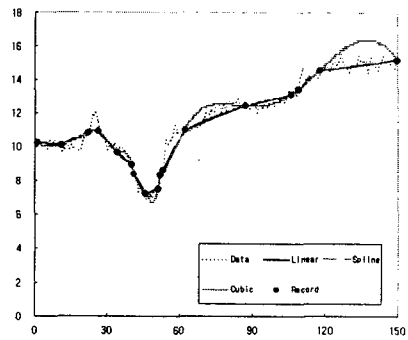


**Fig. A-1. Comparison of reproduction method(linear, spline, cubic)**

**Table A-1. Comparison of 3 reproduction methods(linear, spline, cubic)**

| σ | Square Sum of Error | | |
|---|---|---|---|
| | Linear | spline | cubic |
| 0.6 | 14.0 | 13.8 | 11.8 |
| 0.7 | 13.5 | 46.5 | 13.2 |
| 0.8 | 14.1 | 46.7 | 13.7 |
| 0.9 | 17.5 | 77.5 | 17.2 |
| 1.0 | 24.3 | 57.6 | 24.2 |