데이터 정제와 그래프 분석을 이용한 대용량 공정데이터 분석 방법

박 재 홍
삼성중공업 생산운영팀
변 재 현
경상대학교 산업시스템공학부, 공학연구원

An Analysis Method of Superlarge Manufacturing Process Data Using Data Cleaning and Graphical Analysis

Jai-Hong Park
Production Support Team, Samsung Heavy Industries
Jai-Hyun Byun
Dept. of Industrial and Systems Engineering, Engineering Research Institute, Gyeongsang National University

Key Words: Superlarge Process Data, Data Cleaning, Graphical Analysis, Six Sigma,

Abstract

Advances in computer and sensor technology have made it possible to obtain superlarge manufacturing process data in real time, letting us extract meaningful information from these superlarge data sets. We propose a systematic data analysis procedure which field engineers can apply easily to manufacture quality products. The procedure consists of data cleaning and data analysis stages. Data cleaning stage is to construct a database suitable for statistical analysis from the original superlarge manufacturing process data. In the data analysis stage, we suggest a graphical easy-to-implement approach to extract practical information from the cleaned database. This study will help manufacturing companies to achieve six sigma quality.

1. 서 론

최근 컴퓨터 및 센서 기술이 발달함에 따라 제조공정에서는 대용량의 공정데이터가 실시간으로 수집되고 있다. 여기에는 제품 제조의 전과정에 걸친 공정변수의 값뿐만 아니라 최종제품의 품질변수에 이르기까지 거의 모든 데이터가 포함되어 있다.

제조현장에서 공정관리 및 개선은 측정된 데이터에 근거하여 이루어지므로 정확한 데이터를 수집하여 분석하는 것이 중요하다. 정확하지 않은 측정시스템을 사용하게 되면 측정의 오류가 발생하게 되어 제품의 품질특성을 제대로 알 수 없고, 제조공정에서 발생하는 문제점을 정확히 파악해 낼 수 없다. 그러므로 측정시스템이 제품이나 공정을 정확히 측정하여 올바른 데이터를 산출할 수 있는지에 대한 평가가 반드시 이루어진 후, 데이터가 수집되어져야 한다.[배 도선 외, 1999]

최근 제조현장에서는 측정시스템으로부터 수집된 공정데이터가 대용량인 경우, 초기 공정데이터로부터 통계적 기법을 적용하여 원하는 정보를 얻기란 쉽지가 않다. 그 이유는 결측치(missing value)를 포함하고 있는 데이터가 많고, 제조공정의 불확실성과 잡음의 개입으로 인해 얻어진 데이터의 질이 낮아질 우려가 있으며, 무엇보다도 데이터의 양이 너무 방대하기 때문이다.[김 영상, 1999] 따라서 초기 대용량 공정데이터에 통계적 기법을 적용하기 전에 우선 데이터 정제(data cleaning)를 하여 분석할 수 있도록 만들어야 한다.[Banks와 Parmigiani, 1992] 그런 다음, 제품의 생산주기동안 대량으로 수집된 대용량 공정데이터를 이용하여 의미있는 품질정보를 추출하기 위한 체계적인 분석절차가 제시되어야 한다.

본 연구의 목적은 최근 제조공정으로부터 수집되어 저장된 대용량 공정데이터를 분석에 적합한 데이터베이스로 구축하기 위한 데이터 정제와 이러한 데이터베이스를 기반으로 대용량 공정데이터의 특성을 고려한 체계적 분석방법을 제시함으로써 최근 국내기업에서 활발하게 추진하고 있는 6 시그마 혁신활동의 데이터 분석방법에 도움을 주는 것이다.

2. 데이터 정제(Data Cleaning)

측정시스템으로부터 수집된 데이터가 대용량인 경우, 본격적인 통계분석을 하기 전에 데이터 포맷의 정돈, 결측치의 처리, 이상치(outlier) 제거 등 통계적 분석에 용이한 데이터로 전환시키는 것이 필요하다. 이와 관련하여 Banks와 Parmigiani(1992)는 대용량 공정데이터에 대해 데이터 정제를 위한 12단계의 절차를 제안하였다. 그런데 Banks와 Parmigiani가 제안한 9단계와 10단계는 중복되므로 하나의 단계로 통합하는 것이 좋고, 12단계는 데이터 정제보다는 분석에 해당하는 내용이므로 제외하는 것이 합당하다. 본 논문에서는 Banks와 Parmigiani가 제안한 12단계를 위와 같이 통합, 축소하여 10단계의 데이터 정제단계를 제시함으로써 대용량 공정데이터로부터 의미 있는 정보를 얻는데 도움을 주고자 한다.

단계1: 데이터 입력 확인

데이터 입력은 일관성을 유지하는 것이 필요하다. 수치 데이터의 경우는 실수형(floating point)으로 문자 데이터의 경우 그대로 입력처리를 한다. 이 단계는 데이터 정제의 첫 단계로서 다음 단계에 큰 영향을 미치므로 이 단계의 오류를 제거하기 위하여 주의를 기울여야 한다.

단계2: 시간척도에 따른 데이터 수집 확인

데이터 수집 시 데이터 기록의 식별기준인 시간에 따라 각 변수들의 데이터가 수집되었

는지 확인하는 단계이다. 만약 어떤 변수의 데이터가 동일한 시간에 측정되지 않아 결측치가 발생했다면, 그 변수의 결측치는 정상조건에서 결여되었다는 표시를 하여야 한다.

단계3: 결측치 표시

결측치가 발생하면, 발생원인을 파악하여 모든 결측치에 대해 코드 값을 부여한다. 결측치가 발생하는 것은 측정단위의 차이나 센서의 오 작동에 기인하는 경우가 많다. 측정단위의 차이에 의한 결여란 어떤 데이터는 1시간 단위로 측정되고, 다른 데이터는 30분 단위로 측정되어 시간지표(time index)의 불일치가 발생하여 데이터가 결여되는 경우를 말한다. 발생원인별로 특정한 코드 값을 표시하는 방법은 단계 7의 결측치 도표에 예시하였다.

단계4: 데이터 크기(sample size) 검사

각 변수에 대한 데이터 개수를 확인하는 단계이다. 입력은 예상되는 시계열의 길이와 공정데이터의 개수가 되며, 출력은 각 변수들의 실제 데이터 개수와 예상 데이터 개수에 차이가 있는 변수들의 목록이다.

단계5: 불가능한 값 처리

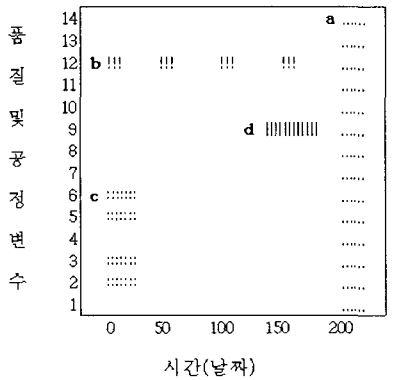
불가능한 값(impossible value)이란 일반적으로 발생할 수 없는 값을 말하며, 이상치와는 다르다. 이러한 값들은 주로 부호(+, -)가 바뀌거나, '0'이 추가되거나 빠지지거나, 데이터 입력 시 입력위치가 잘못되어 나타난다. 불가능한 값의 여부는 반드시 공정 전문가와 협의하여 결정하여야 한다.

단계6: 동기화(synchronization)

데이터가 동기화 되도록 색인(index)을 재조정하는 단계이다. 제품의 공정변수가 품질변수에 영향을 주는 시기를 재조정하여 적합한 분석이 되도록 한다. 예를 들어, 공정변수 값이 바뀌게 될 때 그러한 변화가 10시간 이후에 품질변수에 영향을 줄 때는 공정변수의 변화시기를 10시간 이후로 조정해야 한다.

단계7: 결측치 도표 작성

결측치 도표란 각 변수들의 결측치를 시간(날짜)에 따라 도표로 작성하는 것이다. 공정 엔지니어는 결측치 도표로부터 공장의 조업중단, 하위시스템의 고장 등 결측치 발생원인에 대한 정보를 얻어 결측치의 원인을 제거할 수 있어 데이터 정제 작업을 쉽게 할 수 있도록 도움을 준다. <그림 1>은 결측치 도표의 작성 예를 나타내는데, 결측치의 성향을 기호로 표시하여 나타내면 유용하다. 예를 들어, <그림 1>에서 a:계획적인 공장의 조업중단, b:알 수 없는 원인, c:하위시스템의 고장, d:돌발적인 상황이 발생했는데 발생원인을 알고 있는 경우이다. 이러한 도표를 이용하면 결측치 패턴을 분석하여 공정 모니터링 시스템 개선에 유용하게 쓸 수 있다.



<그림 1> 결측치 도표 작성예제

단계8: 관측도수 차이와 결측치 추정

입력

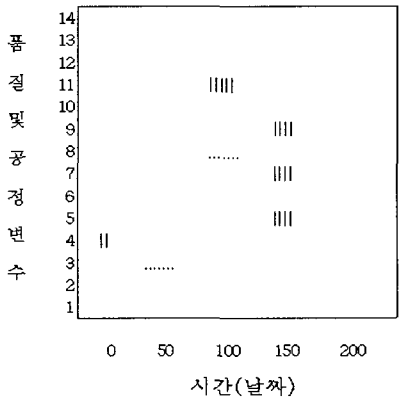
생산공정에서 모든 공정변수가 동일한 빈도로 측정되지 않거나, 센서의 고장 또는 다른 이유로 인하여 일련의 데이터에서 간단한 gap이 생긴다. 이러한 gap을 통계적 방법과 전문가 지식 등을 이용하여 채우는 단계이다. 결측치를 대체하는 방법으로는 평균, 중앙값, 군집평균, 회귀분석 등에 의한 데이터 추정방법이 있다.[Pyle, 1999] Banks와 Parmigiani(1992)는 선형 보간법(linear interpolation)에 의한 데이터 추정방법을 권장한다. 그 이유는 프로그램밍 하기가 쉽고 대부분의 소프트웨어에서 수행이 빠르기 때문이다.

단계9: 극한치 도표 작성

극한치 도표는 공정변수의 데이터 중 $\pm 3\sigma$를 벗어난 데이터를 표시하여 이상치를 탐지하는 데에 도움을 주는 도표이다. <그림 2>에 극한치 도표의 작성 예를 나타내었는데, 기호[|]는 $+3\sigma$ 위로 벗어난 데이터, 기호[.]는 -3σ 아래로 벗어난 데이터를 표시한 것이다. 발견된 이상치는 그 원인을 파악한 후 처리하고 재발하지 않도록 수정조치를 취해야 한다.

단계10. 기술통계량과 기초 탐색

시계열 데이터의 안정성, 공장중단이나 모니터링 장치에 고장이 발생한 전후의 데이터가 일관성이 있는지를 판단하기 위하여 기초적 시계열분석, Q-Q plot 등의 도표분석을 한다. 분석대상인 특정 공정의 데이터로부터 의미 있는 정보를 추출하기 위하여 어떠한 분석 기법을 쓸 것인지를 최종적으로 점검을 하는 단계이다.



<그림 2> 극한치 도표 예

3. 대용량 공정데이터 분석

정제단계를 거친 데이터로부터 유용한 정보를 얻기 위해서는 우선 대용량 공정 데이터의 보편적인 특징을 알아야 한다. 대용량 공정데이터의 특징은 크게 4가지로 볼 수 있다. 첫째, 다단계(multistage)의 공정을 거치면서 수집된 데이터이다. 둘째, 다량의 공정변수가 관계되어 있다. 셋째, 품질변수간, 공정변수간, 품질변수 및 공정변수간 상관관계가 존재할 가능성이 크다. 넷째, 체계적인 조업실험 또는 공정제어를 통하여 얻을 수도 있지만, 공정변수들이 일정범위 내에서 변동할 때 수동적으로 얻는 경우가 많다.

제조공정에서 최종제품이 생산되기까지는 여러 단계의 공정을 거치므로, 최종제품의 품질특성 변동은 모든 공정단계의 변동으로부터 영향을 받는다. 따라서 최종제품의 품질향상을 위해서는 최종단계 뿐만 아니라, 이전 단계의 공정변수들 중 중요한 변수들을 파악하여 적절한 제어를 해야만 품질특성의 변동을 감소시킬 수 있다. 일반적으로 대용

량 공정데이터에는 많은 품질변수 및 공정변수가 포함되어 있으므로 일부 변수들은 크게 연관되어 있을 가능성이 있다. 공정 데이터 분석의 궁극적인 목적은 품질변수에 크게 영향을 미치는 공정변수를 찾아내어 품질향상을 위하여 중요한 공정변수가 가져야 할 값 또는 제어범위를 결정하는 것이다. 본 논문에서는 현장 엔지니어들이 쉽게 이해하고 이용할 수 있는 공정데이터 분석방법인 Covariation Chart, Sliced Inverse Box Plot(SIB) Chart, Brushing Scatter Plot, Box Plot, Multi-vari Chart를 이용한 그래프 분석 기법을 제시하고자 한다.

3.1 Covariation Chart

Covariation chart는 Banks와 Parmigiani (1992)에 의해 개발된 도표로 주어진 공정데이터에 대한 품질변수간의 상관관계, 공정변수간의 상관관계, 품질변수와 공정변수 간 상관관계를 제시한다. 이 도표의 특징은 품질변수 값을 상/하위 10%로 나누어, 이러한 구간에서 품질변수와 공정변수의 상관관계를 볼 수 있기 때문에 공정 개선을 위한 중요한 정보를 얻는 데에 도움이 된다.

3.1.1 Covariation Chart 작성방법

(1) 품질변수와 공정변수의 선택

(2) 4 가지의 covariation chart를 작성하여 변수간 상관관계(상관계수)에 따라 백분율로 분류하여 기호로 표시[Banks와 Parmigiani, 1992]

-4: 90%이상

-3: 20%이상 ~ 90%미만

-2: 0.5%이상 ~ 20%미만

-1: 0.1%이상 ~ 0.5%미만

-0: 0.1%미만

위의 상관계수 분류기준은 데이터 수에 따라 달라질 수 있다. t-검정[식 (1)]을 이용한 상관계수의 유의성 검정을 위하여 유의수준 α=0.01로써 귀무가설(ρ=0)을 기각할 수 있는 최저 상관계수를 <표 1>에 나타내었다. 일반적으로 상관관계가 3, 4를 가질 때 중요하다고 판단한다. 하지만 데이터 수가 아주 클 때는 상관관계 2도 중요하다고 결정할 수 있다.

$$\left| t_0 = r\sqrt{(n-2)/(1-r^2)} \right| > t(n-2, \alpha/2) \qquad (1)$$

3.1.2 Covariation Chart 작성 예제

극소탄소강 철강제품의 경우를 예를 들어 covariation chart를 작성하는 방법을 예시하고자 한다.[박재홍 등, 2001] <표 2>와 <표 3>에 나타나 있는 4개의 품질변수와 10개의 공정변수를 이용하여 냉연공장 전체를 대상으로 한 5755개의 데이터를 이용하여 <그림 3>과 같은 covariation chart를 작성하였다. 상관계수 분류기준은 <표 1>

<표 1> 데이터 개수의 따른 최저상관계수

데이터 수	최저 상관계수	데이터 수	최저 상관계수
50	0.3	2000	0.06
100	0.2	3000	0.05
200	0.17	4000~6000	0.04
300	0.14	7000~10000	0.03
400	0.12	20000~50000	0.02
500	0.1	60000	0.01
600~800	0.09	70000~80000	0.009
900~1000	0.08	90000~100000	0.008

의 최저 상관계수와 상관계수 값의 상대적인 크기에 따라 상관관계를 0~4로 분류하였다.

- 4: 90%이상
- 3: 20%이상 ~ 90%미만
- 2: 10%이상 ~ 20%미만
- 1: 4%이상 ~ 10%미만
- 0: 4%미만

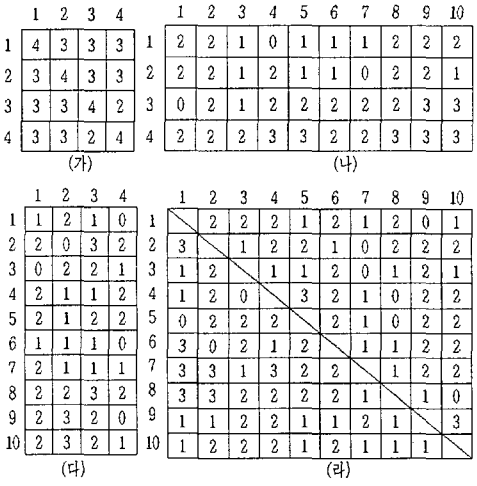
<표 2> 극저탄소강의 품질변수

번호	품질변수
1	TS(Tensile Strength, 인장강도)
2	YP(Yield Point, 항복강도)
3	EL(Elongation, 연신율)
4	HRB(경도)

<표 3> 극저탄소강의 공정변수

번호	공정변수
1	C(탄소)
2	S(황)
3	TI(티타늄)
4	YEH(슬라브 추출온도-예열대)
5	GAH(슬라브 추출온도-가열대)
6	TOT(총 체로시간)
7	CTB(권취온도-BOT부)
8	CTHK(냉연두께)
9	CWTH(냉연코일 폭)
10	WGT(중량)

<그림 3>에서 (가)는 품질변수들 간 상관관계를 나타내고, (나)는 각 품질변수별로 하위 10%에 해당하는 데이터에 대하여 품질변수와 공정변수간 상관관계를 나타낸 것이며, 상위 10% 데이터에 대한 분석은 (다)에 표시되어 있다. (라)는 중요한 품질변수인 YP(항복강도)에 대하여 대각선 위쪽은 품질변수의 데이터 중 하위 10%, 대각선 아래쪽은 상위 10% 데이터에 해당되는 부분에서 공정변수간 상관관계를 나타낸 것이다.



<그림 3> covariation chart 작성 예제

<그림 3>의 결과를 분석해 보면 품질변수간에는 모두 상관관계가 있고 특히 TS와 YP는 다른 품질변수들과 큰 상관관계를 가지고 있음을 (가)를 통해서 알 수 있다. (나)를 보면, HRB가 낮은 값을 가지는 데이터에서는 공정변수 4, 5, 8, 9, 10이 HRB와 높은 상관관계가 있고, 공정변수 9, 10은 또한 EL과도 상관관계가 높다. (다)를 살펴보면, YP의 값이 높을 때 9, 10 공정변수가 상관관계가 크고, EL의 값이 높을 때에는 공정변수 2와 8이 높은 상관관계가 있다. 동일한 방법으로 품질변수 YP에 대한 공정변수간 상관관계도 (라)의 결과를 통하여 해석할 수 있다. 이와 같이 covariation chart는 여러 변수들의 상관관계를 도표로 나타내어 품질변수 값의 특성에 따른 품질변수와 공정변수간 상

관관계와 공정변수간 상관관계를 분석하여 의미 있는 변수를 선택하는 데에 도움을 준다.

3.2 Sliced Inverse Box Plot (SIB) Chart

SIB chart는 X축에 공정변수, Y축에 품질변수를 두는 box plot과는 달리 Y축에 공정변수 값을 놓고, X축에는 품질변수 값으로 둔다. 이렇게 변수들을 두는 목적은 품질변수 값의 변화에 따른 공정변수 분포의 변화를 보기 위함이다[Banks와 Parmigiani, 1992]. 일반적으로 제품의 생산공정에는 다수의 품질변수들이 있고 각 품질변수가 바람직한 범위의 값을 갖는 제품을 생산하기 위해서는 관계되는 중요한 공정변수를 최적의 값 또는 구간으로 제어해야 한다. 그런데 품질변수가 여러 개 있을 경우에는 공정변수의 최적 값을 결정하는 것이 쉽지 않다. 해당 공정변수의 바람직한 값이 품질변수 별로 달라지는 상충(conflict)이 생기기 때문이다. 본 논문은 박 재홍 등(2001)에서 다룬 사례에서 실제 품질변수 3개와 공정변수 3개를 대상으로 SIB chart를 이용, 각 품질변수가 좋게 나타나는 하위 %별로 구간을 나누고, 각 구간별 공정변수별 분포를 분석한다.

분석대상으로 선택된 품질변수 및 공정변수는 <표 4>와 <표 5>에 나타내었다. 그리고 <표 6>은 각 품질변수별로 규격 안에 들어오는 데이터 중 하위 percentile을 기준으로 하여 품질변수의 값을 구간으로 나타낸 것이다.

<표 4> 품질변수 목록

번호	품질변수	최대 기준치
1	TS(인장강도)	27kg/mm^{2}이상
2	YP(항복강도)	17.5kg/mm^{2}이하
3	EL(연신율)	37% 이상

<표 5> 공정변수 목록

번호	공정변수
1	C(탄소)
2	S(황)
3	TI(티타늄)

<표 6> 3개의 품질변수별 구간설정

구간번호	percentile(%)	TS	YP	EL
1	10%이하	27~27.6	17~17.5	44~45.1
2	10%~30%	27.7~28.8	16~16.9	45.2~47.3
3	30%~50%	28.9~30.5	15~15.9	47.4~49.5
4	50%이상	30.1이상	14.9이하	49.6이상

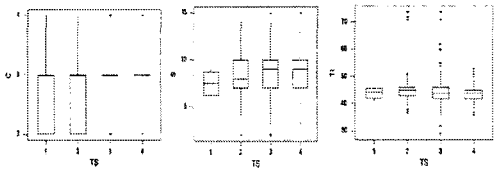
가. 품질변수별 하위%에 따른 구간별 공정변수의 분포분석

<그림 4>~<그림 6>은 각 품질변수의 구간에 속하는 공정변수별 분포를 나타낸 것이다. 공정변수 S, TI의 경우 모든 구간에서 거의 동일한 분포가 나타나지만, 공정변수 C는 구간별로 뚜렷하게 다른 분포를 보이고 있음을 알 수 있다. 각 품질변수별로 공정변수 C의 구간별 분포를 <그림 7>에 나타내었다.

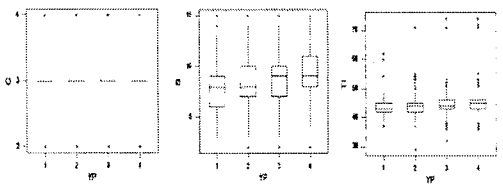
나. 도표를 이용한 공정변수 분포 분석

도표를 이용하여 공정변수의 바람직한 방향을 표시할 때, 공정변수 값의 중심(-)을 기준으로 낮은 쪽이 좋은 경우는 ↓, 높은 쪽인 경우에는 ↑로 표현한다. 예를 들어, C는 2~4 의 정수값을 가지므로 중심값 3을 기준으로 표현한다. <표 7>은 하위0%이하(1구간)에서 각 품질변수에 대하여 공정변수별로 바람직한 방향을 나타낸 것이다. 이와 같은 도포를 이용하면 품질변수가 여러 개 있을 경우에 각 공정변수의 바람직한 값이 품질변수 별로 달라지는 상충현상을 확인하여, 종합적으로 공정변수의 바람직한 값 또는 제어범위를 결정할 수 있다.

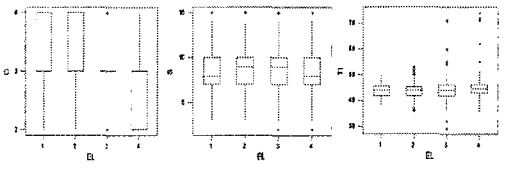
<표 7>을 보면 공정변수 C의 값이 품질변수별로 달라지는 상충이 있기 때문에 C의 바람직한 값을 결정하기 위한 추가적인 분석이 필요함을 알 수 있다.



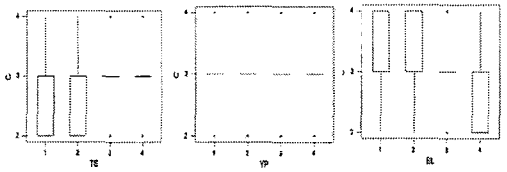
<그림 4> TS의 구간별 각 공정변수의 분포 그림



<그림 5> YP의 구간별 각 공정변수의 분포 그림



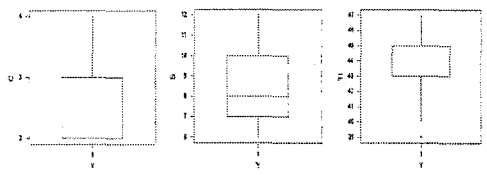
<그림 6> EL의 구간별 각 공정변수의 분포 그림



<그림 7> 품질변수의 구간별 C의 분포

<표 7> 하위 10% 이하에 대한 각 품질변수별 공정변수의 방향

공정변수 \ 품질변수	TS	YP	EL
C	-	↑	↓
S	↓	↓	↓
TI	↓	↓	↓



<그림 8> 공통영역에 해당하는 공정변수별 분포 그림

다. 상충 발생 시 공정변수의 분포분석

공정변수의 바람직한 값이 품질변수별로 달라지는 상충이 생길 때, 품질변수를 모두 만족하는 공정변수의 바람직한 값을 구하기 위해 각 품질변수가 상위 50%이상(하위 50% 이하)에 해당하는 영역의 데이터를 추출하여, SIB chart를 이용하여 공정변수의 분포를 분석한다. 본 논문에서 다루는 사례의 경우, 모든 품질변수의 상위 50%이상의 데이터로 구성된 영역(이하 공통영역)에 해당하는 데이터를 추출, SIB chart를 이용하여 3개 공정변수의 분포를 분석한다. <그림 8>에 SIB chart를 이용한 공통영역에 해당하는 공정변수별 분포를 나타내었다.

3.3 Brushing Scatter Plot

Brushing이란 다차원 데이터(multidimensional data) 일부에 대하여 변수간 상관관계를 분석하기 위한 것으로, 어떤 특정 영역의 데이터를 대상으로 변수들간 상관성을 산점도 행렬(scatter plot matrix)로 나타내는 방법이다.[Becker와 Cleveland, 1987]

brushing에는 4가지의 작업(highlight, shadow highlight, delete, label)이 있다. 그리고 각 작업에서 3가지의 paint mode (transient, lasting, undo) 중 하나를 수행할 수 있다. 본 논문에서는 Minitab에서 이용가능한 highlight의 transient mode에 대해 설명하고자 한다. highlight의 transient mode란 분석하고자 하는 측정점이나 측정점들의 영역을 지정하고(highlight), 그 영역을 이동해 가며(transient), 각 영역별 변수간 상관관계를 그래프로 확인하는 방법이다.

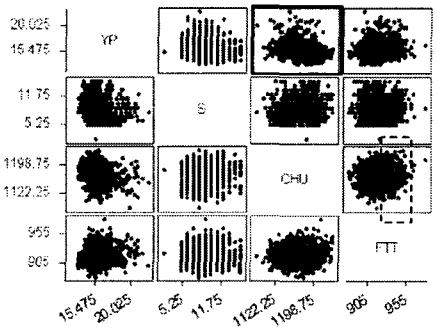
가. 이상치(outlier) 파악

<그림 9>는 박 재홍 등(2001)에서 다룬 사례 중 한 냉연공장의 데이터를 대상으로 YP(항복강도)와 CHU(추출온도)에 관한 2차원 산점도에서 이상치를 highlight하여 이러한 이상치를 발생하게 하는 주요 원인이 되는 공정변수를 파악하는 그림이다. 이 그림으로부터 CHU외에 FFT(열연온도, TOP부)가 주요원인임을 알 수 있다. 표를 보면 YP의 이상치를 발생시키는 FTT의 범위는 높은 구간(920~950 °C)임을 알 수 있다.

대용량 데이터의 경우 데이터의 양이 방대하기 때문에 품질변수의 이상치에 대한 정보를 각 공정변수들로부터 얻기가 쉽지 않다. 따라서 산점도 행렬을 brushing하면 즉시 공정변수들의 데이터에 대한 정보가 나타나기 때문에 쉽게 이상치에 대한 정보를 얻을 수 있다.

나. highlight의 transient mode 적용

산점도 행렬에서 자신이 분석하고자 하는 속성에 따라 데이터를 구간으로 분류하고, 각 구간에 있는 데이터를 highlight하여 transient mode를 적용하면 각 구간별로 공정변수의 값을 어떻게 가지고 가야 하는 지를 알 수 있다. <그림 10>에서 점선으로 표시된 부분은 highlight 된 영역이고 실선은 분석을 위해 표시한 부분이다. 그림을 보면 YP가 하한구간이나 중간구간에 있을 때, TS와 EL이 음의 상관관계가 있음을 알 수 있다.

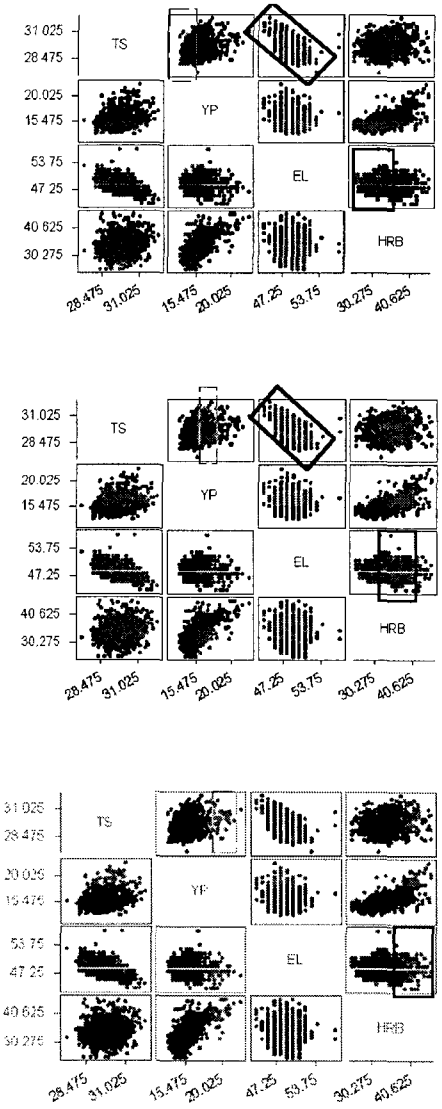


Row	S	CHU	FTT
82	11	1120	906
126	6	1112	944
168	6	1130	910
209	7	1171	930
233	7	1130	925
346	11	1133	898
401	9	1201	946
414	9	1169	938
422	7	1128	932
450	7	1216	937
494	2	1184	921
502	7	1133	928
551	7	1121	904
552	6	1161	921
601	8	1084	917
602	8	1118	913
603	9	1207	927
604	9	1194	931
929	9	1129	954
940	12	1150	920
998	7	1128	919
1000	8	1150	900
1004	8	1084	917
1005	8	1118	913
1011	9	1207	927
1012	9	1194	931
1034	8	1125	924
1309	10	1181	903

<그림 9> 이상치의 원인파악을 위한 예

3.4 Box Plot

대용량 공정데이터는 체계적인 조업실험 또는 공정제어를 통하여 얻을 수도 있지만, 공정변수들이 일정범위 내에서 변동할 때 수동적으로 얻어지는 데이터인 경우에는 현 공정조건을 그대로 반영하고 있다. 이 때에는 현재 공정조건의 공정능력을 시그마 수준으로 환산함으로써 현 공정이 원하는 품질수준을 가진 제품을 어느 정도 생산하고 있는지 알 수 있다. 만약 원하는 품질의 제품이 제대로 생산되고 있지 않다면, 공정조건의 제어를 통한 최적 공정조건을 찾을 수 있는 분석방법이 필요하다. 최적 공정조건을 찾기 위한 분석방법으로 현장 엔지니어들이 쉽게 활용할 수 있는 box plot을 이용하는 방법을 박 재홍 등(2001)에 소개된 사례를 중심으로 설명하기로 한다. 극저탄소강의 재질 중에 개선이 필요한 YP(항복강도)를 주요 품질특성으로 선정하여 그 수준을 하향 안정화시키



<그림 10> 구간별 transient mode 예

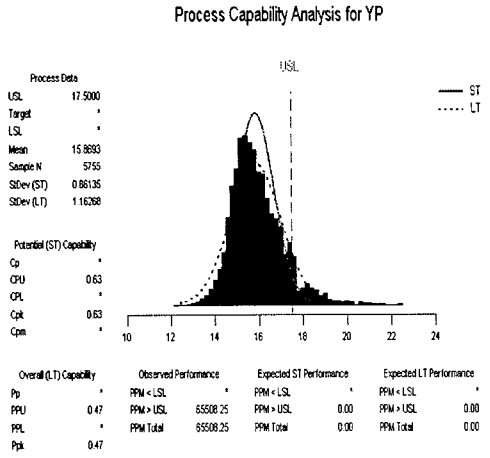
기 위해 주요 공정변수의 제어범위를 정하는 방법을 예시하기로 한다.

가. 전체공정의 공정능력 분석

<그림 11>은 박 재홍 등(2001)의 사례에서 다룬 전체공정의 YP에 대한 공정능력분석 결과이다. 분석결과를 보면 Cpk = 0.63이고, Ppk = 0.47로서, 시그마수준은 2.91σ수준[(Ppk×3)+1.5]이다.

나. 최적구간의 도출

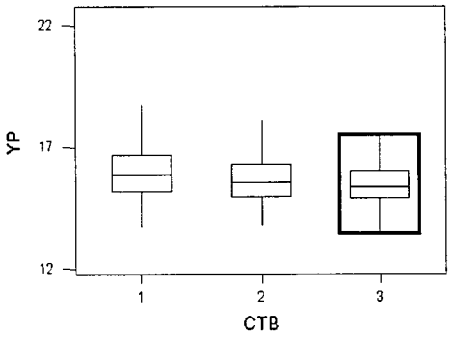
공정변수의 제어범위를 3개의 구간(1: 하한구간, 2:중간구간, 3: 상한구간)으로 분류하고 구간별로 품질변수의 box plot을 도시하여 YP의 값을 가장 크게 줄일 수 있는 구간을 최적구간으로 선정한다. <그림 12>에 냉연공정의 공정변수인 권취온도(CTB)를 예로 들었다. YP를 적정 제어범위인 17.5kg/mm^2 이하로 두기 위해서는 권취온도를 '상한구간'으로 제어해야 됨을 알 수 있다. 다른 주요 공정변수들의 최적구간도 같은 방식으로 구하여, 전체적인 최적 공정조건을 구할 수 있다.



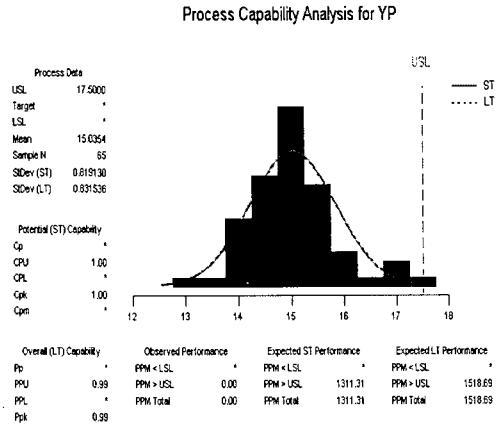
<그림 11> YP의 공정능력분석

다. 최적구간 검증

최적구간의 타당성 검증을 위하여 전체 데이터 중에서 각 주요 공정변수들의 최적구간으로 구성된 데이터를 추출하여 YP에 대한 공정능력을 분석하여, 현 공정에 대한 시그마 수준과 최적구간에서의 시그마 수준을 평가해 보았다. <그림 13>에 최적구간의 공정능력분석 결과를 나타내었는데, 시그마 수준이 2.91σ에서 4.47σ로 향상되었음을 알 수 있다.



<그림 12> CTB의 최적구간 도출예제



<그림 13> 최적구간의 YP 공정능력

3.5 Multi-Vari Chart

Multi-vari chart를 이용한 다변량 분석은 공정변수들이 품질변수에 미치는 효과를 공정변수간 교호작용을 고려하여 도식적으로 나타냄으로써 품질변수에 영향을 미치는 각 공정변수에 대한 품질변수의 변동경향을 분석할 수 있는 방법이다.[De Mast 등, 2001] Multi-vari chart는 공정변수의 데이터가 일정 범위 내의 정수값을 가질 때 품질변수의 변동분석이 용이하다. 예를 들면, 탄소처럼 2~4의 범위 내 정수 값[박재홍 등, 2001]을 가지면 각 값에 대한 품질변수의 변동분석이 용이하다. 하지만 제조현장에서 제어되는 다수의 공정변수는 넓은 범위의 정수값 또는 실수값을 가지는 경우가 많기 때문에 multi-vari chart를 통해 모든 값에 대한 품질변수의 변동분석이 어렵고, 분석을 통해 품질변수의 변동에 영향도가 큰 공정변수의 값을 구한다 하더라도 현장에서 넓은 범위의 값을 가질 수 있는 공정변수를 특정한 값으로 제어하기란 쉽지 않다. 제조현장에서는 세밀한 제어가 힘든 공정변수를 구간으로 제어하는 점을 중시하여, 공정변수의 값을 3개의 구간(1:하한구간, 2:중간구간, 3:상한구간)으로 분류한 다음, multi-vari chart에 이용하고자 한다.

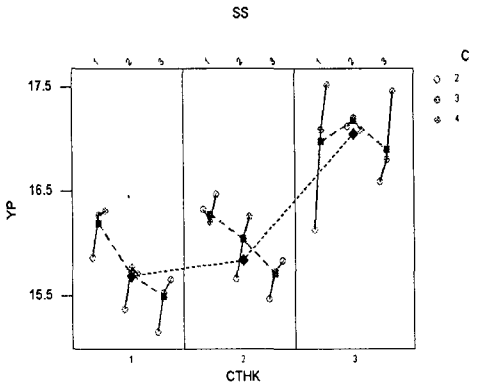
가. Minitab을 이용한 Multi-Vari Chart

박재홍 등(2001)에서 다룬 사례 중 전체공정 데이터를 대상으로 품질변수와 공정변수를 이용하여 multi-vari chart를 이용한 분석 예를 보이고자 한다. <표 8>은 multi-vari chart의 예제에 이용된 품질변수 및 공정변수를 나타낸 도표이고, <그림 14>는 multi-vari chart를 이용한 각 공정변수에 대한 YP의 변동을 분석한 예이다. YP의 적정 제어범위는 17.5 kg/mm^{2}이하이다.

<그림 14>를 통해 각 공정변수의 변화에 대한 품질변수의 변동을 분석해 보면, 우선 YP에 가장 큰 영향을 미치는 공정변수는 제어가 불가능한 CTHK임을 알 수 있다. CTHK가 1 또는 2구간에서는 SS는 3구간, C는 2로 제어하는 것이 좋고, CTHK가 3구간에서는 SS는 1구간, C는 2가 최적 제어범위이다. CTHK 3구간에서 만일 C를 2로 제어할 수 없을 때에는 그림에서 알 수 있듯이 SS를 2구간으로 제어함으로써 YP값의 산포를 줄일 수 있다.

<표 8> multi-vari chart 예제를 위한 품질변수 및 공정변수

변수 구분	변수명	값의 범위	제어가능성 여부
품질변수	YP(항복강도)	17.5 이하	
공정변수	C(탄소)	2, 3, 4	제어가능
공정변수	CTHK (열연두께)	1구간: 0.51 - 0.9 2구간: 0.91 - 1.2 3구간: 1.21 - 1.7	제어불가능
공정변수	SS (소둔온도)	1구간: 821 - 830 2구간: 831 - 840 3구간: 841 - 850	제어가능



<그림 14> multi-vari chart 예제

3.6 그래프 분석기법의 사용 지침

본 논문에 제시된 그래픽 분석기법들은 앞에서 언급된 대용량 공정데이터 특성으로 인해 회귀분석 등 일반적인 통계적 기법으로는 분석하기 어려운 정보를 얻는데 이용할 수 있다. 대용량 데이터를 분석하기 위해 본 논문에서 제시된 그래픽 분석기법들을 어떻게 적용할 것인지 살펴보기로 한다.

다량의 품질변수 및 공정변수가 포함되어 있는 대용량 공정데이터를 수집하였을 때, 우선 공정능력 분석을 통하여 각 품질변수가 원하는 품질수준(시그마 수준)을 가지고 있는지에 대한 평가가 필요하다. 만약 어떤 품질변수들이 원하는 품질수준을 가지고 있지 않다면, 다음과 같은 간단한 분석절차를 적용할 수 있다. 첫 번째로 covariation chart를 이용한 상관관계 분석을 통해 각 품질변수에 영향도가 큰 공정변수를 도출하는 것이 필요하다. 두 번째로 SIB chart와 box plot를 이용하여 covaria- tion chart에 의해 도출된 품질변수와 주요 공정변수간 분포 분석을 통해 품질변수의 값을 제고할 수 있는 주요 공정변수의 최적 제어구간을 도출한다. 그 다음 주요 공정변수들의 최적 제어구간으로 구성된 공통 영역에 있는 데이터를 추출하여 품질변수에 대한 공정능력을 분석하여 공정능력이 원하는 만큼 개선이 되었는지 평가한다. 여기서 제품과 관련된 모든 품질변수가 원하는 품질수준으로 향상되었다면 분석을 종료한다.

공정변수간 상관관계가 큰 경우에는 multi-vari chart 분석을 통해 품질변수에 영향을 미치고 있는 주요 공정변수들의 상호연관성을 고려하여 품질변수의 특성을 만족시키기 위한 주요 공정변수의 최적 제어구간을 선택하여 품질개선을 도모한다.

brushing scatter plot은 위의 분석방법들에 의해 도출된 최적구간이 원하는 품질수준의 결과를 얻지 못할 경우 이용해 볼 수 있는 분석방법이다. 우선 산점도 행렬 상에 brushing 분석방법을 이용하여 이상치를 발견하여 제거하고 주요 품질변수의 다양한 특정영역에서 다른 변수와의 상관관계의 파악을 통해 상관관계가 있는 변수들의 제어조건을 변화시켜 위에서 제시된 분석절차를 순차적으로 적용함으로써 품질수준을 제고할 수 있을 것이다.

4. 결론 및 추후 연구과제

본 논문은 제조공정에서 발생하는 대용량 공정데이터를 이용하여 의미 있는 품질정보를 추출하기 위해 필요한 분석절차를 제시하였다. 우선 통계적 분석에 적합한 데이터베이스를 구축하기 위한 데이터 정제 방법을 제안하였다. 그리고 정제단계를 거친 데이터베이스로부터 유용한 정보를 얻기 위해 대용량 공정데이터의 보편적인 특징을 고려하면서 현장의 엔지니어들이 쉽게 이용할 수 있는 도식적인 데이터 분석방법을 제시함으로써 대용량의 공정데이터를 다루고 있는 국내 기업이 품질수준을 향상하는 데에 도움을 주고자 하였다.

대용량 공정데이터는 제조공정별로 그 특성이 다를 수 있으므로 분석대상에 따라 본 논문에서 제시하는 방법을 수정하여 사용하거나 적합한 분석방법을 추가하여 이용할 수 있으리라고 본다.

참고문헌

[1] 김 영상(1999), "공정모니터링 데이터 분석을 위한 편차최소제곱법과 인공신경망의 비교 연구", 한국과학기술원 산업공학과 석사학위논문.

[2] 박 성현(1998), 「회귀분석」, 민영사.

[3] 박 재홍, 변 재현, 김 창현, 정 창원, 최 영대(2001), "구간세분화 방법을 이용한 철강산업체의 6시그마 프로젝트 추진사례", 「품질혁신」, 제2권, 제1호, pp. 57-66.

[4] 배 도선 외 6인(1999), 「통계적 품질관리」, 영지문화사.

[5] Banks, D. L., Parmigiani, G.(1992), "Pre-Analysis of Superlarge Industrial Data Sets", *Journal of Quality Technology*, Vol.24, pp.115-129.

[6] Becker, R. A., Cleveland, W. S.(1987), "Brushing Scatterplots", *Technometrics*, Vol.29, pp.115-129.

[7] De Mast, J., Rose, K. C. B., Does, R. J. M. M.(2001), "The Multi-Vari Chart: A Systematic Approach", *Quality Engineering*, Vol.13, pp.437-447.

[8] MINITAB(2000), *Minitab Statistical Software: User's Guide* , MINITAB Inc., release 13.

[9] Pyle, D.(1999), *Data Preparation for Data Mining*, Morgan Kaufmann Publishers.