탐색적 데이타분석(EDA) 기능에 관한 통계팩키지 프로그램의 비교검토

허명회* 정진환**

〈요 약〉

본 소고에서는 탐색적 데이타분석(EDA)의 방법들이 국내에서 비교적 쉽게 구할 수 있는 통계 팩키지 프로그램에 어떻게 수용되어 있는지를 비교검토할 것이다. 여기에서 고려된 통계 팩키지 프로그램은 IBM-PC의 XT/AT 기종에서 사용가능한 Minitab, NCSS, SAS, SPSS/PC+, Statgraphics, SYSTAT의 모두 6종이다.

1. 서론 : 탐색적 데이타분석(EDA)과 팩키지 프로그램

탐색적 데이타분석(Exploratory Data Analysis : EDA)이란 데이타의 구조적 특성을 알아내기 위한 통계적 분석기법들을 통칭한다. 이에 관하여는 이미 국내에서도 통계학과 학부 교과로 여러 대학에 개설되었으며 백운붕 · 허명회(1987)에 의하여 교재도 간행된 바 있다. 필자들은 「탐색적 데이타분석」의 교과를 운영하면서 Minitab과 NCSS라는 두 통계 팩키지 프로그램을 사용하여 본 경험을 살려 몇가지 일반통계 팩키지 프로그램(general statistical package program)들의 EDA기능을 비교해 보고자 한다.

물론 EDA를 위하여 반드시 컴퓨터를 활용해야 하는 것은 아니다. EDA의 창시자라고 할 수 있는 튜키 John. W. Tukey는 오히려 '연필-과-종이'(pencil-and-paper) 방식의 중요성을 강조하고 있다(Tukey, 1977). 그 이유는 데이타 분석자가 컴퓨터(특히 팩키지 프로그램)에 지나치게 의존하는 경우, 데이타 분석을 단순히 기계적으로 요식화하는 등 일종의 매너리즘에 빠져 통계적 사고를 오히려 저하시킬 우려가 있기 때문이다. 그러나 근래 개발된 여러 로버스트 통계 기법들과 통계 그래픽 기법들은 컴퓨터를 활용하지 않고는 '그림의 떡'으로 그칠 수 밖에 없으므로, 어느 정도 컴퓨터 팩키지 프로그램에의 의존은 탐색적 데이타의 본래 목적을 위해서도 불가피하다고 하겠다.

필자들은 통계학 전공 학부 학생을 대상으로 주간 3시간을 대략 강의 2시간과 컴퓨터 실습 1시간으로 「탐색적 데이타분석」의 교과를 운영하였는데, 학생들의 성취도와 과목에 대한

* 고려대학교 통계학과 교수
** 고려대학교 대학원 통계학과 석사과정 졸업

만족도는 매우 높은 편이었다. 부수적으로 PC를 처음 접한 대부분의 학생들에게는 자연스럽게 DOS 등의 PC에 관한 기초 지식을 습득하는 기회가 되었다.

2. 통계 팩키지 프로그램의 EDA 기법별 비교

통계 팩키지 프로그램을 비교검토하는 경우에는, 통계분석방법에 관한 것 이외에도 데이타의 편집 · 변환 · 관리 기능이라든가 모니터나 프린터에의 출력기능 등 여러 요소에 대한 고려를 해야 할 것이다. 그러나 우선 이 절에서는 각 EDA 기법이 여러 팩키지 프로그램에 어떻게 수용되어 있는가를 검토하기로 하고 기타 소프트웨어 관련사항에 관한 논의는 3절에서 다루기로 하겠다.

일반적으로 데이타분석을 하기 위한 통계 소프트웨어는 크게 두가지 종류로 나눌 수 있다. 다양한 통계처리를 할 수 있는 일반 통계 소프트웨어(general statistical software)가 하나이고, 통계처리가 특정한 통계기법에 제한된 stand-alone program이 다른 하나이다. 우선 여기서는 stand-alone program들을 비교대상에서 제외하기로 하는데, 그 이유는 일반적인 stand-alone program들이 다른 소프트웨어와 호환이 어렵고 document가 부족하기 때문이다(Jones, 1988).

본 소고에서 고려한 소프트웨어는 국내에서 비교적 쉽게 구할 수 있는 PC버전의 통계 팩키지 프로그램들로서, 가장 널리 알려진 SAS와 SPSS/PC+, 교육용으로 적당한 Minitab, 행동과학(behavioral science)에서 자주 사용되는 SYSTAT, 그리고 그래픽 기능이 뛰어난 NCSS와 Statgraphics, 이상의 여섯 종류이다. 여기서 상정하고 있는 전산환경은 허큘리스 그래픽 카드(Hercules graphic card)와 20메가 바이트 이상의 하드 디스크가 부착되고 운영체제는 DOS버전 3.0 이상인 XT 혹은 AT, 흑백 모니터와 도트 매트릭스 프린터로 국내 대학에서 흔히 볼 수 있는 경제적인 시스템이며 별도의 그래픽 플롯터나 칼러 모니터, 수치보조 연산장치는 생각하지 않는다. 물론 위의 여섯가지 이외에도 BMDP 등 여러 일반 통계 소프트웨어들도 고려할 수 있겠으나 대개 국내에서 구하기가 쉽지 않으며 특히 BMDP는 수치보조 연산장치가 필요하다는 제한 때문에 분석에서 제외하였다.

2절과 3절에서 검토될 IBM-PC의 XT/AT 호환기종에서 사용가능한 여섯가지 통계팩키지를 간단히 살펴보면 다음과 같다(김병천, 1987 ; Berk, 1985, 1987 ; 박낙원, 1989a, 1989b).

- Minitab(Release 6. 11, 1987) : 컴퓨터를 처음 접하는 사용자를 위하여 통계학을 교육하면서 컴퓨터를 통한 데이타 분석법을 가르치는 프로그램이다.
- NCSS(Version 5. 1, 1987) : 메뉴운영방식으로 그래픽기능이 뛰어난 프로그램이다.
- SAS(SAS/Base와 SAS/STAT : Release 6.03, 1987) : 데이타의 편집기능이 뛰어나고 다양한 통계처리를 제공하는 범용 프로그램이다.
- SPSS/PC+ (Version 2.0, 1987) : 명령문이 대화형식으로 되어 있어 팩키지를 어느 정도 다룰 수 있는 사용자들에게 매우 편리한 프로그램이다.
- Statgraphics(Version 2. 6, 1987) : 메뉴운영방식으로 그래픽 기능이 우수하고 APL*

PLUS를 이용하여 사용자가 소프트웨어를 수정하고 그 기능을 향상시킬 수 있도록 만든 프로그램이다.

- SYSTAT(Version 3.0, 1985) : 모듈형식으로 되어 있고 여러 운영체제(DOS, Unix, CP/M)에서 사용할 수 있는 프로그램이다.

구체적으로 어떤 통계적 방법들을 EDA기법으로 분류하느냐 하는 문제에는 약간의 주관성이 내재하기 마련이다. 여기에서는 근래에 나온 관련 참고서적들 – 예를 들어 Vellman and Hoaglin(1981), Hoaglin, Mosteller and Tukey(1982, 1985), Chambers, Cleveland, Kleiner and Tukey(1983)– 에 설명되고 있는 다음의 기법들을 중점적으로 검토하기로 한다. 단, 지면의 제약 때문에 몇가지 특징적인 기법(상자그림, 스캐터플롯 브러싱)에 대해서만 각 팩키지별 수행결과들을 제시하고자 한다.

- 줄기와 잎(stem-and-leaf) : Minitab, SAS, SPSS/PC+, NCSS, Statgraphics, SYSTAT 등에서는 가능하다. 그러나 SPSS/PC+의 경우 MANOVA 모듈을 통해서만 줄기와 잎이 가능하다.
- 확률밀도(probability density 혹은 density trace) : NCSS는 고해상도 그래픽 출력이 가능한 비모수적 확률밀도 추정기법을 수용하고 있다. 반면 다른 팩키지 프로그램들에서는 히스토그램 정도만 가능할 뿐이다.
- 상자그림(box plot) : Minitab, NCSS, SAS, SPSS/PC+, Statgraphics, SYSTAT 등에서 가능하다. 그러나 SAS, SPSS/PC+와 SYSTAT의 출력결과는 텍스트 모드이기 때문에 보기가 그다지 좋지 않다. 그리고 SPSS/PC+의 경우는 상자그림을 얻기 위하여는 MANOVA 모듈을 수행하여야 한다. 각 팩키지별 상자그림은 <그림 1>과 같다.
- 문자 값(letter values) : 중위수, 4분위수, 8분위수, 16분위수 등과 mid, spread 등의 원론적인 EDA 기법으로서의 문자값은 Minitab에만 수용되어 있다.
- 루토그램(rootogram) : Minitab과 Statgraphics에 수용되어 있다. Minitab은 정규분포에의 적합결과인 2배 제곱근 잔차(DRRS)까지 계산하나 Statgraphics는 편차(deviation)만 출력한다. 그 밖의 팩키지 프로그램들에는 루토그램에 관한 기능이 없다.
- 2차원 플롯(two-dimensional plot) : SAS, SPSS/PC+와 SYSTAT는 두 변수의 스캐터플롯(散點圖)을 제공하나 텍스트 모드로 출력되므로 보기에 좋지 않다. 반면 Minitab과 NCSS, Statgraphics는 고해상도의 그래픽 출력을 제공한다.
- 저항성 직선(resistant line) : Minitab은 단순회귀적합에의 저항성 직선이 수용되어 있는데 간접적으로는 다중회귀의 경우로의 일반화도 가능하다.
- 시계열자료의 평활(smoothing techniques for time-series) : Minitab은 4253H twice와 3RSSH twice의 두 저항성 비선형 평활법(resistant nonlinear smoother)을 수용하고 있다. Statgraphics는 3RSS, 3RSSH, 5RSSH, 3RSR 등의 기법을 수용하고 있다.
- 산점도의 평활(scatterplot smoothing) : NCSS는 중위수 평활과 LOWESS평활을, SYSTAT은 LOWESS평활을 수용하고 있다.

- 2원 분석에서의 중위수 다듬기(median polish) : Statgraphics와 Minitab에서만 가능하다.
- 3차원(이상) 플롯(three or higher dimensional plot) : NCSS와 Statgraphics는 산점도행렬(scatterplot matrix, draftsman's display)과 창틀그림(casement display)을 제공한다. 창틀그림의 경우 Statgraphics가 NCSS보다 우수한 플롯을 만들어 낸다. NCSS에서는 이외에도 해바라기(sunflower display)라는 일종의 2차원 데이타의 비모수적 확률밀도 추정이 가능하다. 특히 NCSS는 다이내믹 그래픽 기법인 스캐터플롯 브러싱(scatterplot brushing)과 실시간 3차원 회전(real time three-dimensional rotation)을 지원한다는 점은 매우 특기할 만한 사실이다(Becker and Cleveland, 1987 ; Becker Cleveland and Wilks, 1987). NCSS의 스캐터플롯 브러싱의 결과는 〈그림 2〉와 같다.
- 기타 다변량 그래픽 기법(multivariate graphic techniques) : NCSS에서는 Chernoff face, Andrew curve(trig function), star plot, parallel axis plot이 가능하고, Statgraphics에서는 star plot과 sun ray plot이 가능하다.

〈표 1〉은 이상을 기초로 하여 소프트웨어별, 항목별로 요약 · 평가한 결과를 보여준다.

3. 기타 소프트웨어 관련사항

EDA를 효율적으로 지원하기 위하여는 적어도 두가지 필요조건이 있다. 첫째는 그래픽 기능이 가능하여야 한다는 것이고 둘째는 데이타의 편집 · 변환 · 관리기능이 좋아야 한다는 것이다.

앞의 여섯 소프트웨어 중에서 허큘리스 모니터와 도트 매트릭스 프린터 정도로 그래픽 기능이 다양한 것은 NCSS와 Statgraphics이다. SAS의 경우 SAS/GRAPH를 이용하면 고해상도의 2차원 플롯과 3차원 플롯이 가능하나 2절에서 고려한 대개의 기법은 수용되어 있지 않아 비교대상에서 제외되었다. SPSS/PC+의 경우는 Microsoft Chart가 있어야 그래픽 기능을 살릴 수 있다. SYSTAT의 경우 최근에 나온 그래픽 버전에서는(필자들이 직접 확인하여 보지는 못하였으나) NCSS나 Statgraphics에 견줄만한 그래픽 기능이 있다고 한다.

데이타의 편집 · 변환 · 관리의 측면에서는 EDA 교과에서 흔히 다루는 작은 자료 화일의 경우 앞의 여섯 소프트웨어는 유사한 정도의 성능을 갖고 있다. 물론 교육용으로는 흔히 다루지 않는 매우 큰 자료의 경우에는 처리 용량과 효율에 있어 뚜렷한 차이가 있을 수 있을 것이다. 특히 SAS같은 경우에 있어서는 선행결과를 후속되는 자료분석의 입력자료로 어느 정도 자유롭게 이용할 수 있다는 것은 잘 알려진 사실이다.

여러 소프트웨어를 병용할 때의 실제적 문제점 가운데 중요한 요소는 한 소프트웨어의 출력화일을 다른 소프트웨어의 입력화일로 쓸 수 있느냐 하는 것이다. 이런 호환성을 위해서 ASCII, dBASE, Lotus 1-2-3화일 등을 입출력할 수 있는 기능을 각 소프트웨어가 자체적으로 갖는 것이 바람직하다. 각 소프트웨어별로 가능한 입출력 화일의 종류를 정리하여 보면 〈표 2〉와 같다. 이 표에서 볼 수 있듯이 여섯 소프트웨어 모두 ASCII 화일의 자료를 읽을

수 있다. 그러나 NCSS에서는 ASCII화일로 출력할 수 없고, SAS에서는 이것이 가능하기는 하나 경우에 따라 매우 불편할 수 있다(output 윈도우에서 file 명령어를 사용한 뒤 워드프로세서 등을 써서 자료 이외의 부분을 편집·삭제해야 한다).

4. 결 론

이상에서 검토하여 본 여러 사항을 종합적으로 고려하여 Minitab과 NCSS 또는 Statgraphics를 국내 대학의 전산환경 하에서 탐색적 데이타분석(EDA)의 교과와 실제 자료분석에 적합한 팩키지 프로그램으로 추천한다. 그 이유로는 Minitab이 대개의 EDA 고유의 방법론을 수용하고 있으며 데이타의 변환기능이 뛰어나고 배우기 쉽고 교육적인 소프트웨어이기 때문이다. 한편 NCSS와 Statgraphics는 그래픽 기능이 PC 수준에서 뛰어나다는 점 때문에 둘 중의 적어도 한 팩키지 프로그램은 EDA를 효율적으로 지원하기 위하여 필요하다고 생각된다. NCSS는 Statgraphics에 비교하여 다이내믹 그래픽스가 가능하고 다양한 다변량 그래픽 기법이 가능한 장점이 있는 대신 데이타의 입출력·관리기능 면에서는 열등하고 자체 출력기능이 없다는 단점이 있다.

〈참 고 문 헌〉

(1) 김병천(1987) "개인용 컴퓨터에서의 통계팩키지의 선택과 활용", 응용통계연구 1권 1호, 75-90.

(2) 박낙원(1989a) "통계와 그래픽의 만남", 소프트 월드 4월호, 169-175.

(3) 박낙원(1989b) "사용하기 쉬운 통계팩키지 NCSS", 소프트월드 7월호, 175-181.

(4) 백운붕·허명회(1987) 「EDA-탐색적 데이타분석」, 서울 : 박영사.

(5) Becker, R. A., and Cleveland, W. S.(1987) "Brushing Scatterplots", *Technometrics* 29, 127-142.

(6) Becker, R. A., Cleveland, W. S., and Wilks, A. R. (1987) "Dynamic Graphics for Data Analysis", *Statistical Science* 2, 355-383.

(7) Berk, K. (1985) "Review of statistical software", *Americal Statistician* 39, 67-70.

(8) Berk, K. (1987) "Review of statistical software", *Americal Statistician* 41, 64-67.

(9) Chamber, J, M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983) *Graphical Methods for Data Analysis*, Duxbury Press, Boston.

(10) Hoaglin, D. C, Mosteller, F., and Tukey, J. W. (1982) *Understanding Robustness and Exploratory Data Analysis*, Wiley, New York.

(11) Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (1985) *Exploring Data Tables, Trends, and Shapes*, Wiley, New York.

(12) Jones, P. K. (1988) "State of the art microcomputer software for logistic regression", Americal Statistical Association 1988 Procceedings of the Section on Statistical Education, 26-30.

(13) Ryan, B. F., Joiner, B. L., and Ryan, Jr. T. A. (1985) *Minitab*, Duxbury Press, Boston.

(14) Tukey, J. W. (1977). *Exploratory Data Analysis*, Addison-Wesley, MA : Reading.

(15) Vellman, P. F., and Hoaglin, D. C. (1981). *Applications, Basic and Computing of Exploratory Data Analysis*, Duxbury Press, Boston.

〈표 1〉 EDA 기법별 통계 팩키지 프로그램의 비교

	Minitab	NCSS	SAS	SPSS/PC+	Statgraphics	SYSTAT
줄기와 잎	☆☆	☆	☆☆	☆	☆☆	☆☆
확률밀도		☆☆				
상자 그림	☆☆	☆☆☆	☆☆	☆	☆☆☆	☆
문자 값	☆☆☆	☆	☆	☆	☆	☆
루토그램	☆☆				☆	
2차원 플롯	☆☆	☆☆☆	☆	☆	☆☆☆	☆
저항성 직선	☆☆☆					
시계열 평활	☆☆				☆☆	
산점도 평활		☆☆				☆☆
중위수다듬기	☆☆				☆☆	
3차원 플롯	☆	☆☆☆			☆☆	
다변량그래픽		☆☆☆			☆☆	

☆ 가능함, ☆☆ 좋음, ☆☆☆ 매우 좋음.

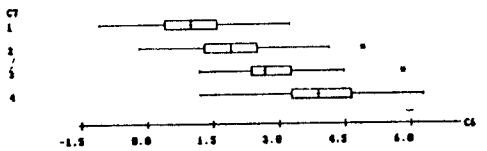
〈표 2〉 팩키지 프로그램의 화일 import및 export 기능

	Minitab	NCSS	SAS	SPSS/PC+	Statgraphics	SYSTAT
IMPORT ASCII	Y	Y	Y	Y	Y	Y
dBASE	N	N	Y	Y	Y	Y
Lotus	Y	N	N	Y	Y	Y
EXPORT ASCII	Y	N	Y/N	Y	Y	Y
dBASE	N	N	Y	Y	Y	N
Lotus	Y	N	N	Y	Y	N

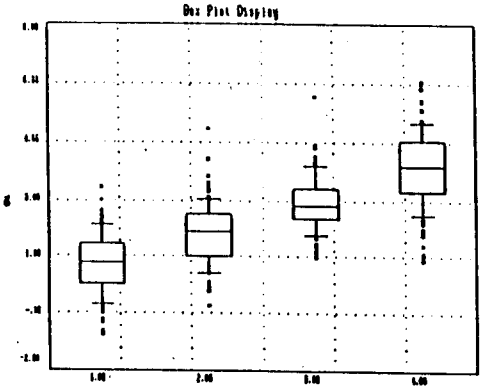
Y : yes, N : no.

〈그림 1〉 각 통계 팩키지별 상자그림

a) Minitab

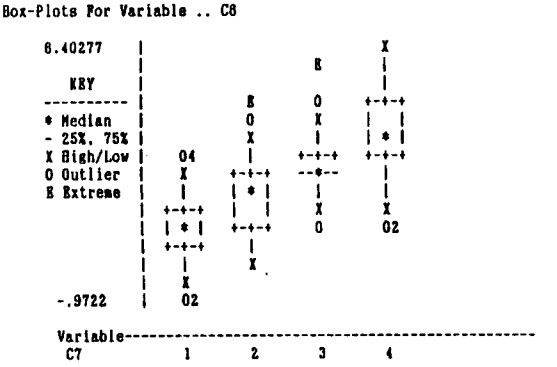


b) NCSS

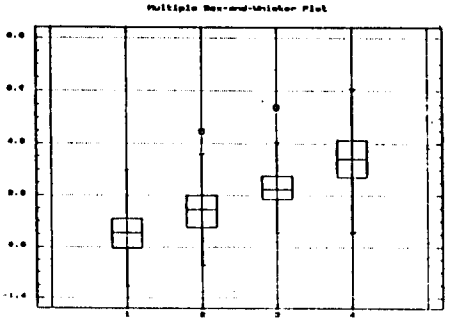


c) SAS

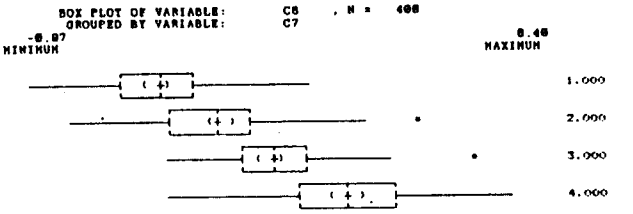
d) SPSS



d) Statgraphics

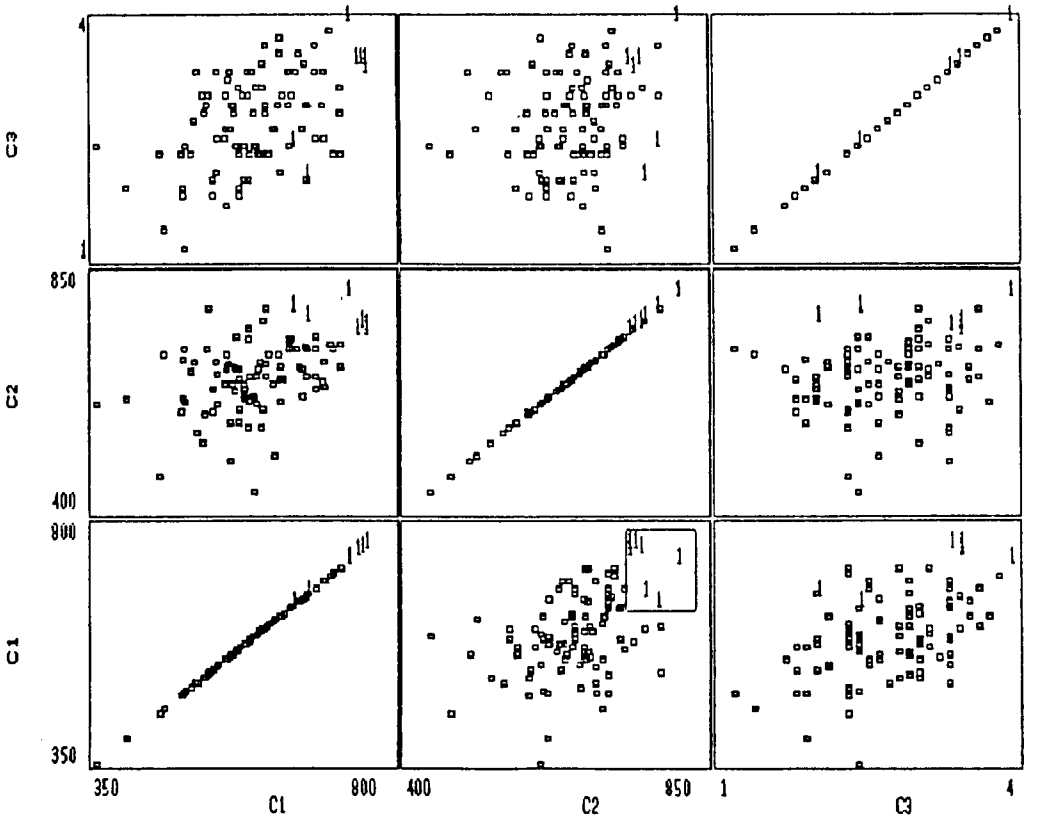


e) SYSTAT



자료출처: Normal(1,1), Normal(2,1), Normal(3,1), Normal(4,1) 에서 각각 100개의 표본을 임의로 추출한 가상의 자료임.

〈그림 2〉 NCSS의 스캐터플롯의 브러싱



C1=VERB, C2=MATH, C3=GPA.

자료출처: Ryan, Joiner, and Ryan(1985)의 Grades Data: Sample A, 309-312.

Software Review of Statistical Package Programs on EDA Aspects

Myung-Hoe Huh*, Jin-Whan Jung**

〈Abstract〉

We will compare several statistical package programs on aspects of exploratory data analysis(EDA). Specifically, PC versions of Minitab, NCSS, SAS, SPSS/PC+, Statgraphics, and SYSTAT are reviewed.

*Dept. of Statistics, Korea University, Anam-dong, Seoul 136-701, Korea.
**Dept. of Statistics, Korea University, Anam-dong, Seoul 136-701, Korea.