열람자 주의사항

본 저작물의 저작권자는 인구서머세미나 사무국이며, 본 자료를 다운로드하여 블로그, SNS에 업로드하여 타인과 공유하는 행위는 저작권을 침해하는 행위입니다.

- (저작권법 제136조의 벌칙) ‘저작재산권 그 밖에 저작권법에 따라 보호되는 재산적 권리를 복제, 공중송신, 배포, 대여, 2차적저작물 작성의 방법으로 침해한 자에게는 5년 이하의 징역 또는 5천만원 이하의 벌금에 처하거나 이를 병과할 수 있다 ’

Caution

The copyright of this mateiral is owned by the 7th Summer Seminar on Population Secretariat. If permission from the copyright holder is not acquired before posting their material online (e.g. posting the material on SNS or distributing by e-mail), this activity could be considered a copyright infringement.

GIS를 이용한 인구자료 분석

홍성연(syhong@khu.ac.kr)

2020년 8월 10일 (월) ~ 14일 (금)
대한상공회의소 중회의실B

The 7th KOSTAT-UNFPA
Summer Seminar on Population

Contents

01. 강의소개 ---------- 5
02. GIS의 개념과 공간데이터 ---------- 13
03. GIS 소프트웨어 ---------- 33
04. 좌표 체계 ---------- 41
05. 근접성 분석과 중첩 분석 ---------- 65
06. 공간데이터의 시각화 ---------- 81
07. 공간적 자기상 관성의 개념 ---------- 95
08. 공간 가중 행렬 ---------- 107
09. 국지적 측도 ---------- 119
10. R과 RStudio 소개 및 설치 ---------- 133
11. 데이터 입력과 객체, 클래스 ---------- 159
12. 파일 데이터 열기 ---------- 173

The 7th KOSTAT-UNFPA
Summer Seminar on Population

01

강의소개

1. 강의 소개

1) 강의 개요
2) 강사 소개
3) 세부 일정

강의 개요

- 본 워크숍의 목표는 GIS와 통계분석 소프트웨어를 사용하여 다양한 인구통계 데이터를 지도에 나타내고, 인구 분포를 분석하는 공간통계 기법을 소개하는 것
  - GIS 프로그램을 처음 접하는 인구 및 사회과학 분야 연구자를 대상으로 하며, GIS 프로그램의 기초적인 사용법부터 설명함
  - 강의 및 실습 내용에는 단계구분도, 카토그램, 커널 밀도 지도, 모란 I, LISA, 상이지수 등이 포함됨
  - 실습은 오픈소스 GIS 프로그램인 QGIS를 위주로 진행하며, 후반부에는 통계 프로그램 R도 사용함

강사 소개

- 홍성연
  - 경희대학교 지리학과 조교수(2016년 2월 ~ 현재)
  - 연구 분야: 거주지/활동공간 분리(segregation in neighbourhoods and activity spaces), 도시분석학(urban analytics)
  - 경력:
    - 인천대학교 도시행정학과 조교수(2015년 2월 ~ 2016년 1월)
    - 일본 도쿄대학교 도시공학과 JSPS 박사후연구원(2012년 11월 ~ 2014년 11월)
    - 뉴질랜드 오클랜드대학교 환경학부 지리학 전공 박사(2011년 10월)
    - 뉴질랜드 오클랜드대학교 통계학과 학사(2007년 5월)

강사 소개

- 최창락
  - 경희대학교 지리학과 박사과정(2020년 ~ 현재)
  - 연구 분야: 데이터 마이닝(data mining), 빅데이터 분석(big data analysis)
  - 이메일: hihi7100@khu.ac.kr

- 정예원
  - 경희대학교 지리학과 석사과정(2020년 ~ 현재)
  - 연구 분야: 공간분석(spatial analysis), 거주지 분리(residential segregation)
  - 이메일: elpis28@khu.ac.kr

세부 일정

- 1일차: 지리정보시스템과 공간데이터, QGIS 프로그램 소개

시간	유형	주제	비고
10:00 – 10:50			
11:00 – 11:50			
13:30 – 14:20	세부 일정은 강의 시간에 안내		
14:30 – 15:20			
15:30 – 16:20			
16:30 – 17:20			

세부 일정

- 2일차: 데이터 편집과 시각화

시간	유형	주제	비고
09:00 – 09:50			
10:00 – 10:50			
11:00 – 11:50	세부 일정은 강의 시간에 안내		
13:30 – 14:20			
14:30 – 15:20			
15:30 – 16:20			

세부 일정

- 3일차: 공간적 자기상관의 개념과 전역적, 국지적 측도

시간	유형	주제	비고
09:00 – 09:50			
10:00 – 10:50			
11:00 – 11:50	세부 일정은 강의 시간에 안내		
13:30 – 14:20			
14:30 – 15:20			
15:30 – 16:20			

세부 일정

- 4일차: 공간데이터와 R

시간	유형	주제	비고
09:00 – 09:50			
10:00 – 10:50			
11:00 – 11:50	세부 일정은 강의 시간에 안내		
13:30 – 14:20			
14:30 – 15:20			
15:30 – 16:20			

세부 일정

- 5일차: 거주지 분리의 개념과 측정

시간	유형	주제	비고
09:00 – 09:50			
10:00 – 10:50			
11:00 – 11:50		세부 일정은 강의 시간에 안내	
13:30 – 14:20			
14:30 – 15:20			
15:30 – 16:20			

The 7th KOSTAT-UNFPA
Summer Seminar on Population

02

GIS의 개념과 공간데이터

2. GIS의 개념과 공간데이터

1) 지리정보시스템의 정의
2) 벡터 데이터
3) 래스터 데이터
4) 속성 테이블
5) 메타데이터

지리정보시스템

- 지리정보시스템의 정의
  - 특별한 목적을 위해 실세계로부터 공간 데이터를 수집, 저장하고 원하는 형태로 검색, 변형, 표현하기 위한 일련의 도구 (Burroughs, 1986)
  - 컴퓨터 데이터베이스에 저장된 공간 및 비공간정보를 조작, 요약, 쿼리, 수정, 시각화하기 위해 사용되는 정보시스템(Goodchild, 1997)
  - 공간 데이터를 수집, 저장, 검색, 분석, 표현하기 위한 자동화된 시스템(Slocum et al., 2005)

공간데이터 → *GIS SW* → 정보

데이터와 정보의 구분

- 데이터와 정보, 지식의 구분(McDonough, 1963)
- 데이터(data)
  - 현실세계에서 측정 등을 통해 얻은 값으로 객관적 실제를 반영
  - 모든 기록은 데이터가 될 수 있음
  - 예: 날씨, 온도, 전국의 치과병원 주소와 전화번호
- 정보(information)
  - 자료를 특정한 목적과 문제해결에 도움이 되도록 가공한 것
  - 예: 우리 동네 날씨, (치과를 찾고 있는 상황에서) 내게 가까운 치과의 주소와 전화번호

2. GIS와 공간데이터

데이터와 정보의 구분

- 동적인 관계
  - 모든 데이터는 특정한 상황에서 정보가 될 수 있음
  - 데이터 ⊃ 정보
- 정보의 가치는 절대적이지 않음
  - 어떤 데이터는 특정 시점에 누군가에 매우 유용한 가치의 정보가 될 수 있으나, 다른 사람에게는 아무 가치 없는 단순한 자료일 수 있음
  - 정보는 우리가 원하는 시간에 제공되어야 하며, 적시에 제공되지 못하는 정보는 그 가치가 매우 떨어짐

지리정보과학

- 공간데이터를 수집, 처리, 표현, 분석하는 방법에 대해 연구하고 개선을 추구하는 학문 분야
  - 지리정보과학이 발전함에 따라 우리는 원하는 정보를 보다 정확하고 신속하게 얻을 수 있음
- 지리정보과학의 다양한 연구 주제(https://giscience.org/)
  - 데이터 수집과 가공: Geo-APIs, high-performance computing algorithms for spatial-temporal data, location privacy, spatial data infrastructures, uncertainty quantification and error propagation
  - 데이터 시각화: Geovisualization and visual analytics
  - 데이터 분석과 시뮬레이션: GeoAI, geosimulation and spatio-temporal modelling, spatial aspects of social computing, spatially-explicit machine learning

데이터 모형

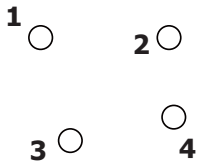
- 데이터 모형(data model)은 실제 세계의 사상(features)을 컴퓨터에서 나타내는 방법을 의미함
  - 크게 벡터(vector) 모형과 래스터(raster) 모형으로 구분됨

벡터 모형

래스터 모형

벡터 데이터

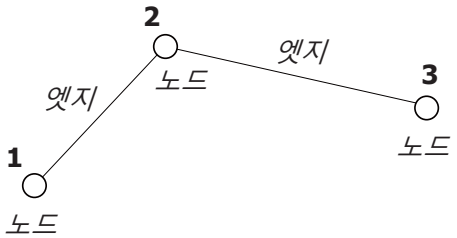
- 벡터 모형의 데이터는 실세계의 다양한 대상물을 컴퓨터에서 표현하기 위해 점, 선, 면을 사용함
  - 가옥은 점, 도로는 선, 필지 등은 면을 사용해 나타낼 수 있음
- 점(point)
  - 벡터 지도의 기본 단위로, 하나의 (x, y) 좌표와 속성 정보로 구성됨



Point ID	x	y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
4	x_4	y_4

벡터 데이터

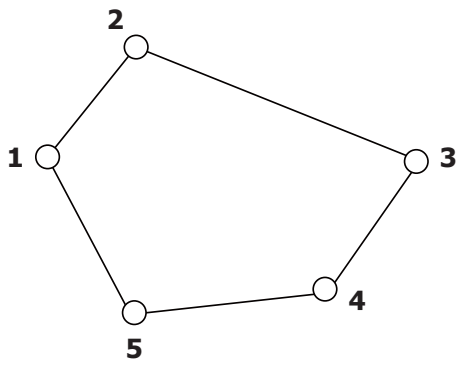
- 선(line)
  - 위치와 길이라는 속성을 갖는 1차원 사상으로 최소 두 개의 점을 갖고 있음
  - 일반적으로 각각의 점을 노드(node)라 부르며, 두 개의 노드를 잇는 선분을 엣지(edge)라고 부름



Point ID	x	y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3

벡터 데이터

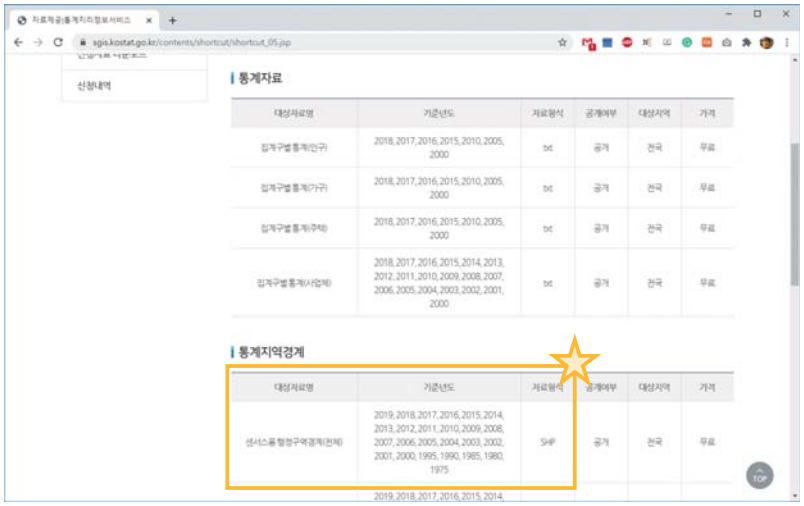
- 면(polygon)
  - 연속된 선으로 둘러싸인 2차원 객체로, 크기와 둘레 정보를 가질 수 있음



Point ID	x	y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
4	x_4	y_4
5	x_5	y_5
6	x_1	y_1

벡터 데이터의 예

- 센서스용 행정구역 경계, 도로명주소 전자지도 등이 모두 벡터 데이터의 예임



벡터 데이터 포맷

- Esri Shapefile
  - Esri에서 개발한 벡터 데이터 파일 포맷으로 ArcGIS를 비롯한 많은 GIS 소프트웨어에서 실질적(de facto) 표준으로 사용되어 왔음
  - 1990년대 초에 개발되어 ArcView GIS 2.0부터 사용되기 시작함
  - 다음의 표와 같이, 하나의 (벡터) 공간데이터가 같은 이름을 갖는 여러 개의 파일로 이루어짐

파일 확장자	기능
.shp	좌표 등 공간데이터가 기록된 파일
.shx	공간데이터의 빠른 읽고 쓰기를 위한 인덱스 파일
.dbf	dBase IV 형식으로 기록된 속성값

벡터 데이터 포맷

- Esri Shapefile (이어서)
  - 이 외에도 많은 파일이 부수적으로 만들어질 수 있음(예: 투영법 정보가 들어가는 prj 파일, 컴퓨터가 읽고 쓰기 편리한 바이너리 형식으로 인덱스가 기록된 sbn 파일, 메타 데이터가 들어가는 shp.xml 파일 등)
  - 거의 대부분의 GIS 소프트웨어에서 사용 가능하다는 장점이 있으나, 단점도 많음
    - 공간데이터를 구성하는 각각의 파일 크기는 최대 2GB로 제한됨
    - 속성 테이블의 필드 이름은 최대 10자를 넘을 수 없고, 필드 수도 255개 이내로 제한됨
    - 속성 테이블에서 값이 없음(NULL)은 0으로 기록됨(문자열의 경우는 공백으로 기록되고, 날짜의 경우 테이블 상에서는 <null>로 표시되나 실제 기록은 0으로 됨)

벡터 데이터 포맷

- Esri Geodatabase
  - 기존의 Shapefile이 가지는 여러 가지 문제를 해결하기 위해 ESRI에서 새롭게 개발한 공간데이터 저장·관리 방법
    - 기본적으로 1TB 크기의 공간데이터까지 지원하나, 고해상도 위성영상과 같은 빅데이터를 위해 용량 제한을 최대 256TB까지 확장 가능
    - 각각의 속성 테이블은 최대 65,534개의 필드를 가질 수 있음
    - 각 필드의 이름 또한 64자까지 기록이 가능하며, 빈 값을 0으로 기록하는 문제도 없음
  - 많은 장점을 가지고 있으나, Esri 소프트웨어에서만 사용이 가능하다는 제한이 있음

벡터 데이터 포맷

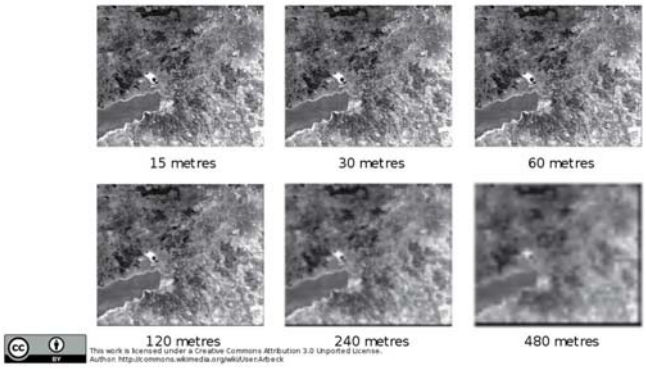
- OGC GeoPackage
  - 플랫폼이나 소프트웨어에 관계 없이 사용할 수 있는 개방된(open) 공간데이터 포맷
  - Esri Geodatabase와 마찬가지로 벡터 데이터와 래스터 데이터 모두 기록할 수 있음
  - QGIS의 기본 공간데이터 포맷이며, Esri ArcGIS, R 등 많은 소프트웨어에서 지원하고 있음
  - Shapefile과 마찬가지로 파일 형태(.gpkg)로 저장되고 사람들과 공유할 수 있으나, 내부 구조는 데이터베이스 형식임
  - 단일 파일의 최대 용량은 140TB이나 운영체제의 파일 시스템에 따라 실제 사용가능한 용량은 이보다 적을 수 있음(FAT32의 경우는 4GB)

래스터 데이터

- 래스터 모형의 데이터는 전체 지역을 일정 크기의 단위 셀(cell), 또는 픽셀(pixel)로 분할하고, 각 셀에 속성값을 부여하여 실세계의 다양한 대상물을 표현함
  - 넓게 보면 위치 정보를 가지고 있는 모든 사진 파일을 래스터 데이터로 생각할 수 있음
  - 위성영상, 항공사진 등이 대표적인 래스터 데이터의 예
  - 웹 브라우저를 통해 사용자에게 보여지는 구글 지도, 카카오 지도 등도 (사용자 입장에서는) 모두 래스터 데이터임
- 일반적으로 같은 면적을 나타내는 벡터 데이터보다 많은 저장 공간을 필요로 함

래스터 데이터

- 래스터 데이터에서 셀 하나가 나타내는 지표면의 면적을 공간해상도(spatial resolution)이라 함
  - 셀 하나가 나타내는 지표면의 면적이 좁을수록 공간해상도는 높아짐
  - 공간해상도가 15m x 15m인 영상에서는 공간해상도가 60m x 60m인 영상보다 더 작은 물체를 정확하게 인식할 수 있음



This work is licensed under a Creative Commons Attribution 3.0 Unported License.
Author: http://commons.wikimedia.org/wiki/User:Arbeck

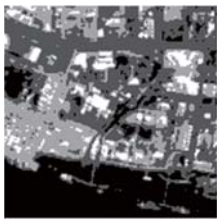
래스터 데이터

- 각 셀에 입력할 수 있는 속성값의 범위가 넓을수록 다양한 색상을 표현할 수 있음
  - 1비트 데이터는 셀의 속성값이 0 또는 1로만 기록됨
  - 8비트 데이터에서는 0부터 255까지의 값이 사용될 수 있으며, 색상도 보다 자연스럽게 나타낼 수 있음

256단계(8비트)

64단계

4단계



2단계

http://www.crisp.nus.edu.sg/~research/tutorial/image.htm#ifov

래스터 데이터

- 래스터 데이터가 여러 개의 밴드로 구성되어 있다면, 밴드를 조합하여 영상 판독을 보다 수월하게 할 수 있음

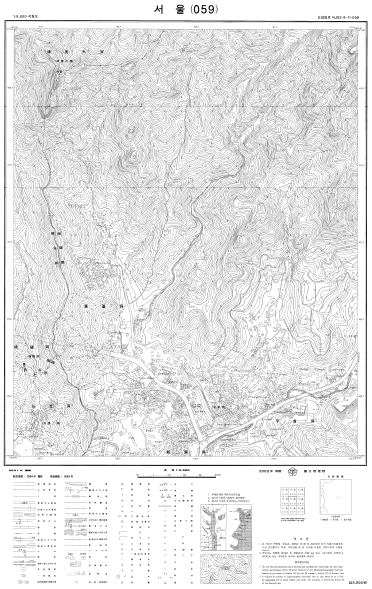
가시광선 데이터

근적외선 데이터

공간해상도가 높고, 각 셀이 나타낼 수 있는 속성값의 범위가 넓으면서, 여러 개의 밴드로 구성된 래스터 데이터 ➔ 높은 품질의 래스터 데이터

래스터 데이터의 예

- 국토정보플랫폼에서 제공하는 항공사진, 정사영상, 구지도 등은 모두 래스터 데이터의 예임
  - 국토정보플랫폼: http://map.ngii.go.kr/
- GIS 프로그램에서 배경으로 사용하는 지도 또한 모두 래스터 데이터임
  - OpenStreetMap (OSM)
  - ArcGIS Map Service의 Light Gray Canvas Base



래스터 데이터 포맷

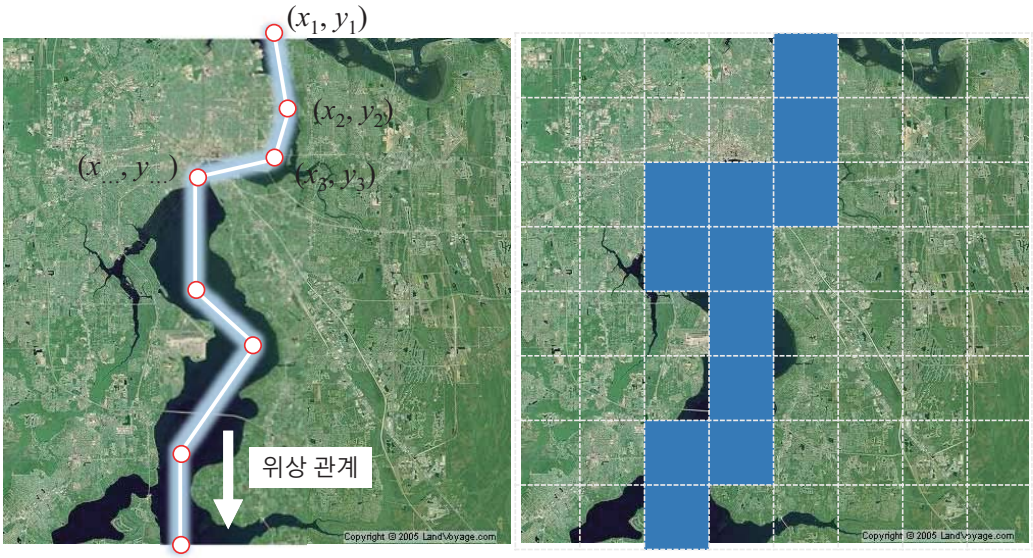
- TIFF (Tagged Image File Format)
  - 앨더스(Aldus, 지금은 Adobe에 인수된 회사)와 마이크로소프트가 공동 개발한 이미지 저장 포맷으로 알려져 있음
  - FileFormatInfo에 따르면 1986년 앨더스에서 TIFF 포맷의 기술 문서를 처음 발간하였으나, 첫 번째 버전은 아님!!
  - LZW, ZIP와 같은 비손실(loseless) 압축 기법을 사용할 수 있기 때문에 데이터 저장 포맷으로써 유용함
    - 압축을 아예 하지 않거나(파일 크기 커짐), JPEG과 같은 손실 압축 기법을 사용하는 것도 가능함(데이터 손실)
  - 내부적으로는 이미지 파일 헤더(Image File Header), 이미지 파일 디렉토리(Image File Directory), 그리고 실제 픽셀 단위 값이 기록된 부분(Bitmap)으로 구성되어 있음

래스터 데이터 포맷

- GeoTIFF (Geographic Tagged Image File Format)
  - 국토지리정보원을 비롯한 많은 국내외 기관, 기업, 단체에서 GeoTIFF 형식으로 위성영상, 항공사진 등 다양한 래스터 데이터를 제공하며, 플랫폼에 관계 없이 사용이 가능하다는 장점을 갖고 있음
  - 기본적으로 TIFF와 동일한 구조이나, 헤더 부분에 좌표체계 (coordinate systems), 준거타원체(ellipsoids), 투영법(projections)과 정보가 추가로 기록되어 있음
  - EPSG (European Petroleum Survey Group)에서 제공하는 투영법 목록과 동일한 방식으로 좌표체계 정보를 기록함
  - GIS 프로그램이 아닌 일반적인 이미지 소프트웨어에서는 보통의 TIFF 파일과 GeoTIFF 파일이 구별되지 않음
  - 기본적으로 32비트 파일 형식이기 때문에, 최대 파일 크기가 4GB로 제한됨 ➔ BigTIFF

벡터 vs. 래스터

- 벡터 데이터와 래스터 데이터의 비교

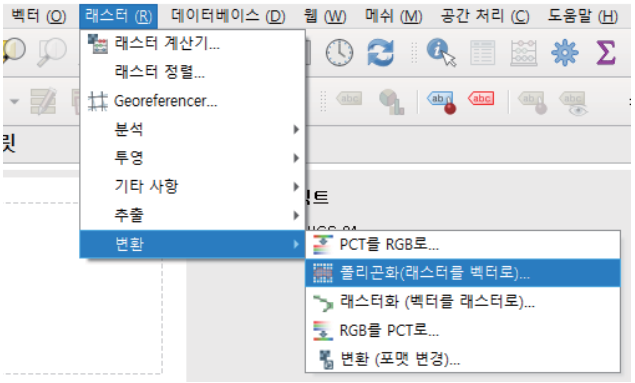


벡터 vs. 래스터

- 벡터 모형과 래스터 모형은 각각 장단점을 갖고 있기 때문에, 어떤 모형이 더 좋은가 하는 것은 상대적인 문제
  - 일반적인 지도의 표현에는 벡터 모형이 효율적일 수 있음
  - 위성영상과 같은 사진을 나타내는 데에는 래스터 모형이 보다 적합함
- 대부분의 GIS 소프트웨어는 벡터 데이터와 래스터 데이터 간의 변환 기능을 제공하며, 따라서 필요에 따라 변환 가능
  - 벡터 데이터를 래스터 데이터로 변환(rasterisation)하는 것은 비교적 간단한 작업으로, 셀 크기에 따라 정확한 결과를 얻을 수 있음
  - 래스터 데이터를 벡터 데이터로 변환(vectorisation)하는 것은 복잡하고 시간이 오래 걸리며, 특히 영상 데이터를 변환하는 경우에는 좋은 결과를 얻기가 어려움

벡터 vs. 래스터

- QGIS에서는 크게 두 가지 방법으로 데이터 모형 변환이 가능함
  - 메뉴에서 래스터 (R) ➔ 변환 ➔ 폴리곤화(래스터를 벡터로)... / 래스터화(벡터를 래스터로)... 선택
  - 공간 처리 툴박스에서 GRASS ➔ 래스터 ➔ r.to.vect / GRASS ➔ 벡터 ➔ v.to.rast 선택



벡터 vs. 래스터

- 공간해상도가 같은 두 래스터 데이터의 중첩은 지도 방정식(map algebra)을 사용해 매우 직관적으로 수행할 수 있음

2019년 강수량

7	5	5	3
1	1	1	1
1	1	1	1
1	1	1	2

2020년 강수량

5	1	2	1
1	1	1	1
1	1	1	2
1	2	2	2

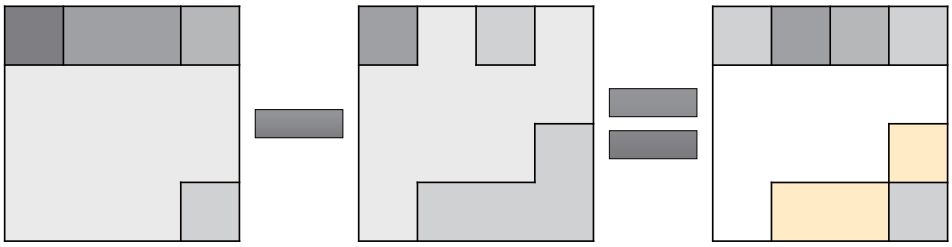
변화량

2	4	3	2
0	0	0	0
0	0	0	-1
0	-1	-1	2

벡터 vs. 래스터

- 위상(topology) 관계*가 포함되지 않은 벡터 데이터는 상대적으로 중첩 분석이 어려움

2019년 강수량 2020년 강수량 변화량



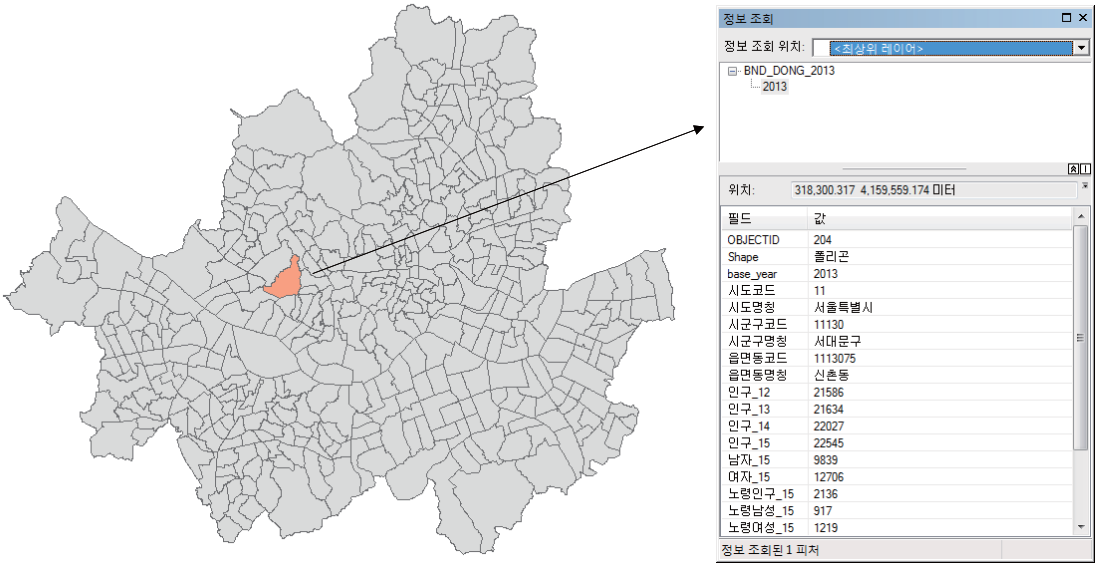
속성 정보

- 공간데이터에는 객체의 위치를 나타내는 좌표는 물론, 해당 객체의 특징을 나타내는 속성 정보도 포함될 수 있음
  - 예: 행정동 경계를 나타내는 벡터 데이터에 포함된 인구 데이터, 국가별 연 평균 강수량 등

건물의 위치 정보
- x, y 좌표 등

건물의 속성 정보
- 건물 유형, 건축 일시, 소유주 이름, 용도 등

속성 정보



벡터 데이터의 속성 테이블

- 점, 선, 면으로 표현된 객체 하나 하나가 고유한 ID 값을 갖게 되며, 해당 ID를 통해 속성 정보가 입력된 테이블과 연결됨

공간 정보

Point ID	*x*	*y*
1	22.7	45.6
2	76.3	48.7
3	32.7	15.8
4	77.1	19.5

속성 정보

Object ID	업종	상호명	규모
1	치과	올치과의원	2
2	치과	제이치과의원	3
3	치과	성신치과의원	2
4	치과	백동준치과의원	1

공통의 필드를 사용해 연결

필드

벡터 데이터의 속성 테이블

테이블

BND_DONG_2013

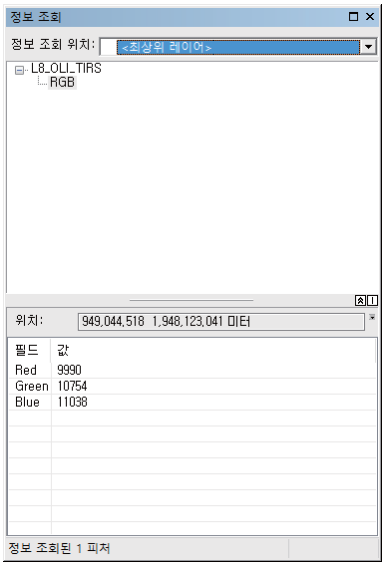
OBJECTID *	Shape *	base_year	시도	시도	시군구	시군구	읍면동	읍면동명칭	인구_1	인구_1	인구_1	인구_1	남자_1	여자_1
1	폴리곤	2013	11	서울특	11010	종로구	1101053	사직동	10477	10327	10132	10033	4689	5344
2	폴리곤	2013	11	서울특	11010	종로구	1101054	삼청동	3485	3367	3268	3174	1506	1668
3	폴리곤	2013	11	서울특	11010	종로구	1101055	부암동	11478	11469	11356	11217	5433	5784
4	폴리곤	2013	11	서울특	11010	종로구	1101056	평창동	19816	19779	19595	19414	9209	10205
5	폴리곤	2013	11	서울특	11010	종로구	1101057	무악동	8598	8626	8445	8619	4070	4549
6	폴리곤	2013	11	서울특	11010	종로구	1101058	교남동	8342	5450	5005	4875	2339	2536
7	폴리곤	2013	11	서울특	11010	종로구	1101060	가회동	5520	5333	5214	4961	2307	2654
8	폴리곤	2013	11	서울특	11010	종로구	1101061	종로1·2·3·4가	8984	8797	8920	8786	5040	3746
9	폴리곤	2013	11	서울특	11010	종로구	1101063	종로5·6가동	6438	6245	6069	6057	3322	2735
10	폴리곤	2013	11	서울특	11010	종로구	1101064	이화동	9521	9435	9312	9216	4372	4844
11	폴리곤	2013	11	서울특	11010	종로구	1101067	창신1동	7669	7237	7097	6862	3666	3196
12	폴리곤	2013	11	서울특	11010	종로구	1101068	창신2동	11984	11707	11311	11040	5564	5476
13	폴리곤	2013	11	서울특	11010	종로구	1101069	창신3동	8502	8310	8091	7933	3981	3952
14	폴리곤	2013	11	서울특	11010	종로구	1101070	숭인1동	7692	7563	7348	7232	3617	3615
15	폴리곤	2013	11	서울특	11010	종로구	1101071	숭인2동	9560	9944	10078	9983	5094	4889
16	폴리곤	2013	11	서울특	11010	종로구	1101072	청운효자동	15219	14696	14472	14354	6802	7552
17	폴리곤	2013	11	서울특	11010	종로구	1101073	혜화동	19863	19582	19631	20066	9520	10546
18	폴리곤	2013	11	서울특	11020	중구	1102052	소공동	1764	1766	1895	1869	1054	815
19	폴리곤	2013	11	서울특	11020	중구	1102054	회현동	6581	6563	6505	6263	3422	2841
20	폴리곤	2013	11	서울특	11020	중구	1102055	명동	4113	4013	3614	3521	1589	1932

1 (0 out of 423 선택함)

BND_DONG_2013

래스터 데이터의 속성 테이블

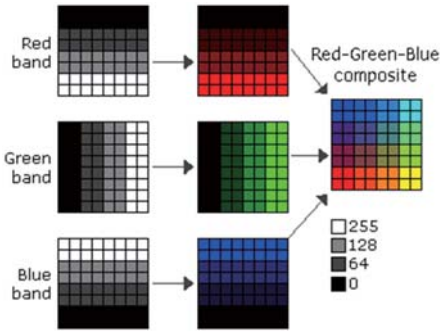
- 래스터 데이터의 각 셀에도 속성 정보가 입력되어 있음



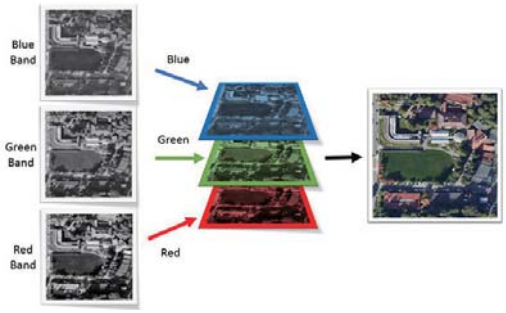
2. GIS와 공간데이터

래스터 데이터의 속성 테이블

- 벡터 데이터와 달리 래스터 데이터, 특히 영상 데이터의 속성 정보는 단순히 셀의 색상을 정의하는 값임



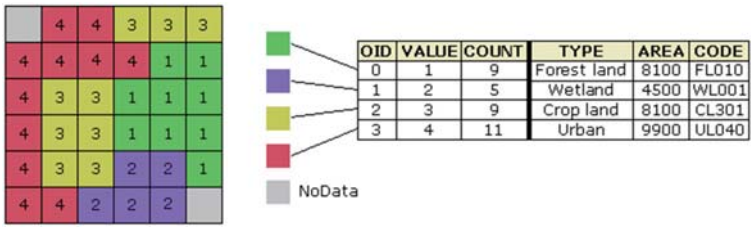
http://desktop.arcgis.com/en/arcmap/10.3/manage-data/raster-and-images/what-is-raster-data.htm



http://gsp.humboldt.edu/olm_2015/Courses/GSP_216_Online/images/3-bands.jpg

래스터 데이터의 속성 테이블

- 벡터 데이터와 마찬가지로 속성 데이터(테이블)의 생성이 가능
  - 다만 16비트 영상 데이터는 밴드 당 2^{16} = 65536개의 서로 다른 값이 나올 수 있으며, 밴드 3개를 조합하는 경우 65536^3 개가 됨
  - 각각의 색상에 대해 속성을 부여하는 것이 무슨 의미가 있을까?
- 대부분의 GIS 소프트웨어에서는 셀 값의 범위가 일정 수준 이하이거나, 중복되는 값을 제외하고 실제 데이터에 포함된 값의 개수가 적을 때에만 속성 테이블을 생성함



OID	VALUE	COUNT	TYPE	AREA	CODE
0	1	9	Forest land	8100	FL010
1	2	5	Wetland	4500	WL001
2	3	9	Crop land	8100	CL301
3	4	11	Urban	9900	UL040

http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=Raster_dataset_attribute_tables

메타데이터

- 데이터에 관한 데이터
  - 데이터의 특성을 기술하거나 설명하고 그 위치 등을 나타내는 구조화된 정보
  - 자원에 대한 조회, 사용, 이해, 판독, 관리 등을 용이하게 하기 위한 또 다른 정보
  - 데이터 신뢰성 및 공유성 보장하는 역할
- 메타 데이터의 주요내용
  - 설명: 검색을 위한 제목, 키워드, 작성자, 개요 등의 제보
  - 권한: 데이터에 대한 접근 및 사용 권한 정보
  - 관리: 데이터의 갱신 빈도, 파일 포맷, 파일 크기 등

메타데이터의 중요성

- 서비스 제공자 측면
  - GIS 분야에서 데이터는 전체 비용 중 70% 가량을 차지
  - 불필요한 중복을 방지하고 정보의 손실을 피함
  - 데이터 관리에 필요한 의사결정을 지원
    - 데이터가 업데이트 되어야 하는 시기는 언제인가?
  - 저작권 등 법적인 문제를 문서화
  - 결과적으로 메타 데이터 구축을 통해 시간과 비용을 절감할 수 있음
- 사용자 측면
  - 데이터의 탐색을 보다 수월하게 하고, 원하는 데이터를 신속하게 검색하는데 도움을 줌

The 7th KOSTAT-UNFPA
Summer Seminar on Population

03

GIS 소프트웨어

3. GIS 소프트웨어

1) GIS 소프트웨어란?
2) 대표적인 GIS 소프트웨어
3) 도메인 GIS 소프트웨어
4) 영상 분석 소프트웨어

GIS 소프트웨어란?

- 래스터 데이터는 기본적으로 사진 파일과 같은 구조(+ 좌표 체계)이기 때문에, 간단한 이미지 뷰어 프로그램으로도 열어볼 수 있음
  - 다만 공간데이터로써 가장 중요한 정보인 좌표 체계 정보가 인식되지 않기 때문에 분석 용도로는 사용할 수 없음
- 벡터 데이터의 열람과 수정, 편집, 분석을 위해서는 전문적인 GIS 소프트웨어의 사용이 필요함
  - 과거 지도 제작이 GIS 소프트웨어의 주요 활용 분야였던 시절에는 MicroStation이나 AutoCAD와 같은 CAD 소프트웨어를 대신 사용하기도 했음
  - 요즘은 GIS 소프트웨어와 CAD 소프트웨어의 역할이 비교적 뚜렷하게 구분되고 있음

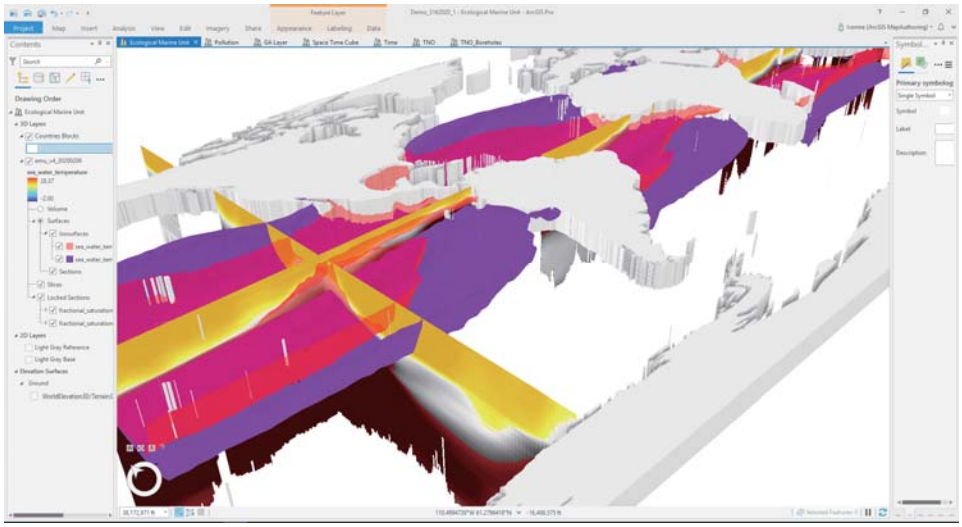
대표적인 GIS 소프트웨어

- ArcGIS 계열 제품
  - Esri社에서 개발한 상용 소프트웨어로 전세계적으로 높은 시장 점유율을 차지하고 있음
  - 데스크톱 소프트웨어인 ArcGIS Desktop, ArcGIS Pro는 물론, 클라우드 기반의 GIS 플랫폼인 ArcGIS Online도 운영하고 있음
  - 공간데이터의 생산부터 관리, 분석, 시각화에 이르는 기능들을 많이 포함하고 있으나, 높은 구매 비용이 진입 장벽으로 작용함

3. GIS 소프트웨어

대표적인 GIS 소프트웨어

- ArcGIS 계열 제품



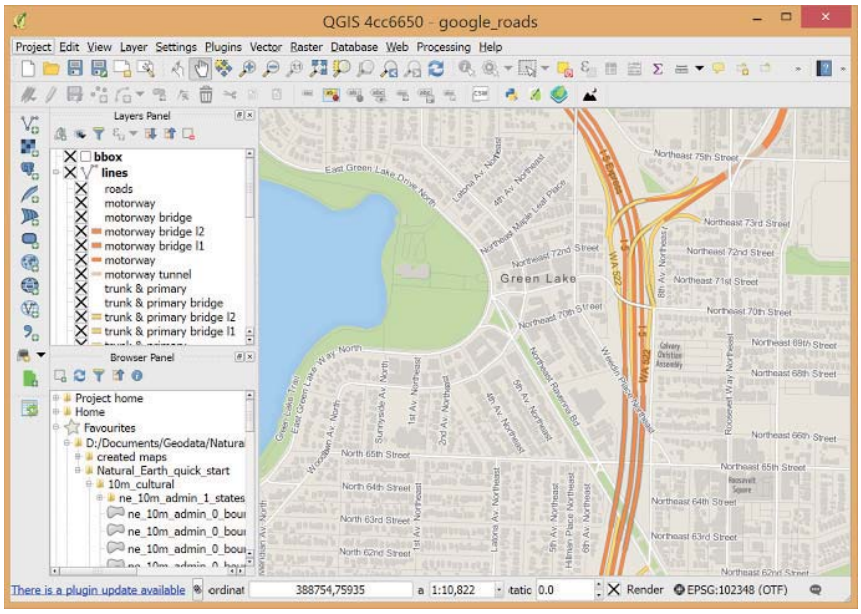
https://www.esri.com/arcgis-blog/products/arcgis/announcements/q3-2020-arcgis-release/

대표적인 GIS 소프트웨어

- QGIS
  - 대표적인 오픈소스 GIS 소프트웨어로, 윈도우는 물론 맥OS와 리눅스 등 여러 운영체제에서 사용할 수 있음
  - 최근 공간 분석 기능은 물론, 사용자 인터페이스(UI)와 안정성에서도 많은 개선이 이루어져 사용자가 증가하고 있음
  - 다양한 플러그인을 통해 기능을 쉽게 확장할 수 있고, Python 스크립트를 사용해 반복적인 작업을 효율적으로 수행할 수 있음

3. GIS 소프트웨어

대표적인 GIS 소프트웨어



https://qgis.org/en/site/about/index.html

대표적인 GIS 소프트웨어

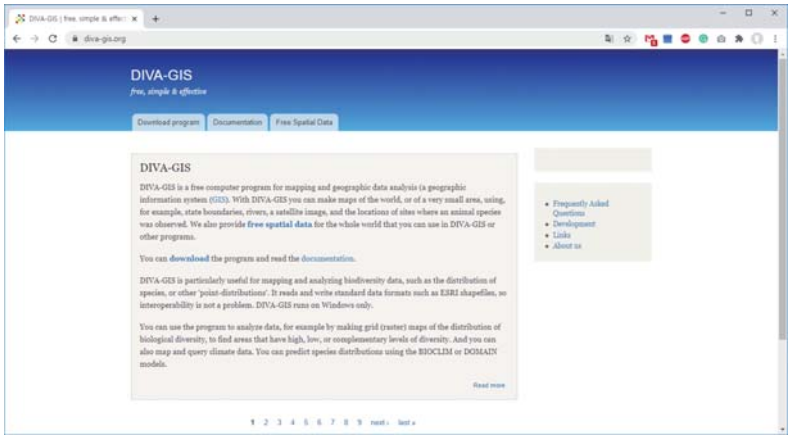
- GeoDa
  - 오픈소스 GIS 소프트웨어로 탐색적 공간데이터 분석(ESDA) 기능에 중점을 두고 개발되었음
    - 소프트웨어 다운로드: https://spatial.uchicago.edu/software
    - 소스 코드: https://github.com/GeoDaCenter/geoda/
  - 공간데이터의 생성이나 편집, 관리보다 분석과 시각화 기능에 초점을 맞춘 프로그램으로 비교적 손쉬운 사용이 가능함

대표적인 GIS 소프트웨어

- GRASS GIS
  - GRASS는 Geographic Resources Analysis Support System의 약자로, 오픈소스 GIS 소프트웨어이며 1980년대부터 개발이 시작되었음
    - 소프트웨어 다운로드: https://grass.osgeo.org/download/
    - 소스 코드: https://github.com/OSGeo/grass
  - 불편한 사용자 인터페이스 등으로 인해 2000년대 이후 사용자 수가 줄었으나, GRASS GIS에 구현된 기능들은 QGIS 툴박스나 R 패키지를 통해 여전히 활용되고 있음
    - QGIS Desktop 3.14.0 with GRASS GIS 7.8.3
    - R package `rgrass7`

도메인 GIS 소프트웨어

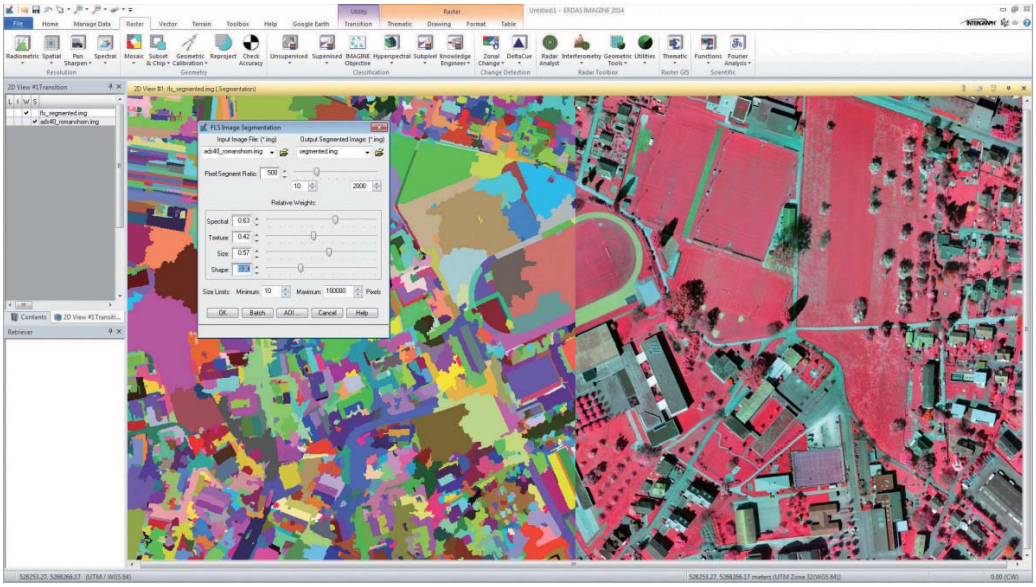
- DIVA-GIS
  - 생태학 분야의 연구자들이 주로 사용되는 GIS 소프트웨어로 무료로 내려 받아 사용할 수 있음(오픈소스는 아님)



영상 분석 소프트웨어

- ERDAS IMAGINE
  - 원격탐사 영상과 같은 래스터 데이터 가공과 분석에 특화된 소프트웨어로 영상 분류(image classification)와 같은 기능이 잘 갖춰져 있음
    - 전통적으로 사용되어 온 K-Means나 ISODATA 분류 기법은 물론, 객체 기반 영상 분류 기법과 인공지능 기반의 알고리즘까지 다양한 분류 방법이 지원됨
  - Shapefile은 물론 Esri Geodatabase도 열 수 있으며, 기본적인 벡터 데이터 가공과 편집도 가능
  - ArcGIS Pro와 유사한 리본(ribbon) 인터페이스를 갖고 있어 처음 접하는 사람도 비교적 쉽게 기능을 찾을 수 있고, ArcGIS에서 플러그인 형태로 사용할 수도 있음
  - 상용 소프트웨어로 비용 부담이 있음

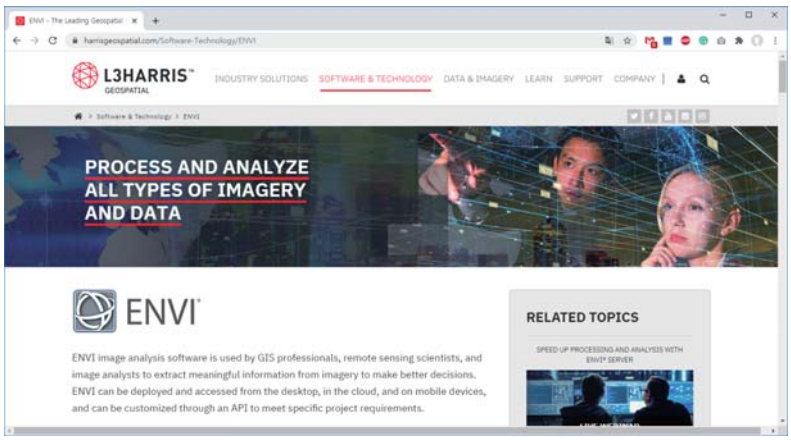
영상 분석 소프트웨어



https://www.hexagongeospatial.com/brochure-pages/erdas-imagine-brochure

영상 분석 소프트웨어

- ENVI
  - ERDAS IMAGINE과 마찬가지로 원격탐사 영상 분석에 특화된 GIS 소프트웨어



The 7th KOSTAT-UNFPA
Summer Seminar on Population

04

좌표 체계

4. 좌표 체계

1) 좌표 체계의 필요성
2) 지구의 형상과 타원체
3) 경위도 좌표계
4) 직각 좌표계
5) 대표적인 지도 투영법
6) 공간 참조 체계

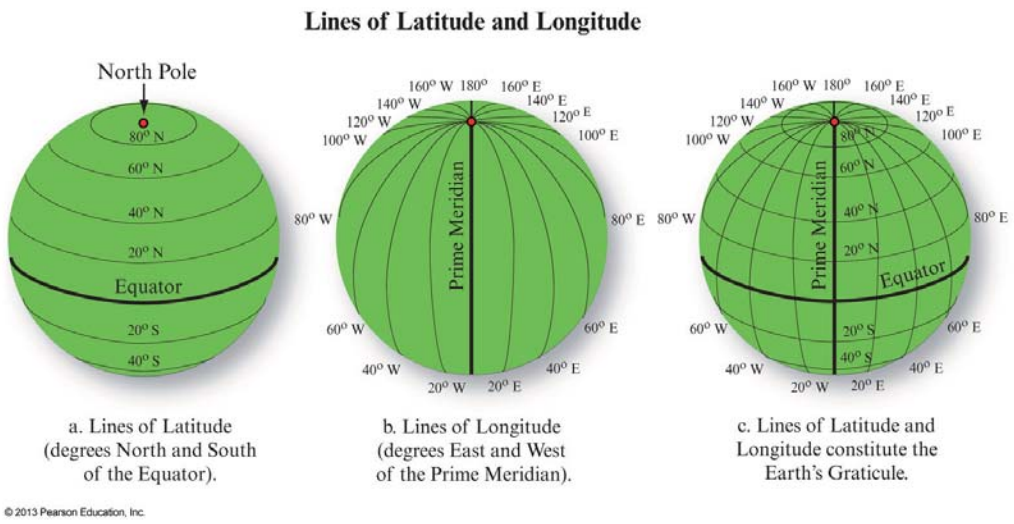
좌표 체계의 필요성

- GIS 소프트웨어에서 사용하는 모든 공간데이터는 정확한 위치 정보를 가지고 있어야 함
  - 지도 상에 표현된 객체를 현실 세계의 위치와 대응시킬 수 있는가?
- 위도, 경도와 같은 좌표를 사용해 임의의 점이 지표 위의 어느 지점에 위치해 있는지 표현할 수 있음
  - 그러나 이를 위해서는 좌표의 원점과 거리 단위, 방향 등을 모두 알아야 함 ➔ 좌표 체계의 필요성

Point ID	x	y
1	24.4	58.3
2	26.1	58.0
3	24.7	57.1
4	26.8	57.2

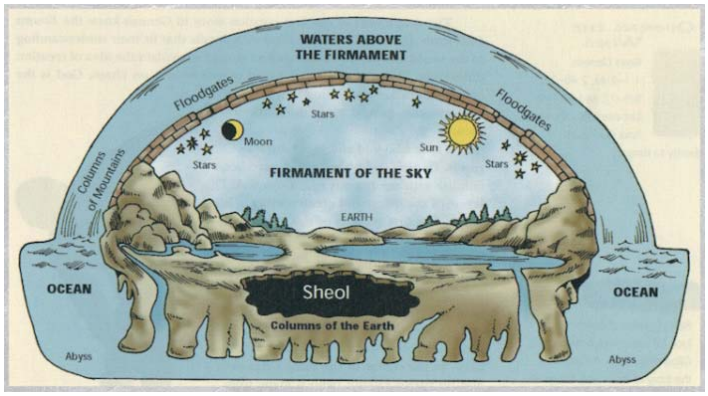
좌표 체계의 필요성

- 위도와 경도는 이미 정해진 것이 아닌가?

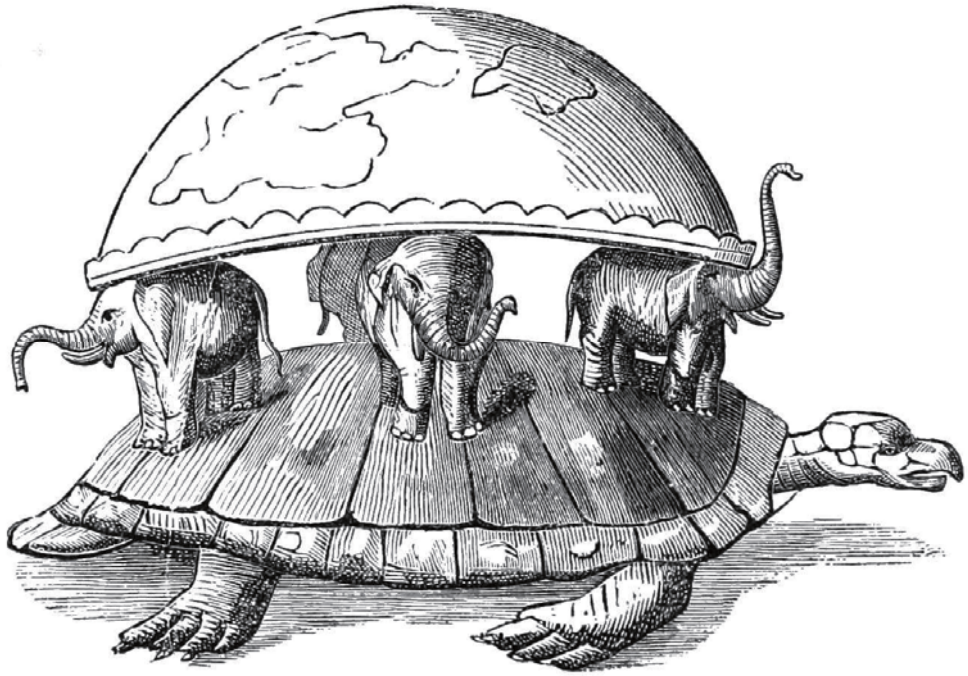


지구의 형상

- 지구의 형상을 어떻게 정의하는지에 따라, 같은 경위도가 다른 장소를 나타낼 수도 있음
- 지구는 어떻게 생겼을까?



http://www.testingtheglobe.com/images/EnclosedEarth1.jpg



https://upload.wikimedia.org/wikipedia/commons/c/c0/PSM_V10_D562_The_hindoo_earth.jpg

지구의 형상

- 고대에는 지구가 평평하다고 믿는 사람들이 있었음
  - 물론 지금도 ...
- 인공위성에서 본 지구의 모습
  - 그렇다면 지구는 완전한 원형일까?
  - 뉴턴은 지구 자전의 효과로 지구가 남북 축의 반경보다 동서 축의 반경이 긴 타원체의 형상을 나타낸다고 주장!

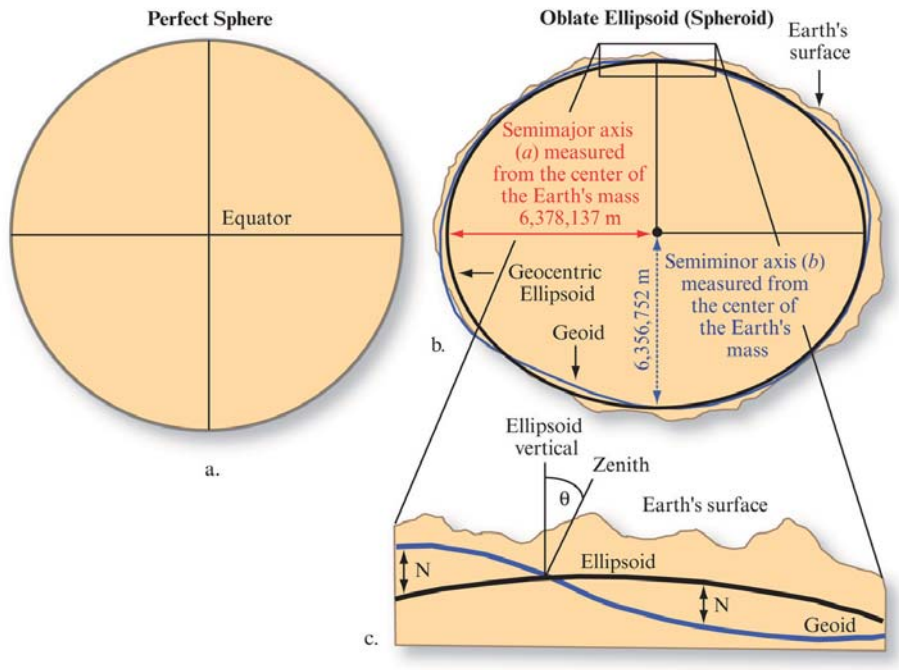
지구의 형상

- 지구 타원체(Earth ellipsoid)
  - 적도의 반경과 편평률(flattening)로 정의되는, 타원체 모양을 한 가상의 지구
  - 편평률:

$$f = \frac{(a-b)}{a}$$

a는 적도의 반경(장축), b는 극 반경(단축)

  - 지구가 완전한 기하학적 타원체가 아니기 때문에 학자들에 따라 적도 반경과 극 반경을 바탕으로 산출되는 타원체의 편평률은 약간씩 다르며 국가마다 사용하는 편평률에도 차이가 있음

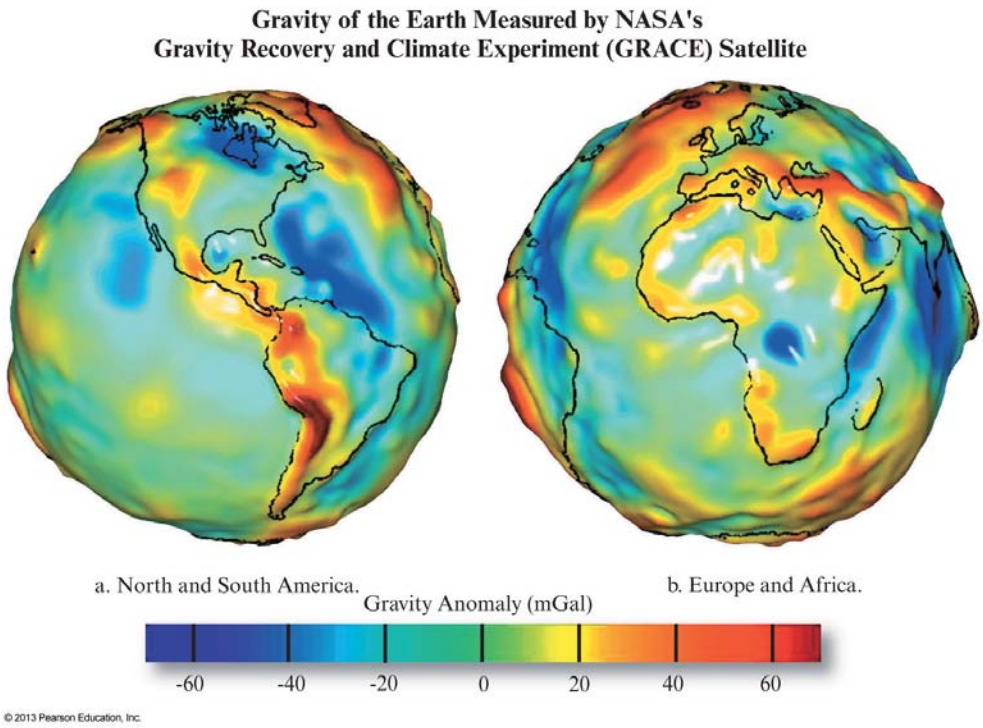


© 2013 Pearson Education, Inc.

타원체 이름	적도 반경(m)	편평률	주 사용국가
Everest (1830)	6,377,276	1/301	인도
Bessel (1841)	6,377,397	1/299	일본, 독일, 한국
Airy (1844)	6,377,563	1/299	영국
Clarke (1866)	6,378,206	1/295	북아메리카
Clarke (1880)	6,378,249	1/293	프랑스, 남아프리카
Hayford (1909)	6,378,388	1/297	북아프리카, 유럽
International (1924)	6,378,388	1/297	국제적으로 채택
Krasovsky (1938)	6,378,254	1/298	러시아
GRS67 (1967)	6,378,160	1/298	남아메리카, 호주
WGS72(1972)	6,378,135	1/298	미국
GRS80 (1980)	6,378,137	1/298	국제적으로 채택
WGS84	6,378,137	1/298	국제적으로 채택

지구의 형상

- 지오이드(Geoid)
  - 실제 지구에 작용하는 중력을 나타낸 물리적인 표면을 의미
    - 조석이나 파도가 없는 상태의 평균 해수면을 말함
    - 육지의 경우 해수면이 없기 때문에, 육지와 해수면이 닿는 부분부터 터널을 팠다고 가정했을 때 흘러 들어온 해수면의 높이가 지오이드(Geoid) 면이 됨
    - 만유인력에 의해 일반적으로 바다에서는 지오이드 면이 지구 타원체보다 낮은 반면, 대륙에서는 높게 나타남
  - 인공위성을 통한 중력 탐사를 바탕으로 지오이드 구축
    - 지각의 구성요소와 밀도가 장소에 따라 다르기 때문에 중력의 세기 또한 다르게 나타남



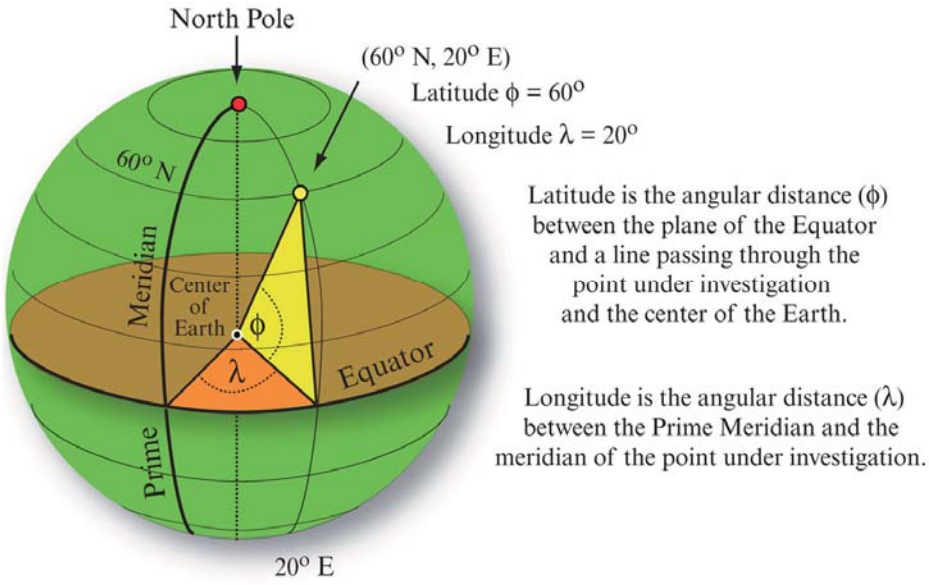
준거 타원체

- 준거 타원체(reference ellipsoid)
  - 지표상의 특정 지점에서 지구 타원체와 지오이드 면 간의 차이가 최소화 될 수 있도록 만든 타원체
    - 특정 지역 또는 국가에서 채택한 준거 타원체는 국제적 표준이 되는 지구 타원체와는 다를 수 있음
  - 서로 다른 타원체 사용에 따른 문제점을 줄이기 위해 점차 세계적으로 WGS84(World Geodetic 1984) 타원체를 사용하는 추세
  - 우리나라도 기존 베셀(Bessel) 타원체에서 2007년 1월 1일부터 WGS84와 비슷한 GRS80 체제로 전환을 법제화

경위도 좌표계

- 경위도 좌표계의 원점은 본초 자오선(Prime Meridian)과 적도(Equator)가 만나는 곳으로 정의함
- 위도(latitude)는 지구 중심으로부터 적도의 남쪽, 또는 북쪽 지점 간의 각도로 정의함
  - 적도의 위도를 0°, 북극과 남극은 각각 +90°와 -90°가 됨
- 경도(longitude)는 본초 자오선(Prime Meridian)을 기준으로 동서 방향 각도로 정의함
  - 본초 자오선의 경도는 0°, 동쪽과 서쪽으로 각각 180°와 -180°까지 값을 가짐

Determining Latitude and Longitude



© 2013 Pearson Education, Inc.

경위도 좌표와 타원체

- 베셀(Bessel)이 1841년에 지구의 크기와 형상을 산출한 타원체
  - 장반경(적도 반경)은 아래와 같이 약 6,377,397 m, 단반경(극 반경)은 약 6,356,075 m로 편평도는 약 1/299.1528
  - 우리나라의 경우 오랫동안 베셀 타원체를 사용했음
- Google Earth 등에서 사용되는 WGS84 타원체와는 아래 표와 같이 다소 차이가 있음

	베셀 타원체	WGS84 타원체
적도 반경	6,377,397.155 m	6,378,137.0 m
극 반경	6,356,078.963 m	6,356,752.3 m
편평도	1 / 299.15281535	1 / 298.25722356

http://en.wikipedia.org/wiki/Bessel_ellipsoid

경위도 좌표와 타원체

- 오래 전에 블로그에서 찾은 글:

서울서 둘째로 잘하는집 | 먹다죽은귀신때깔굿 2005/05/18 10:56

http://blog.naver.com/jorpen56/100013033898 본사

제목 : [서울 - 삼청동] 서울서 둘째로 잘하는집
이름 : 발꼬락
게시일 : 2003-05-21 17:17:58
조회 : 3938
추천 : 1

삼청동 먹자골목주변은 깨끗하고 조용하고 오래된 음식점들이 많습니다.
인도가 좁아 겨우 두사람이 발 맞추어 갈 수 있는 정도죠.
경복궁이나 삼청공원에 산책삼아 가셨다가 나오는 길에 들러 보시면 아주 정겹게 먹어 볼 수 있는 찻집을 소개합니다.

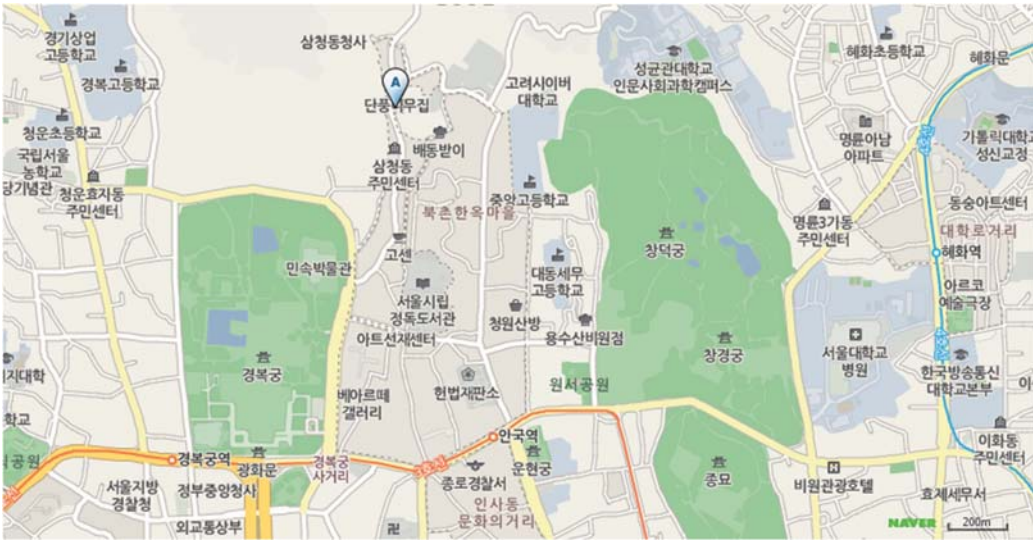
간판부터 궁금하게 만듭니다.
첫째도 아니고 왜 굳이 둘째로 맛있는 집이라고 했을까요?
세상에서 제일 맛있는 음식은 임금님께로 가고 두번째로는 이집이 제일 맛있다는 의미랍니다.
아담한 크기에 맛을 즐기러 오신 분들이 옹기 종기 모여 앉아 기다립니다.
벽은 한지로 꾸며두었고 마치 시골 간이역에 있는 다방처럼 의자도 유행이 이미 지나버린 쇼파들이 앉아 있습니다.
좁은 공간인데도 조용한 걸 보니 모두들 아주 진지하고 차분하게 음식을 기다립니다.

네비게이션 경위도 좌표(BESSEL) : E 126°59'02" N 37°35'02"

http://blog.naver.com/jorpen56?Redirect=Log&logNo=100013033898

경위도 좌표와 타원체

- 네이버 지도에서 해당 경위도 좌표로 검색한 결과:



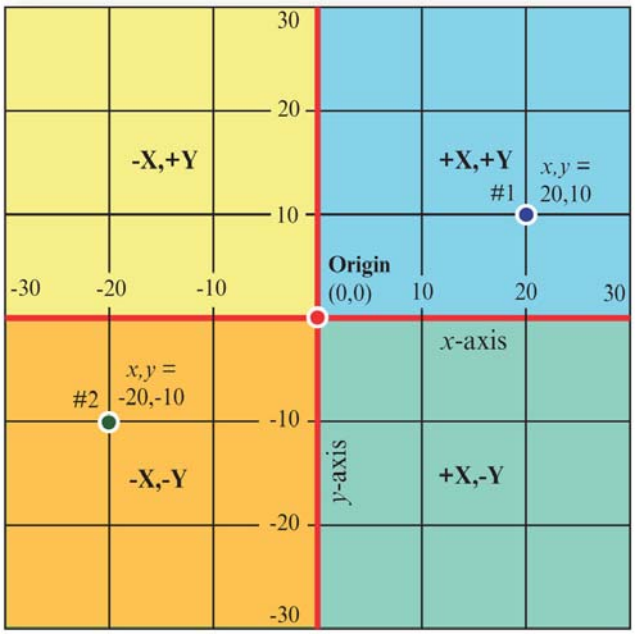
경위도 좌표와 타원체

- 같은 좌표를 Google Earth에서 검색한 결과:

직각 좌표계

- 평면상의 모든 점에 대해 2개의 좌표를 부여하며, 좌표가 원점 (0, 0)으로부터 x축과 y축 방향으로의 거리
  - 좌표 원점을 정하여 면상에서 (x, y) 미터로 위치 표시
- 이해하고 적용하기 수월하나, 측량 범위가 넓지 않을 때에만 사용
- 좌표가 양수와 음수를 모두 가질 수 있음
  - 경우에 따라 연구 지역의 범위가 양의 범위에만 놓이도록 강제적으로 가산값을 부여하기도 함
  - 동쪽 방향 가산값(false-easting)과 북쪽방향 가산값(false-northing)을 사용

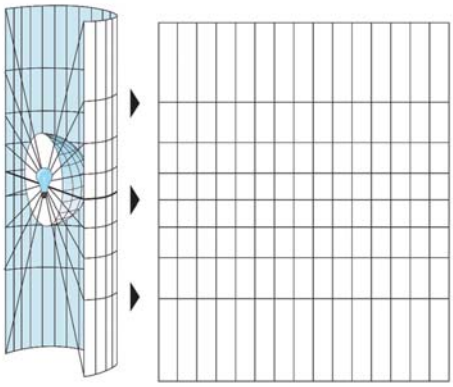
A Cartesian Coordinate System



© 2013 Pearson Education, Inc.

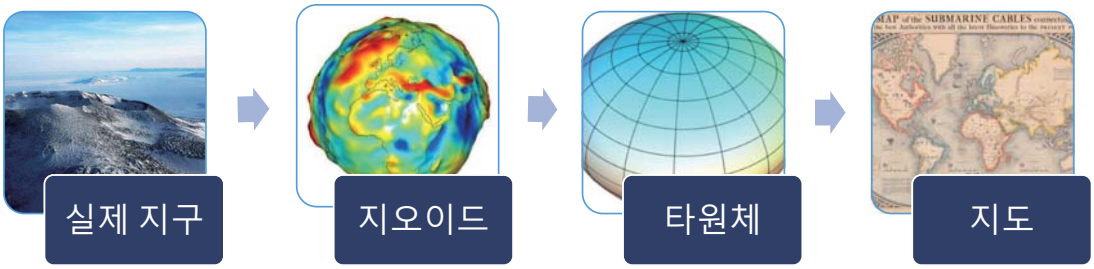
지도 투영법의 개념

- 지도 투영이란 3차원의 타원체에 그려진 경위도 선을 평면(2차원)의 지도에 나타낼 수 있도록 체계적으로 변환하는 것을 말함
  - 개념적으로는 경위도 좌표가 그려진 투명한 지구본을 광원으로 투시하여 투영면에 비춰진 그림자로 지도를 만드는 것



지도 투영법의 개념

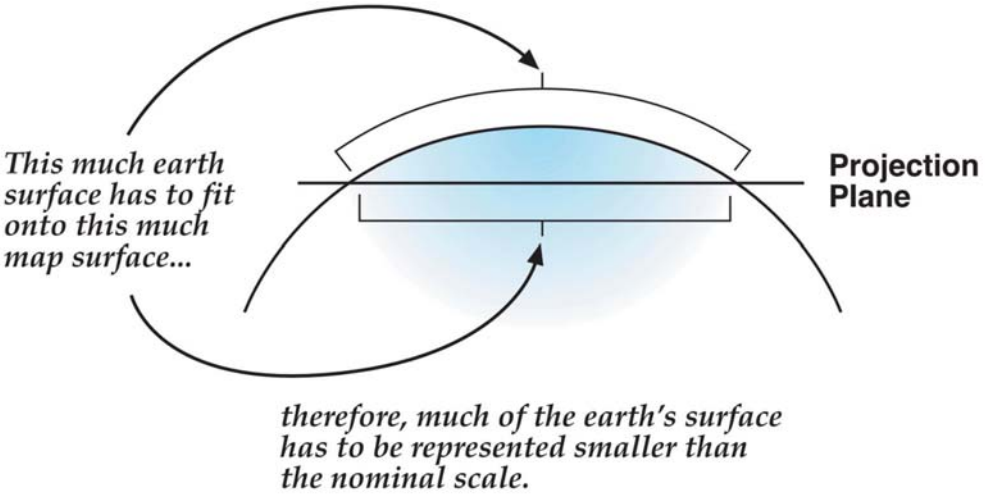
- 실제 지구를 지도로 나타내는 전체 과정은 다음과 같이 도식화 할 수 있으며, 투영법은 타원체를 지도로 옮기는 과정에서 사용됨



지도 투영법의 문제점

- 지표면(3차원)에 위치한 객체의 형상과 면적, 거리, 방향 등을 2차원의 지도에서 그대로 정확하게 나타내는 것은 불가능함
- 지도 제작 시에 고려해야 하는 요소
  - 정형성(지도 상에서의 모양이 지표면 상의 실제 모양과 동일한가?)
  - 정적성(지도 상에서의 면적이 지표면 상의 실제 면적과 같은 비례로 나타나는가?)
  - 정거성(지도 상에서 두 지점들 간의 거리가 지표면과 같은 관계를 유지하는가?)
- 지구본과 같은 3차원의 구체(또는 타원체)를 평면으로 투영하게 되면 이러한 특성이 왜곡되어 나타남

지도 투영법의 문제점



지도 투영법의 선택

- 왜곡이 없는 지도를 제작하는 것은 불가능하기 때문에 지도를 제작하기에 앞서 사용 목적을 고려하여 어떠한 속성을 유지시킬지 판단하는 것이 필요
- 목적과 관심 지역에 따라 다양한 투영법이 사용되고 있음
  - ArcGIS 10.8을 기준으로 기본 지원하는 투영법은 모두 72개 (http://desktop.arcgis.com/en/arcmap/latest/map/projections/list-of-supported-map-projections.htm)

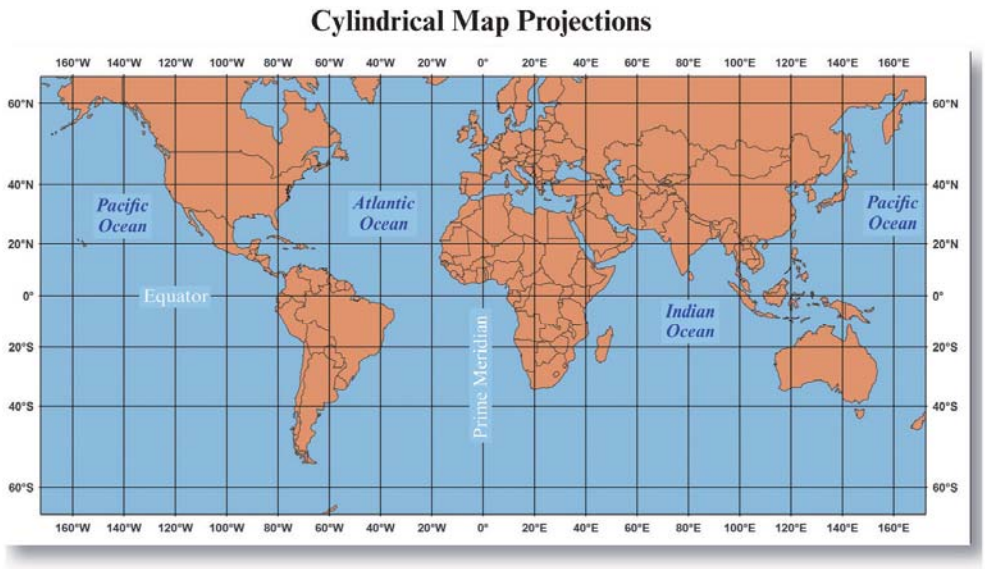
지도 투영법의 분류

- 투영 방법에 따른 분류
  - 투영면 기준을 기준으로 원통(cylindrical), 원추(conic), 방위(azimuthal)
  - 투영면의 축 기준을 기준으로 정축(polar aspect), 횡축(equatorial aspect), 사축(oblique aspect)
  - 광원의 위치를 기준으로 심사(gnomonic), 평사(stereographic), 정사(orthographic)
- 왜곡 유형에 따른 분류
  - 정각(conformal), 정적(equal-area), 정거 (equidistant)

대표적인 지도 투영법

- 메르카토르(Mercator) 도법
  - 네덜란드의 지도학자 게르하르두스 메르카토르(Gerardus Mercator, 1512-1594)가 1569년 발표한 지도 투영법
  - 경선의 간격은 고정되어 있으나 위선의 간격을 조절하여 각도 관계가 정확(정각)하도록 설정
  - 적도에서 멀어질수록 면적(축척)이 크게 왜곡되기 때문에 위도 80-85˚ 이상의 지역에 대해선 일반적으로 사용하지 않음
  - 지도 상 임의의 두 지점을 직선으로 연결하면 항정선과 같아지기 때문에 항해용 지도로 많이 사용됨

대표적인 지도 투영법



a. Mercator conformal map projection.

© 2013 Pearson Education, Inc.

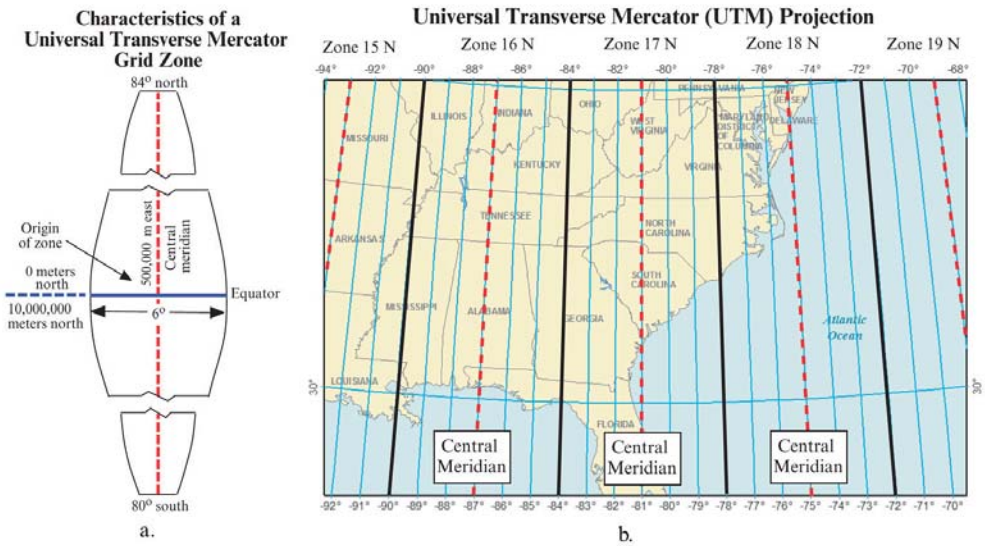
대표적인 지도 투영법

- 횡축(Transverse) 메르카토르 도법
  - 적도 대신 지구본을 옆으로 뉘어서 투영하는 메르카토르 도법
  - 지구를 원형으로 가정한 상태에서의 투영법은 요한 람베르트(Johann Lambert)가 1772년 개발하였음
  - 타원체에서의 투영법은 여러 가지 종류가 있으나, 우리나라에서 사용하는 1:50,000 지도와 UTM 좌표계는 가우스 크뤼거(Gauss-Krŭger)의 횡축 메르카토르 도법을 이용해서 제작함
  - 좁은 경도대에서는 축척의 증가가 매우 작기 때문에 정각성이 뛰어난 대축척지도에서 유용하게 쓰임

대표적인 지도 투영법

- UTM 좌표계(Universal Transverse Mercator Coordinate System)
  - 지구를 경도 6° 간격의 60개 세로 띠로 나누어 횡축 메르카토르 도법으로 그린 뒤, 위도 8° 간격으로 다시 20개의 격자로 구분하는 좌표계
  - 각각의 세로 띠는 북위 84°부터 남위 80°까지 8° 간격으로 나누어지나, 가장 북쪽의 격자(북위 72°-84°)는 12°로 분할됨
  - 각 세로 구역마다 설정된 원점에 대해서 종, 횡 좌표로 위치 표현
  - 종(y) 좌표는 대상점의 적도로부터의 투영 거리, 횡(x) 좌표는 대상점이 속한 구역의 중앙에 위치한 경선으로부터 대상점까지의 투영 거리
  - 좌표는 미터로 표시하며, 좌표가 음수로 표시되는 것을 방지하기 위해 횡좌표에는 500,000m를 더하여 표시하며, 남반구의 종좌표에는 10,000,000m를 더하여 표시함

대표적인 지도 투영법

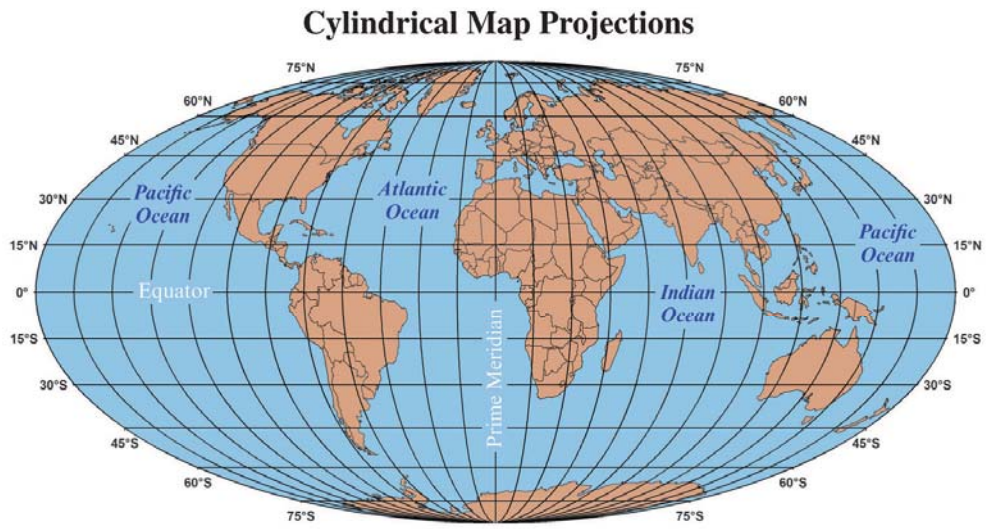


© 2013 Pearson Education, Inc.

대표적인 지도 투영법

- 몰바이데(Mollweide) 도법
  - 독일의 카를 몰바이데(Karl Mollweide)가 1805년 개발한 정적도법
  - 1650년에 고안된 시뉴소이드(Sinusoidal) 도법을 보완한 것으로, 경선들은 등간격으로 극에서 수렴하는 반타원형으로 표현됨
  - 중앙 경선을 제외한 모든 경선들은 타원의 호 형태를 갖고 있음
  - 위선들은 적도와 평행한 직선으로 표현되는데 위선 간격은 극으로 갈수록 실제 지표보다 좁게 나타나게 하여 정적성을 유지함
  - 두 위선과 두 경선에 의해 이루어지는 면적이 지표상에서와 같도록 정적성을 유지하기 위해 위선의 간격을 조정
  - 고위도의 대륙과 바다에 분포도를 표현하기에 적합한 투영 방법

대표적인 지도 투영법



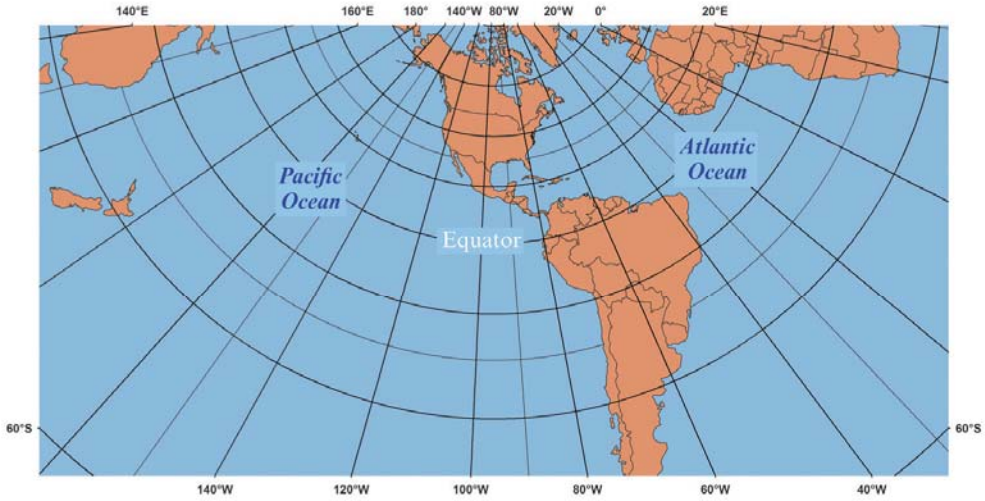
a. Mollweide pseudocylindrical equal-area.

© 2013 Pearson Education, Inc.

대표적인 지도 투영법

- 람베르트(Lambert)의 정형(각)원추도법
  - 경선은 극으로 수렴하는 등간격의 직선이며, 위선은 동심원의 호로 나타남
  - 위선 간격을 조정하여 정각성을 유지하도록 고안되었으며, 표준 위선 사이와 부근에서는 형상과 면적의 왜곡이 최소화됨
  - 동서 방향의 범위를 갖는 국가나 지역의 지도를 제작하는데 사용
- 알버스(Albers)의 정적원추도법
  - 경선은 등간격으로 극에 수렴하는 방사상의 직선으로 나타나며, 위선은 람베르트 정형원추도법과 마찬가지로 동심원의 호로 나타남
  - 람베르트 정형원추도법과는 격자선의 간격이 다름
  - 위선 간의 간격이 경선 간의 축척의 변화를 상쇄하도록 조정함으로써 정적성을 유지시킴

대표적인 지도 투영법

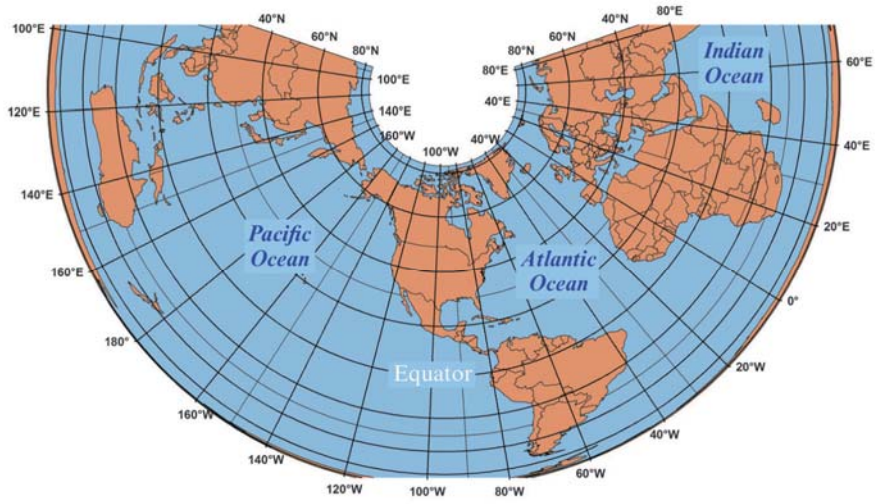


b. Lambert Conformal Conic.

© 2013 Pearson Education, Inc.

대표적인 지도 투영법

Conical Map Projections



a. Albers Equal-area Conic.

© 2013 Pearson Education, Inc.

공간 참조 체계

- 타원체, 좌표 체계(직각/경위도), 투영법이 조합되어 하나의 공간 참조 체계(spatial reference system; SRS)를 구성함
  - 공간 참조 식별자(spatial reference identifiers; SRID), 또는 좌표 참조 체계(coordinate reference systems; CRS)라고도 함
- EPSG에 등록된 SRS는 6,000개 이상이며 Korea로 검색해서 찾을 수 있는 SRS도 40여 개에 달함(https://epsg.io/)
  - 우리나라에서 사용되는 SRS의 구체적인 정의는 다음 링크에서 확인할 수 있음: https://www.osgeo.kr/17

공간 참조 체계

- 언젠가 한 번은 마주칠 수 있는 공간 참조 체계와 EPSG 번호…

분류	EPSG	이름	비고
전 지구 대상	4326	WGS84 경위도	GPS
	4004	Bessel 1841 경위도	
	4019	GRS80 경위도	WGS84와 유사
	3857	Google Mercator	Google Map
UTM 계열	32652	UTM52N	
	32651	UTM51N	
KATEC 계열	5178	UTM-K (Bessel)	새주소지도
	5179	UTM-K (GRS80)	통계청

공간 참조 체계

- 언젠가 한 번은 마주칠 수 있는 공간 참조 체계와 EPSG 번호...

분류	EPSG	이름	비고
국토지리정보원(Bessel)	2096	동부원점	
	2097	중부원점	
	2098	서부원점	
	5173	보정된 서부원점	2002년 이전에 주로 사용
	5174	보정된 중부원점	
	5175	보정된 제주원점	
	5176	보정된 동부원점	
	5177	보정된 동해(울릉)원점	

공간 참조 체계

- 언젠가 한 번은 마주칠 수 있는 공간 참조 체계와 EPSG 번호...

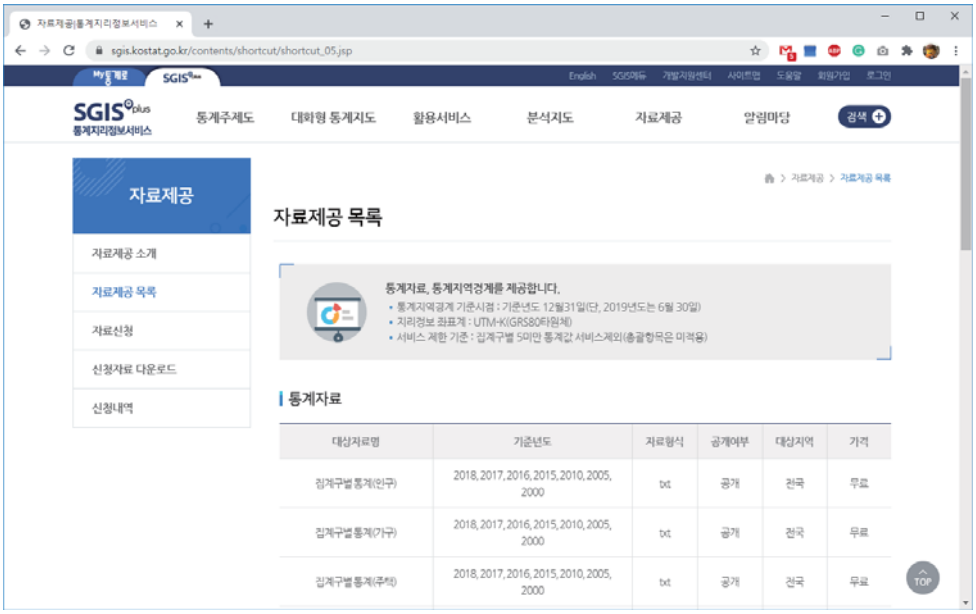
분류	EPSG	이름	비고
국토지리정보원(GRS80)	5180	서부원점-falseY:50000	
	5181	중부원점-falseY:50000	
	5182	제주원점-falseY:50000	
	5183	동부원점-falseY:50000	
	5184	동해(울릉)원점-falseY:50000	
	5185	서부원점-falseY:60000	2002년 이후에 사용
	5186	중부원점-falseY:60000	
	5187	동부원점-falseY:60000	

공간 참조 체계

- 언젠가 한 번은 마주칠 수 있는 공간 참조 체계와 EPSG 번호...

분류	EPSG	이름	비고
	5188	동해(울릉)원점-falseY:60000	
전 지구 대상	4166	Korean 1995	WGS84
	4162	Korean 1985	Bessel 1841
	4737	Korean 2000	GRS80

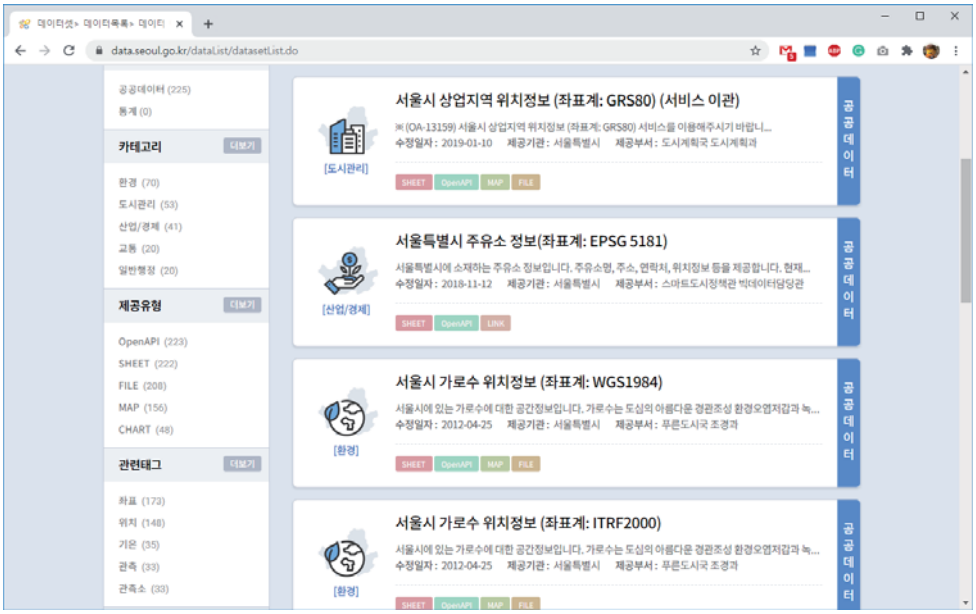
공간 참조 체계



공간 참조 체계

데이터 유형		공간(벡터)		
데이터 설명		연속지적도형정보를 기반으로 건물 공간정보와 건축행정시스템(세움터)의 건축물대장 속성정보를 건물단위로 통합하여 구축한 공간(도지)기반의 건물통합정보		
생산주체	관리기관/부서	국토교통부 / 공간정보제도과	시스템명	부동산종합공부시스템
	생산기관/부서	시군구 / 지자체	갱신주기	1년 상시
공간정보	타원체	Bessel	투영법	횡메르카토르투영법(TM proj.)
	좌표계	평면직각좌표계	객체형태	면
데이터 건수		14,391,694	데이터 용량 (MB)	6,605
취합주체	관리기관/부서	국토교통부 / 국가공간정보센터	시스템명	국토정보시스템
	수집주기	변경발생시	수집방법	연계
구축	지리적 범위	전국	시간적 범위	최종버전(갱신일 기준)
공개	개방시스템	국가공간정보포털		
	배포 데이터 좌표계	Bessel/TM, EPSG:5174	배포 데이터 포맷	SHP
유통	가격정책	무상공급	판매방법	웹 다운로드 서비스, API 서비스
비고				

공간 참조 체계



The 7th KOSTAT-UNFPA
Summer Seminar on Population

05

근접성 분석과 중첩 분석

5. 근접성 분석과 중첩 분석

1) 공간 분석이란?
2) 근접성 분석과 버퍼
3) 중첩 분석

공간 분석이란?

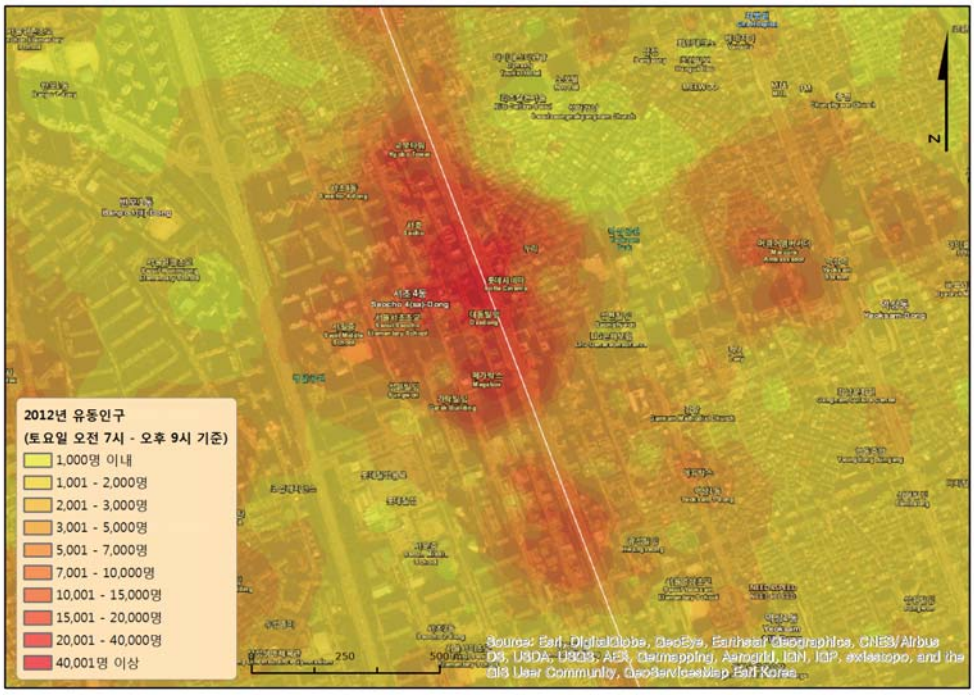
- 지리적인 사상(事象)의 공간적 분포와 다른 사상과의 관계 등을 살펴봄으로써 새로운(필요한) 정보를 추출하는 과정
  - 1990년대 후반 이후 탐색적 공간데이터 분석(exploratory spatial data analysis; ESDA)의 필요성이 강조되면서, 사회과학 분야에서도 GIS의 중요성이 커짐
- 인구 관련 공간 분석의 예:
  - 생활인구 데이터를 활용한 노인인구 공간적 분포 및 군집분석: 서울시를 중심으로(이지혜·김형중, 2019)
  - 공간통계 기법을 이용한 현주인구 추정 모델링(이건학·김감영, 2016)
  - 공간계량기법을 이용한 학령별 인구의 공간적 분포 및 지역특성 영향요인 연구(김리영·서원석, 2016)
  - 부산시 고령인구의 공간적 분포 변화(이유미·구동회, 2012)

공간 분석의 시작

- 공간 분석 수행에 앞서 생각해야 하는 것 두 가지:
  1. 무엇을 분석할 것인가?
     - 분석의 목적을 정확하게 이해하는 것이 필요함
     - 분석을 통해 답하고자 하는 문제가 무엇인지 매우 구체적으로 적을 수 있어야 함
  2. 어떻게 분석할 것인가?
     - 문제에 대한 답을 찾기 위한 과정을 세분화하고 각각의 단계에 맞는 분석 방법을 선택하는 것이 필요함
     - 실제 분석을 시작하기 전 충분한 시간을 들여 필요한 데이터를 최대한 확보하는 것이 중요

공간 분석과 데이터

- 새로 개점하는 커피 전문점의 위치는 어디가 되어야 할까?
  - 용도지역(주거, 상업 ...)
  - 기존의 커피 전문점 위치
  - 유동인구 데이터
  - 배후지역 상주인구의 연령대, 소득 수준 등
- 실제 우리가 수행할 수 있는 분석의 범위와 수준은 사용 가능한 데이터, 그리고 분석 능력에 많은 영향을 받게 됨



근접성 분석

- 근접성 분석의 예:
  - 우리 집에서 가장 가까운 지하철역은 어디일까?
  - 소방서를 중심으로 반경 1 km 이내에 몇 개의 주유소가 있을까?
  - 하천이 범람했을 때 영향을 받게 되는 가옥은 모두 몇 채일까?
- 같은 동네에 위치한 A 고등학교와 B 고등학교, 과연 어느 학교의 입지가 더 좋을까?
  - 입지가 좋은 학교는 학교 주변에 유해업소가 적은 학교라 정의
  - 학교 주변을 정의하는 면과 유해업소를 나타내는 점(포인트)의 중첩 분석을 수행할 수 있음 ➔ 공간적 조인(spatial join)
  - 그런데 학교 주변은 어떤 기준으로 정의할 수 있을까? ➔ 도메인 지식
  - 해당 기준을 GIS에서 어떻게 나타낼 수 있을까? ➔ 버퍼(buffer)

근접성 분석

- 학교환경위생정화구역
  - 학교의 보건·위생 및 학습환경보호를 위하여 학교 주변에 학교보건위생에 지장이 있는 행위 및 시설을 제한한 지역
- 설정 범위

근접성 분석

- 다음과 같이 고등학교와 여러 음식점이 분포해 있다고 할 때:

근접성 분석

- 반경 200미터의 버퍼를 생성

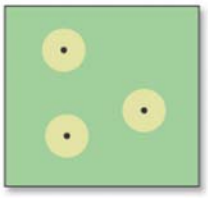
근접성 분석

- A 학교가 더 좋은 학교로 볼 수 있음

버퍼의 개념

- 버퍼(buffer)란 특정한 객체를 중심으로 일정한 폭을 가지는 구역을 말함
  - 근접성 분석을 할 때 관심 대상이 되는 지역(근접한 지역)의 경계를 설정하는 역할을 할 수 있음

점 버퍼

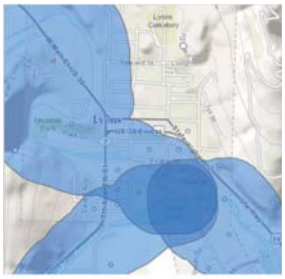


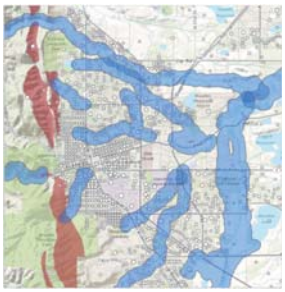
선 버퍼

면 버퍼

버퍼의 활용

- 버퍼는 중첩과 함께 GIS의 주요 기능 중 하나로 매우 폭넓게 사용되고 있음
- 버퍼의 활용 예:
  - 천연자원, 또는 특정한 동식물의 보호와 관리를 목적으로 개발제한구역을 설정하는 경우
  - 하천을 중심으로 버퍼를 산출하여 홍수와 같은 자연재해로부터 취약한 지역을 파악하고, 이를 통해 재해 예방과 피해 최소화에 기여할 수 있음





버퍼의 활용*

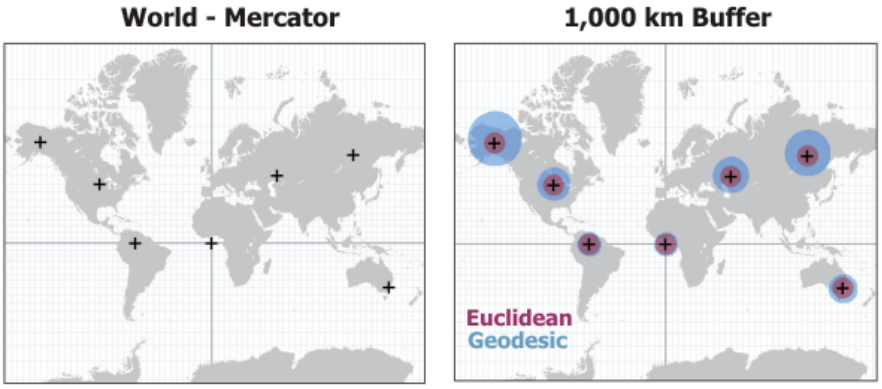
- 영어로 된 것도 한 번 읽어 봅시다!
  - Buffering of fire station and hydrant on the map gives potential service of current situation. When two buffers of hydrant actually overlap, in that area, the two hydrants can be used in case of fire. If some area on the map is not included in any fire service buffer, this area is potentially vulnerable to fire hazard.
  - If you set your retail store as a point in the map and setting several ring buffers (say 1 km, 5 km and 10 km), you may distinguish your potential customer by distance.
  - Setting a centerline of a road and setting buffer equivalent to width of the road may give the full road width. This operation could be useful to determine the required land acquisition of road widening.

버퍼와 거리의 측정

- 거리 측정의 방식:
  - PLANAR 방식은 지점 간의 직선거리(유클리드 거리)를 기반으로 버퍼를 생성하게 됨
  - GEODESIC 옵션을 사용하는 경우 지오이드 상에서의 거리(측지 거리)를 바탕으로 버퍼를 생성함
- 버퍼의 크기가 크지 않은 경우 PLANAR 방식과 GEODESIC 방식의 구분에는 큰 의미가 없을 수 있지만, 다음의 예와 같이 매우 큰 버퍼를 만드는 경우에는 선택에 주의가 필요함

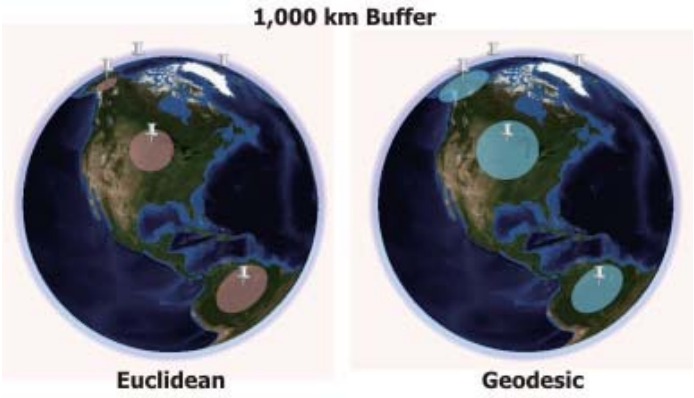
버퍼와 거리의 측정

- 거리 측정 방법에 따른 버퍼의 변화(2차원 지도를 통해 확인했을 때의 결과를 보면 유클리드 거리 기반의 버퍼가 보다 일정한 크기를 갖는 것으로 보이지만 ...)



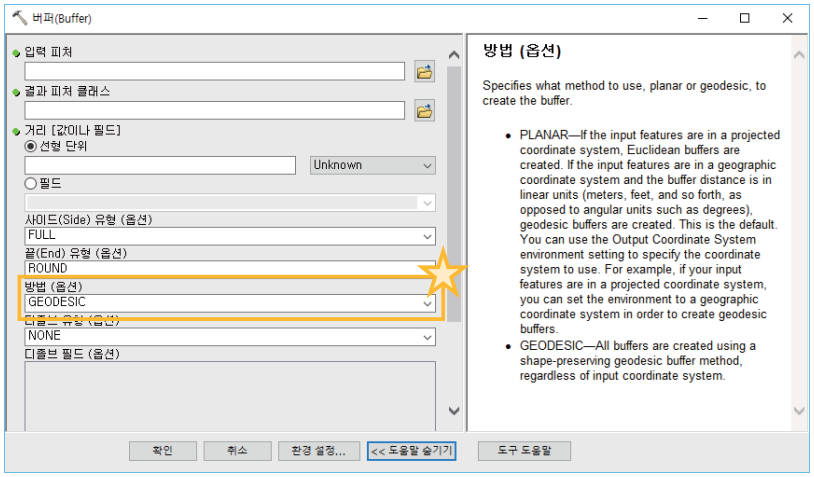
버퍼와 거리의 측정

- 거리 측정 방법에 따른 버퍼의 변화(지구본을 통해 확인했을 때의 결과를 보면 GEODESIC이 보다 일정함을 알 수 있음)



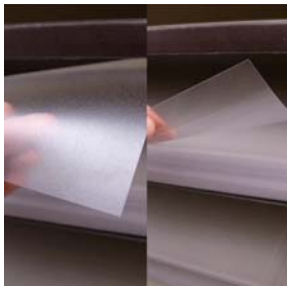
버퍼와 거리의 측정

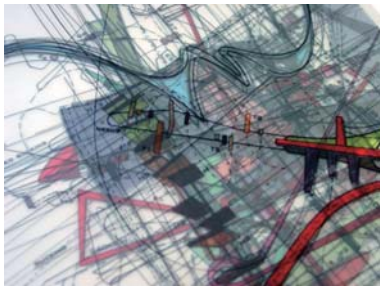
- QGIS는 기본적으로 PLANAR 기준으로 버퍼를 생성함
- ArcGIS는 다음과 같이 PLANAR와 GEODESIC 중 선택이 가능함

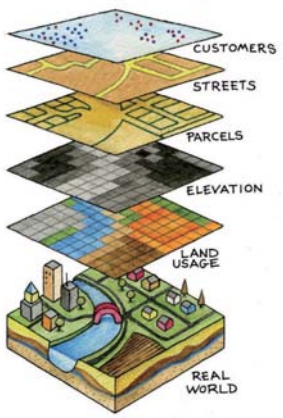


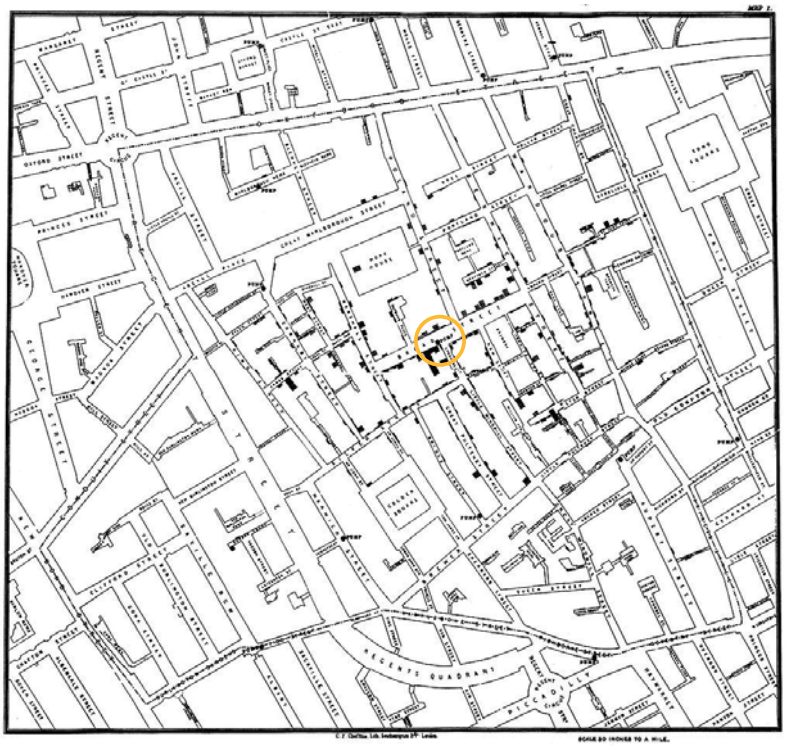
중첩의 개념

- GIS의 분석 기능 중 가장 중요한 기능 가운데 하나
  - 지도 위에 또 다른 지도를 올려놓고 두 지도에 나타난 사상들 간의 관계를 (주로 시각적으로) 살펴보는 방법
  - 현대적인 지리정보시스템의 등장 이전부터 투명한 종이 등을 사용하여 중첩 분석 수행
  - John Snow의 콜레라 지도가 대표적인 예



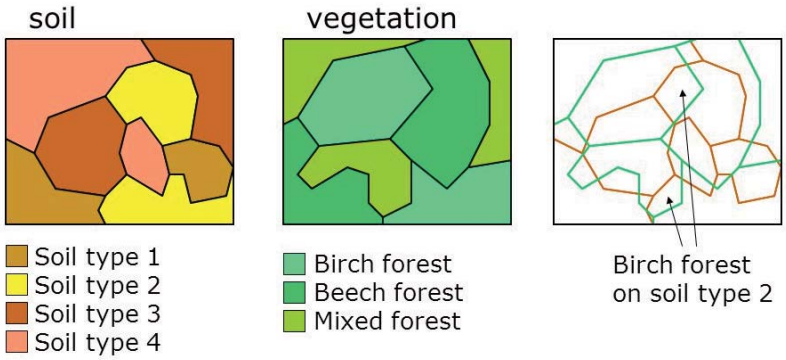




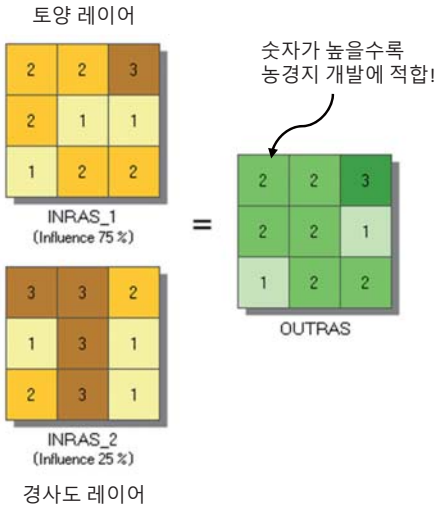


중첩의 활용

- 중첩을 통해 공간상에서 대응관계를 갖고 있는 사상들 간의 관계, 특히 인과관계를 간접적으로 살펴볼 수 있음
  - 탐험적 공간자료분석(ESDA)의 예
  - 토양 레이어와 식생 레이어를 중첩하여 자작나무가 특히 잘 자라는 토양을 살펴볼 수 있음



중첩의 활용



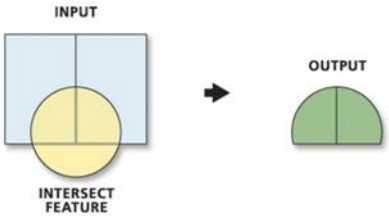
- 여러 개의 공간데이터(레이어)를 조합하여 새로운 정보를 추출할 수도 있음
  - 농경지 개발에 가장 적합한 구역을 찾기 위해 토양과 경사도 데이터를 중첩
  - 농경지 개발에는 토양이 경사도보다 중요하므로 0.75의 가중치를 부여하고, 경사도에는 가중치 0.25를 부여함
  - 다른 예로, 인구 수와 연령대, 소득수준 데이터를 중첩하여 특정한 업종이 위치하기 좋은 지역을 찾을 수 있음

중첩의 유형화

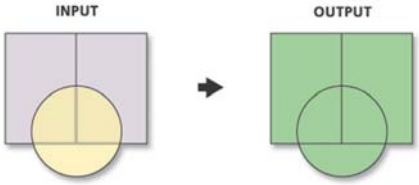
점과 면의 중첩 (points-in-polygon)	선과 면의 중첩 (line-in-polygon)	면과 면의 중첩 (polygon-and-polygon)
• 예: 서울시 자치구별 강력범죄 발생률을 나타내는 단계구분도와 CCTV 대수를 나타내는 점 지도 중첩 • QGIS에서는 공간적 조인(Spatial Join) 또는 위치로 선택(Select By Location)을 사용할 수 있음	• 예: 홍수에 취약한 지역을 찾기 위해 지역 내에 선으로 표현된 강(江)이 포함된 곳을 중첩 분석을 통해 찾을 수 있음 • QGIS에서는 위치로 선택(Select By Location) 사용	• 예: 토지이용계획을 세우기 위해 농업지역, 유적지, 특정한 종의 서식지 등 다양한 정보를 중첩 • 중첩이 사용되는 가장 전통적이고 일반적인 경우 • 교차, 결합 등 다양한 연산 기능을 사용할 수 있음

대표적인 중첩 기능

- 교차(intersect)
  - 두 개의 레이어 A와 B를 중첩하는 경우 두 레이어에 공통적으로 포함된 부분만이 별도의 레이어로 저장됨($A \cap B$)
  - 관련 실습 내용 참고!

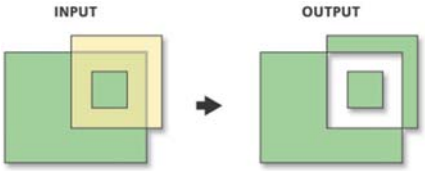


- 결합(union)
  - 두 개의 레이어 A와 B를 결합하는 경우 두 레이어 간에 겹치거나 부분적으로 교차하는 모든 형상들이 포함된 레이어가 새롭게 저장됨($A \cup B$)

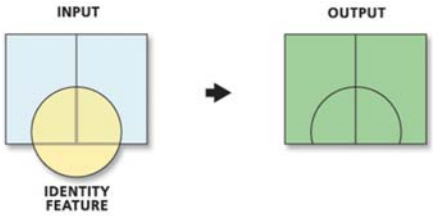


대표적인 중첩 기능

- 여집합(symmetric difference)
  - 두 개의 레이어 A와 B를 중첩하였을 때 두 레이어에 공통적으로 포함되어 있는 부분을 제외한 나머지 영역이 별도의 레이어로 저장됨 ($(A \cap B)^C$)

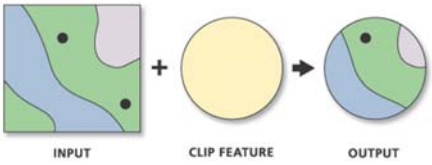


- 동일성(Identity)
  - 레이어 A와 레이어 B를 중첩하였을 때 레이어 A의 모든 객체들은 그대로 유지되지만, 레이어 B의 객체들은 레이어 A와 교차하는 부분만 유지됨

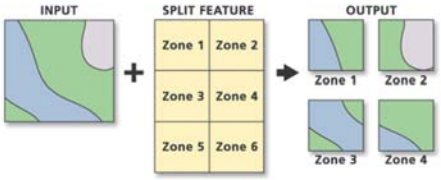


대표적인 중첩 기능

- 클립(clip)
  - 두 번째 레이어 B를 이용해 첫 번째 레이어 A를 잘라내는 기능으로 전체 데이터에서 원하는 부분(예를 들어 연구 지역)만을 추출하고자 할 때 사용



- 분할(split)
  - 두 번째 레이어 B를 이용해 첫 번째 레이어 A를 작은 여러 개의 조각으로 분할하는 기능



The 7th KOSTAT-UNFPA
Summer Seminar on Population

06

공간데이터의 시각화

6. 공간데이터의 시각화

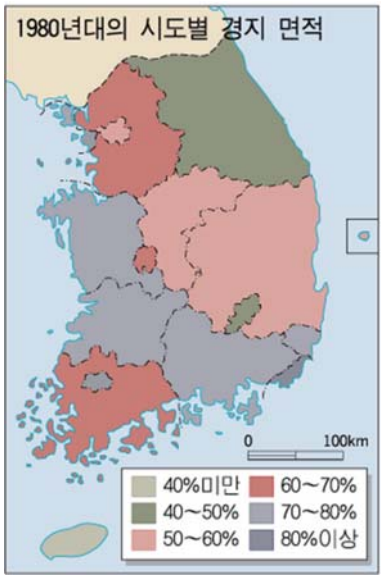
1) 단계구분도
2) 점 지도
3) 차트 지도
4) 카토그램

공간데이터의 시각화

- 공간데이터의 시각화(지도 제작)에 앞서 고민해야 하는 요소
  - 지도를 통해 전달하고자 하는 핵심적인 내용은 무엇인가?
  - 해당 내용을 전달하는데 있어 지도가 반드시 필요한가(일반적인 그래프 등으로 충분하지는 않은가)?
  - 지도의 주 사용자층은 누구인가?
    - 전문가를 위한 지도에는 보다 많은 정보가 함축적으로 포함될 수 있음
  - 양적 데이터, 질적 데이터 등 데이터의 유형은 무엇인가?
    - 유형에 따라 효과적인 지도 형태가 달라짐

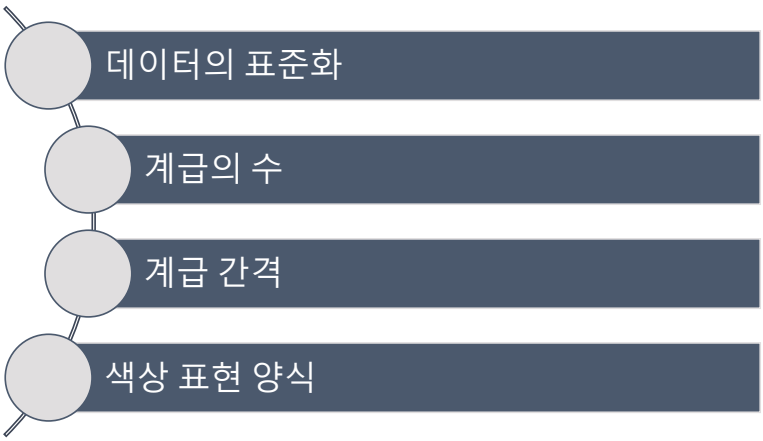
단계구분도

- 19세기 이후 가장 폭넓게 사용되는 주제도 유형의 하나
- 특정한 통계 수치 등을 어떤 기준에 따라 몇 단계로 구분하고, 색이나 농도 등을 달리하여 나타내는 지도
  - 오른쪽의 예에서는 행정구역 내의 경지 면적을 6단계로 구분하여 나타냄
  - 자치구별 재정자립도 등 단위구역에 따라 차이를 나타내는 현상을 표현하는데 유용함



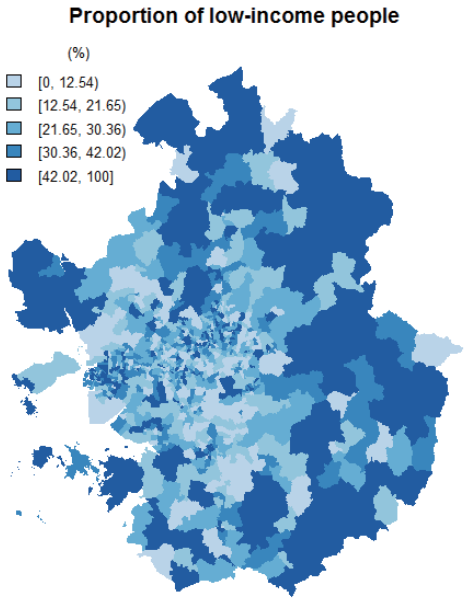
단계구분도 제작 방법

- 단계구분도 제작에 있어 고려해야 하는 중요 항목:



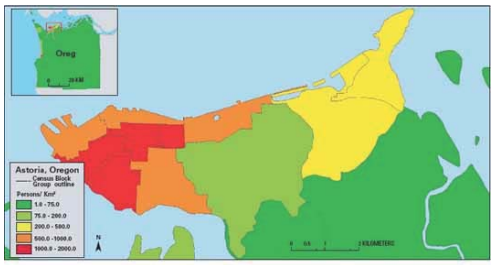
데이터의 표준화

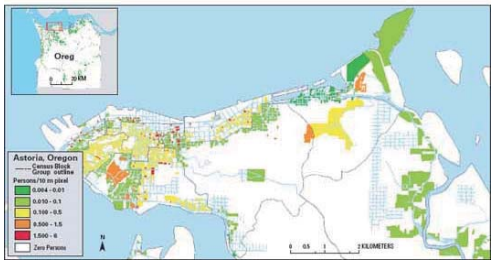
- 단계구분도는 지도에 사용된 공간 단위가 전반적으로 유사한 경우에 효과적임
  - 지도에서 공간 단위의 면적, 모양이 매우 다른 경우에는 정보 전달에 왜곡이 있을 수 있음
  - 왼쪽 지도는 수도권의 저소득 가구 분포를 나타내는 단계구분도로, 행정 경계의 크기가 작은 서울의 분포는 확인이 어려운 것을 알 수 있음



데이터의 표준화

- 대시메트릭(dasymetric) 매핑
  - 특정한 공간 단위(예를 들어 행정동) 위에 나타난 데이터를 위성영상, 토지피복도 등의 보조 데이터를 사용하여 보다 세밀한 단위로 재분류하여 표현하는 지도학적 기법

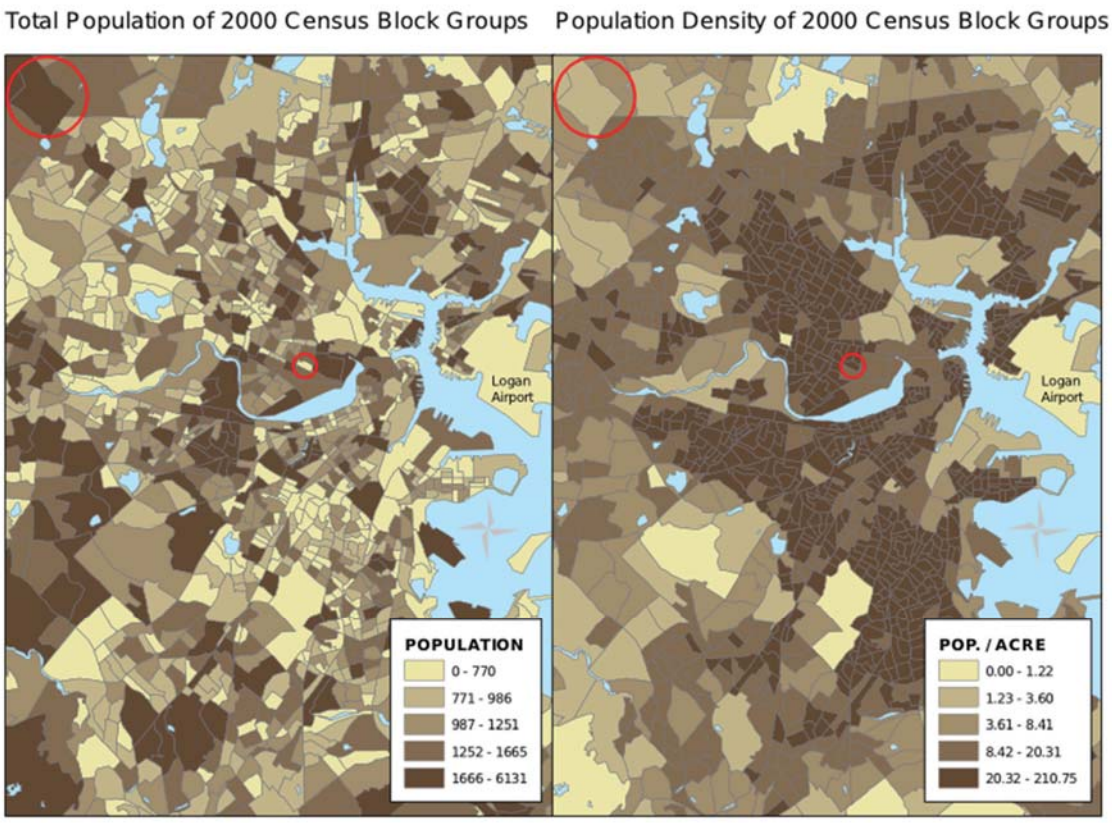




http://pubs.usgs.gov/tm/tm11c2/

데이터의 표준화

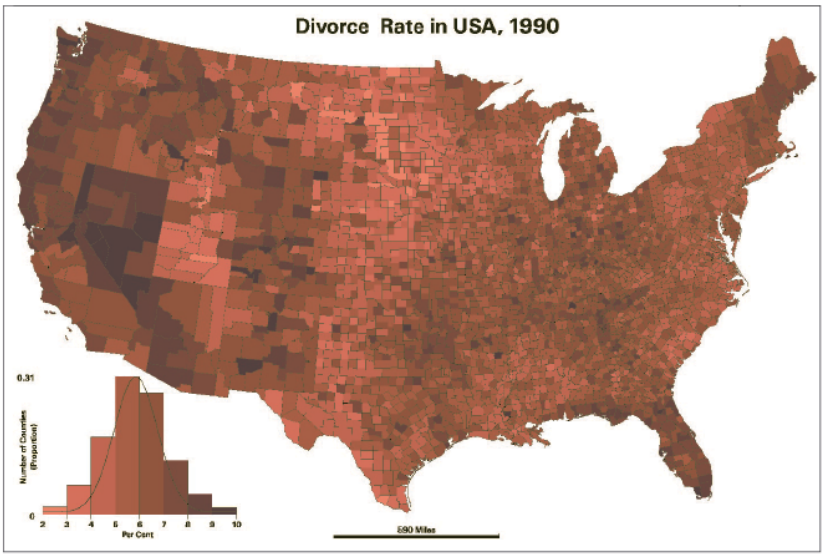
- 인구 데이터의 표준화
  - 지역별 인구 분포를 단계구분도로 나타낸다고 할 때, 일부 지역의 면적이 다른 지역에 비해 현저히 크다고 한다면 인구 수를 그대로 나타내는 것보다 면적 등으로 표준화하여 나타내는 것이 좋을 수 있음(인구 수 ➔ 인구 밀도)
  - 다른 예로, 미국의 주(州)별 사망률을 단계구분도로 나타내는 경우 플로리다 주의 사망률이 연도별로 증가하는 것을 확인할 수 있음
  - 이유가 무엇일까?



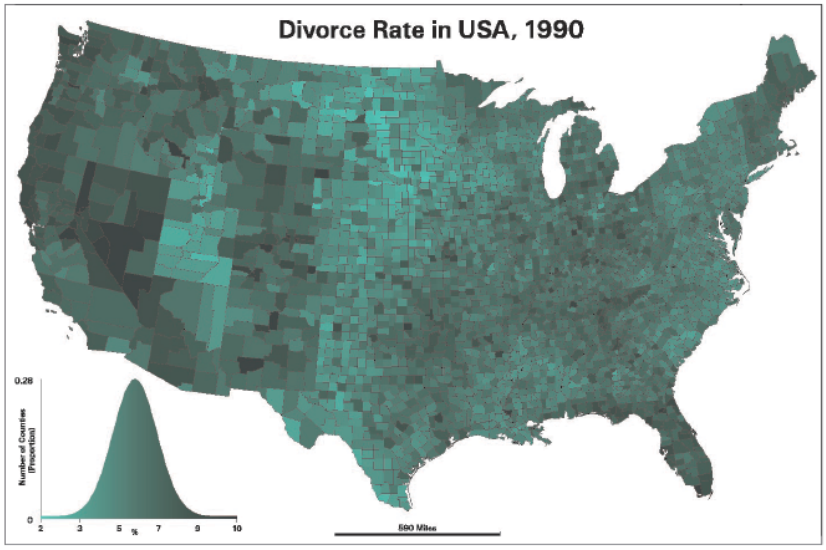
계급의 수

- 단계구분도 제작에서 가장 어려운 문제 중 하나가 계급의 수를 결정하는 일
  - 계급 수가 너무 적으면 표현하고자 하는 정보가 지나치게 일반화 또는 단순화되어 현상의 공간적 분포에 관한 정보가 정확하게 독자에게 전달되지 않을 수 있음
  - 너무 많은 수의 계급이 사용되면 지도가 너무 복잡해질뿐만 아니라 오히려 효과적인 정보의 전달을 방해할 수도 있음
- 일반적인 계급 수의 범위는 5개에서 15개 정도이며, 데이터의 수(집계 구역의 수)가 증가할수록, 그리고 데이터 값의 범위가 커질수록 계급의 수도 늘어나는 것이 일반적임

계급의 수



계급의 수



계급의 수

- Scott (1979)이 제안한 최적의 계급 수(k):

$$k = 3.5sn^{-\frac{1}{3}}$$

  – 여기서 s는 데이터의 표준편차, n은 데이터의 수를 의미함

- 계급을 너무 세분화하는 것은 의미가 없을 수 있음
  – 일반적으로 사람들은 15개 정도까지의 음영 단계만을 구분 가능

- 계급 수를 결정할 때에는 논리적으로 같은 그룹에 속해서는 안되는 데이터 값이 동일한 계급에 속하지 않도록 주의해야 함
  – 기온 자료를 지도화할 때 계급의 수를 너무 적게 설정하여 하나의 계급에 –10℃ 에서 10℃ 까지가 모두 속하는 일이 없도록 해야 함

계급 간격의 설정

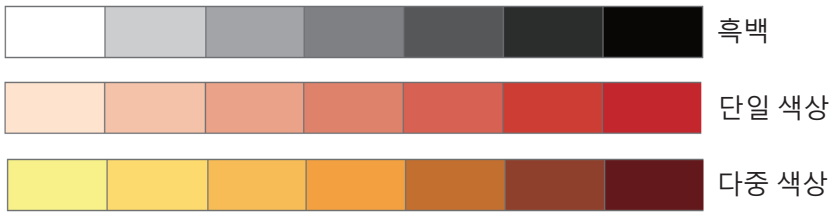
- 수동(manual)
  - 구분점 값을 수동으로 설정하는 방법
  - 임의의 구간을 강조하고 싶을 때 유용할 수 있음
- 등간격(equal interval)
  - 속성값의 최대값과 최소값 범위를 동일한 간격으로 나누어 구분점 값을 설정하는 방법
  - 인구 수나 소득 분포와 같이 지역 간 차이가 많이 나타나는 데이터를 분류하는 방법으로는 적절하지 않음
  - 계급 간에 포함되는 데이터의 수가 매우 다를 수 있기 때문
- 정의된 간격(defined interval)
  - 등간격 방식과 유사하나 구분점의 수가 아닌 구분 간격을 지정

계급 간격의 설정

- 등도수(quantile)
  - 각 구간에 전체 데이터의 일정 비율이 포함될 수 있도록 분류
- 기하학적인 간격(geometrical interval)
  - 승수에 기반을 둔 등비수열(geometric sequence)을 갖는 속성값 분류
- 표준편차(standard deviation)
  - 정규분포를 지닌 데이터를 표현하는데 적합한 방법
  - 평균값을 기준으로 좌우 대칭적으로, 표준편차 간격으로 구분점 설정
- 최적 분류(natural breaks)
  - Jenks (1967)가 고안한 수학적 최적화 분류 방법을 토대로 구분점 값을 설정하는 방법이며, 대부분의 GIS 소프트웨어에서 기본값으로 사용됨

색상 표현 양식

- 흑백을 사용할 것인가, 컬러를 사용할 것인가?
  - 지도를 종이에 인쇄하여 사용할 것인가, 아니면 컴퓨터 스크린에서 사용할 것인가?
- 순차적(sequential) 색상 표현
  - 낮은 값에서 높은 값으로 이어지는 양적 자료의 표현에 적합한 방식
  - 일반적으로 색상이 어두울수록 높은 값을, 또는 부정적인 현상을 나타내는 경우가 많음



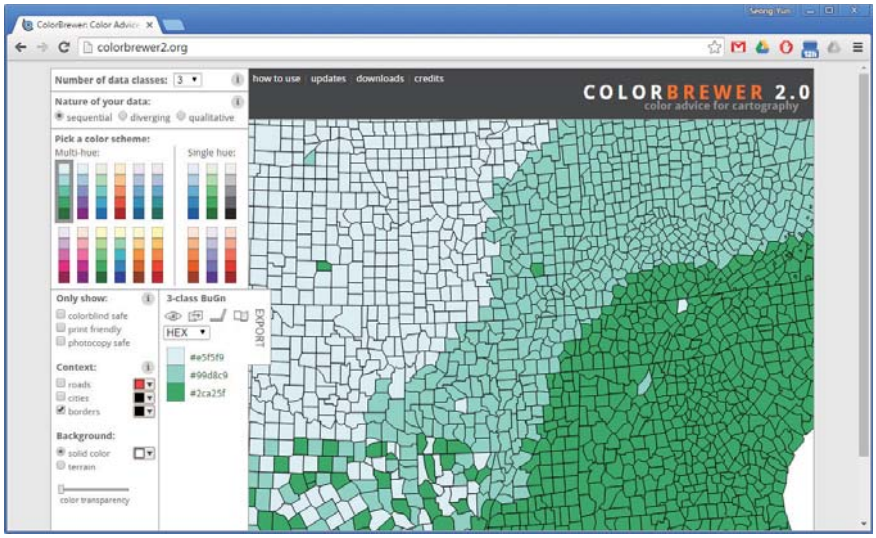
색상 표현 양식

- 갈래형(diverging) 색상 표현
  - 데이터의 중간 범위와 양쪽 끝부분의 값들을 동시에 강조
  - 데이터의 가운데에 위치하는 계급은 옅은 색으로 표현하고, 가장 낮거나 높은 데이터 값은 서로 대비되는 진한 색으로 표현함

- 분광형(spectral) 색상 표현
  - 질적 자료의 표현에 유용한 색상 표현

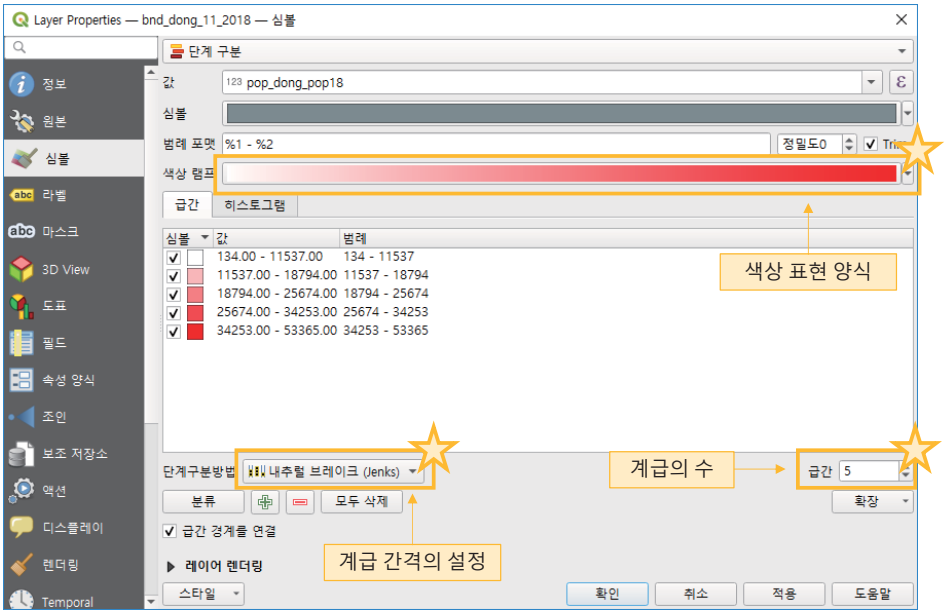
색상 표현 양식

- 유용한 사이트: colorbrewer2.org



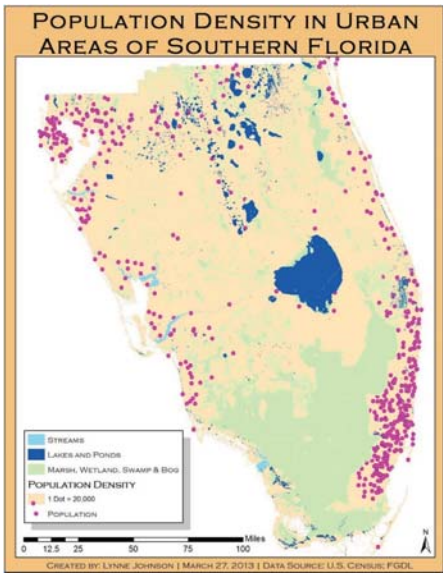
https://colorbrewer2.org/#type=sequential&scheme=BuGn&n=3

QGIS와 단계구분도



점 지도

- 점(point)을 이용한 주제도 중 가장 단순한 형태의 지도
- 같은 양을 대표하는, 동일한 크기의 점을 반복적으로 그려 속성값의 분포를 나타냄
  - 분포의 상대적 밀도를 시각적으로 쉽게 전달
  - 서로 다른 색이나 형태를 사용하여 여러 종류의 속성값을 동시에 나타낼 수 있음



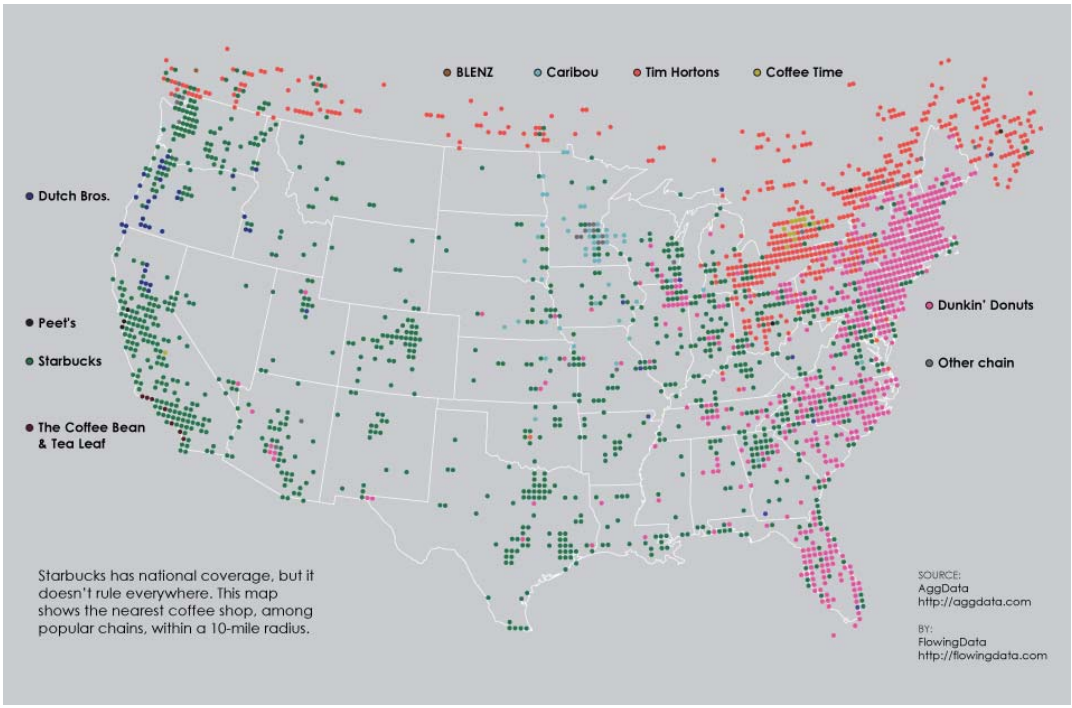
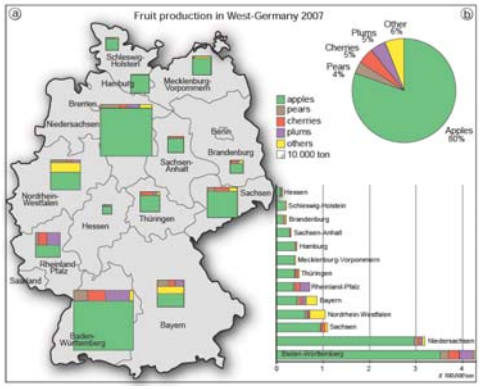
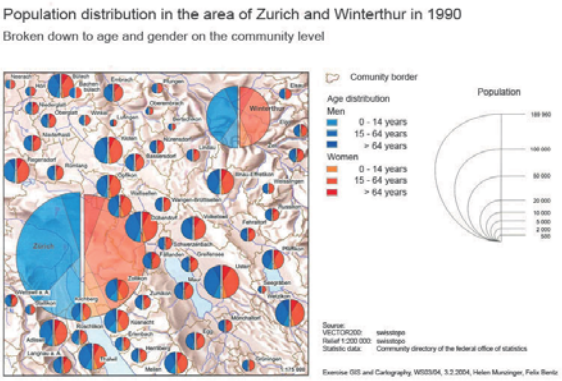


차트 지도

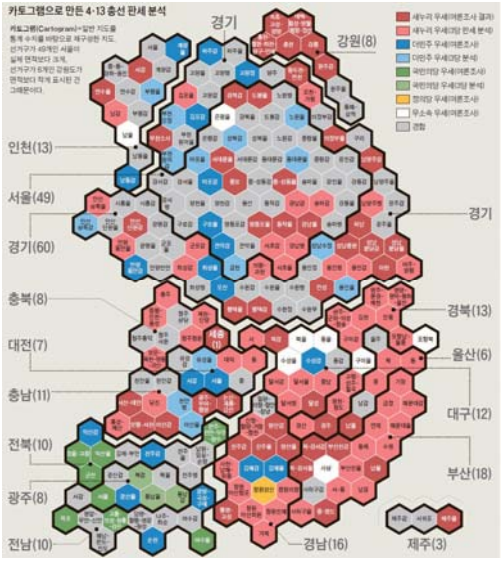
- 지도 위에 막대그래프나 원형그래프와 같은 다양한 형태의 차트를 올려서 데이터를 나타내는 지도
  - 좁은 지도 위에 너무 많은 시각적 요소가 있는 경우, 지도의 정보전달력이 떨어질 수 있기 때문에 주의가 필요함





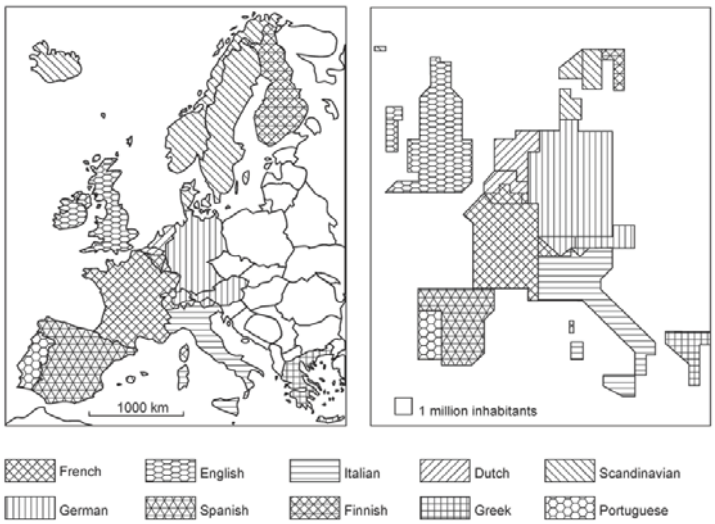
카토그램

- 단계구분도와 유사하나, 집계구역의 크기를 다른 속성값에 비례하여 변경한 지도
  - 오른쪽의 지도는 선거구의 면적을 모두 동일하게 재조정한 카토그램
- 전문적인 지도 제작 소프트웨어가 필요함
  - QGIS에는 플러그인이 있음
  - ArcGIS에서도 ArcScript를 사용(arcscripts.arcgis.com), 또는 ScapeToad와 같은 별도의 응용프로그램을 사용



카토그램

- 유럽의 지역별 모국어 분포를 나타낸 후, 지역별 인구를 바탕으로 면적을 재조정한 카토그램



The 7th KOSTAT-UNFPA
Summer Seminar on Population

07

공간적 자기상 관성의 개념

7. 공간적 자기상관성의 개념

1) 공간적 자기상관성
2) 전역적 모란 지수
3) 기어리 지수
4) Joins-count 검정

공간적 자기상관성의 개념

- Any spatial data set is likely to have characteristic distances (or lags) at which it is correlated with itself.
  - This property is known as self-correlation, or autocorrelation.
  - The ubiquity of spatial autocorrelation is the reason why spatial data are special.
- The degree to which data are similar or different over short or long ranges is fundamental to all branches of spatial analysis.
  - Autocorrelation is likely to be most pronounced at short distances.
  - "Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970).

공간적 자기상관성의 측정

- One reason for developing analytical approaches to spatial autocorrelation is to provide a more objective basis for deciding:
  1. Whether or not there really is a spatial pattern, and if so,
  2. How unusual that pattern is.
- In a sense, many of the point pattern measures we have already discussed can be considered as measures of autocorrelation.
  - Ripley's K function can describe the degree of autocorrelation for the occurrence of point events.
  - There are a number of methods developed specifically for the measurement of spatial autocorrelation.

전역적 모란 지수

- One of the most widely used measure is Moran's I (Moran, 1950), which is a translation of a non-spatial correlation measure to a spatial context.
  - It is often applied to areal units where numerical ratio or interval data are available, but it can also be used for marked point patterns.
  - The essential idea behind this approach is to assess how similar or different attribute values at geographic locations are relative to how spatially close or distant are the associated locations.

전역적 모란 지수

- Moran's I (Moran, 1950) is defined as below:

$$I = \left[\frac{n}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right] \times \left[\frac{\sum_{i=1}\sum_{j=1} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}\sum_{j=1} w_{ij}}\right]$$

where the subscripts i and j refer to different areal units or zones in the study, and y is the data value in each.

n refers to the total number of areal units, and w_{ij} indicates the spatial proximity between the units i and j.

7. 공간적 자기상관성

전역적 모란 지수

- Moran's I (Moran, 1950) is defined as below:

Covariance term

$$I = \left[\frac{n}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right] \times \boxed{\left[\frac{\sum_{i=1}\sum_{j=1} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}\sum_{j=1} w_{ij}}\right]}$$

where the subscripts i and j refer to different areal units or zones in the study, and y is the data value in each.

n refers to the total number of areal units, and w_{ij} indicates the spatial proximity between the units i and j.

전역적 모란 지수

- Moran's I (Moran, 1950) is defined as below:

Proximity between the spatial units

$$I = \left[\frac{n}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right] \times \left[\frac{\sum_{i=1}\sum_{j=1} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}\sum_{j=1} w_{ij}}\right]$$

where the subscripts i and j refer to different areal units or zones in the study, and y is the data value in each.

n refers to the total number of areal units, and w_{ij} indicates the spatial proximity between the units i and j.

전역적 모란 지수

- The spatial weights matrix W contains information about the spatial relationship between all pairs of locations.

$$W = \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{bmatrix}$$

The element in row i, column j of the weights matrix, denoted w_{ij}, represents the relationship between location i and location j.

- For point patterns, each w_{ij} value can be the Euclidean distance between the two points, i and j.

전역적 모란 지수

- Moran's I (Moran, 1950) is defined as below:

Normalise by the overall data set variance

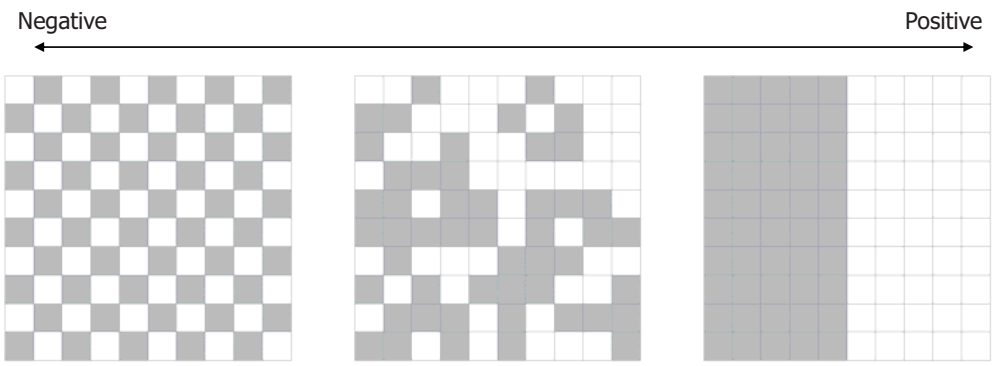
$$I = \boxed{\left[\frac{n}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right]} \times \left[\frac{\sum_{i=1}\sum_{j=1} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}\sum_{j=1} w_{ij}}\right]$$

where the subscripts i and j refer to different areal units or zones in the study, and y is the data value in each.

n refers to the total number of areal units, and w_{ij} indicates the spatial proximity between the units i and j.

전역적 모란 지수

- Theoretically, Moran's I ranges between -1 and 1.
  - A positive value of Moran's I indicates a positive autocorrelation, and a negative value a negative or inverse correlation.



전역적 모란 지수

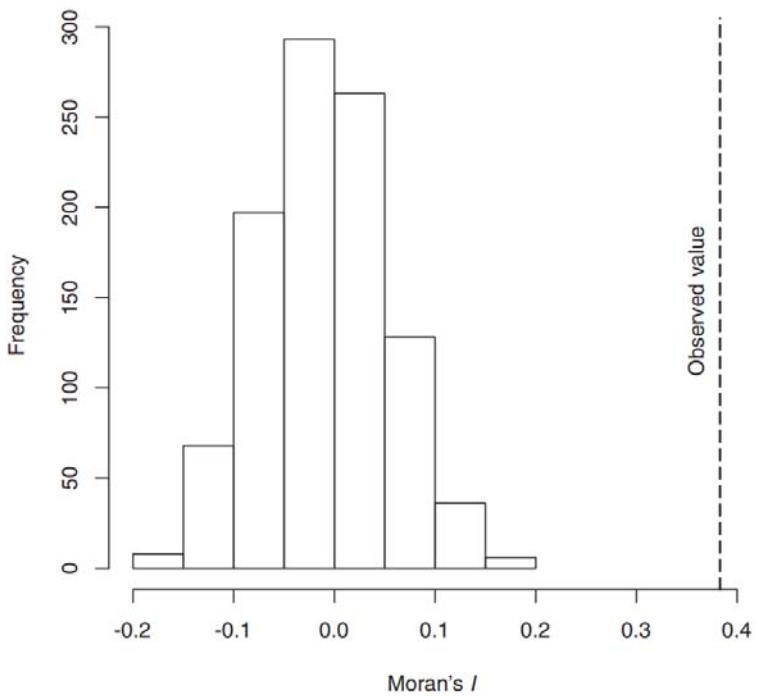
- In practice, however, it is very unusual to see such extreme values, as it is difficult for a map to be perfectly autocorrelated.
  - Generally speaking, an index score of 0.3 or more, or of 0.3 or less, is an indication of relatively strong autocorrelation.

통계적 유의성

- Some attention must be paid to the statistical significance of any results from Moran's I.
  - Is the observed spatial autocorrelation is significantly different from random? Could the apparent pattern have occurred by chance?
- Statistical significance of Moran's I should be tested before we start developing elaborate theories to explain the pattern in a map.
  - The pattern (we think we see) in the map!
  - That apparent pattern could be no more than a chance occurrence.

통계적 유의성

- A Monte Carlo approach is often used to associate p-values with observed values of Moran's I.
  - The location attribute values can be permuted any required number of times (999 is typical), that is, the attribute values observed in the map are randomly assigned to the map locations.
  - Moran's I is recalculated each time for the randomly-generate map.



기어리 지수

- Moran's I is not the only spatial autocorrelation measure.
- An alternative is Geary's C, which is defined as below:

$$C = \left[\frac{n-1}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right] \times \left[\frac{\sum_{i=1}\sum_{j=1} w_{ij}\,(y_i - y_j)^2}{2\sum_{i=1}\sum_{j=1} w_{ij}}\right]$$

  - As with Moran's I, the first term is a variance normalisation factor to account for the numerical values of y.
  - The second term has a numerator based on the square of the difference in y between the two areas under consideration.
  - The second term increases when there are large differences between adjacent locations.

기어리 지수

- The interpretation of Geary's C is different from I; it is actually quite opposite.
  - A value of 1 indicates no autocorrelation.
  - Values less than 1 (but greater than or equal to 0) indicate positive autocorrelation.
  - Values more than 1 indicate negative autocorrelation.
- The reason for this is clear if you consider that the $(y_i - y_j)^2$ term in the calculation is always positive but gives smaller values when similar values are neighbors.

Joins-count 검정

- The joins count test is one possible approach in situations:
  - Where interval or ratio data are not available, or
  - Where some threshold value of the attribute is of particular interest, so that areas above and below the threshold can be treated as binary outcomes
- Based on counting the number of occurrences of neighbouring pairs of polygons in the various different possible categories:
  - In the binary case where we can characterise the two available states as black and white, we arrive at counts of the number of black–black, white–white, and black–white neighbor joins.
  - The observed counts can be compared to the expected numbers to assess the type and strength of autocorrelation present.

Joins-count 검정

- Results from the joins count test
  - Positively autocorrelated maps will have more black–black and white–white joins than expected.
  - Negatively autocorrelated maps will have fewer such joins and more black–white joins than expected.
- Limitations
  - Applicable only to categorical data
  - Not easy to handle when there are more than a small number of categories because of the large number of possible types of join that quickly arise
    - With 6 categories, 15 join types are possible, and with 12, there are 66!

The 7th KOSTAT-UNFPA
Summer Seminar on Population

08

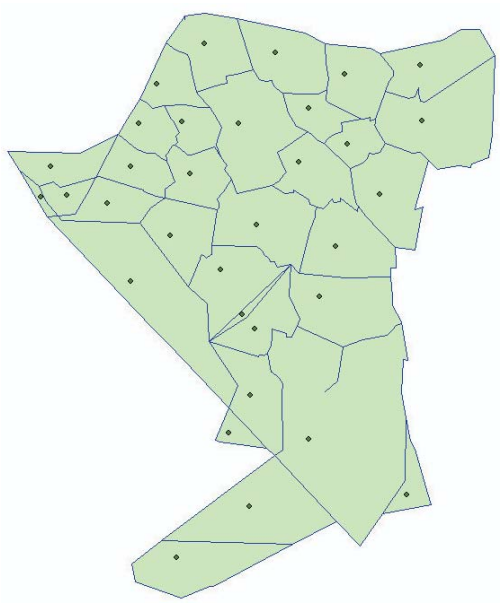
공간 가중 행렬

8. 공간 가중 행렬

1) 공간 단위 간의 거리
2) 공간 가중 행렬의 구축
3) 공간 가중 행렬의 중요성
4) 이상적인 공간 가중 행렬

공간 단위 간의 거리

- The autocorrelation concept is applicable to all the types of spatial objects, not limited to points.
- For areal data:
  - The same distance metrics we discussed in the previous lecture can be applied, if the polygon areas can be represented as points at the polygon centroid.



공간 단위 간의 거리

- Areas are some of the more complex object types commonly analysed.
  - Natural areas: Entities modeled using boundaries defined by natural phenomena such as the shoreline of a lake, the edge of a forest stand, or the outcrop of a particular rock type
  - Imposed areas: Areas imposed by human beings, such as countries, provinces, states, counties, or census tracts
    - The boundaries of imposed areas are defined independently of any phenomenon, and attribute values are enumerated by surveys or censuses.
    - Common in GIS work that involves data about human beings.

공간 가중 행렬

- Adjacency can be thought of as the nominal, or binary, equivalent of distance.
  - Two spatial entities are either adjacent or they are not.
  - This is an important idea in the measurement of autocorrelation effects when a region is divided into areal units.
- In a spatial weights matrix, W, the w_{ij} values will be 1 if the two locations i and j are adjacent, and 0 if they are not.

$$W = \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{bmatrix}$$

  - This is, however, not as straightforward as it seems …

공간 가중 행렬

- How adjacency should be determined is not necessarily clear!
  - The most obvious case is a set of polygons, in which we consider any two polygons that share an edge to be adjacent.
  - An equally simple formulation is to decide that any two entities within some fixed distance of one another (say 100 m) are adjacent to one another.
  - Alternatively, we might decide that the six nearest entities to any particular entity are adjacent to it.
  - We might even decide that only the single nearest neighbour is adjacent.

인접성 기반 정의

- Even in the most obvious case:
  - We may choose to require areas to share an edge in order to consider them adjacency (the Rook's case), or
  - We may consider it sufficient that they only meet at a corner vertex (the Queen's case).

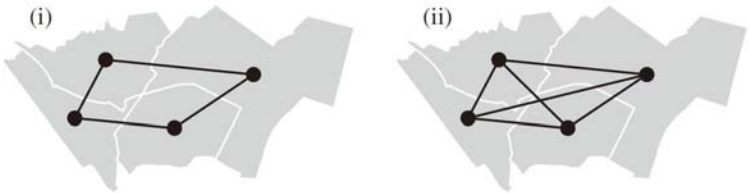
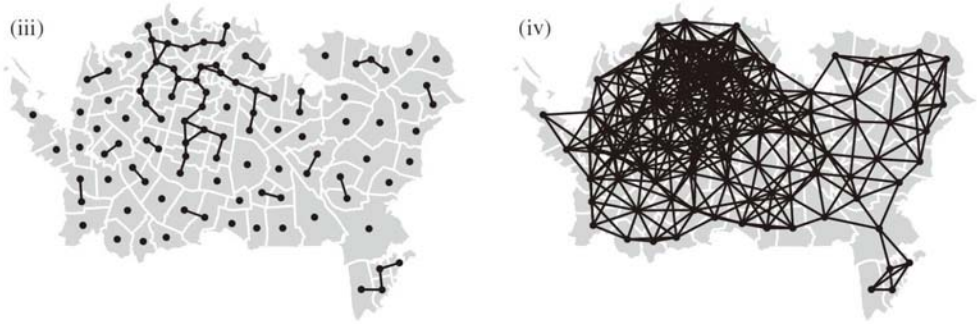


Figure 7.5 (i) Rook's and (ii) Queen's case adjacencies among polygons.

Source: O'Sullivan & Unwin (2010), p. 201

거리 기반 정의

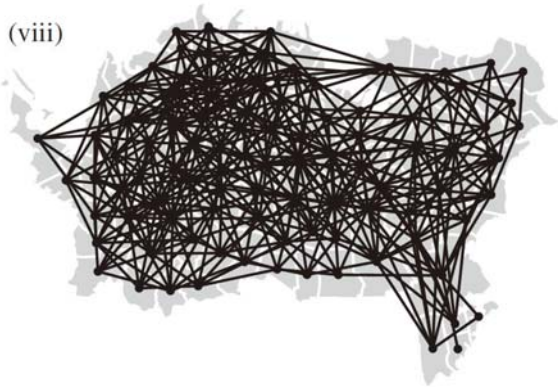
- If we use some measure of distance between polygons:
  - Based on some distance threshold d, we consider two cases adjacent if $d_{ij} < d$ and not otherwise.
  - Alternatively, we may wish to include only the nearest neighbors.



Source: O'Sullivan & Unwin (2010), p. 202

그 밖의 다른 방법

- Zones adjacent at lag two are those that are neighbours "once removed" across an intervening zone.
  - In practice, it is not clear how meaningful analyses based on more remote lags are likely to be.



Source: O'Sullivan & Unwin (2010), p. 202

고려해야 하는 요소

- In the cases discussed so far, adjacency remains a binary quantity.
  - The w_{ij} values were either 1 (connected) or 0 (not connected).
- If some relationships are considered stronger than others, we can make the w_{ij} values range from 0 (for weak interaction) to 1 (for strong interaction):
  - By assuming an inverse-power relationship of their separation distance, or
  - By using the length of shared boundaries between adjacent locations.

고려해야 하는 요소

- Two important considerations in the construction of the weights matrix are:
  - How we deal with the relationship between each location and itself,
  - How we enforce symmetry.
- Because we are not interested in the relationship between each location and itself, elements on the main diagonal of the matrix (i.e., w_{11}, w_{22}) are usually set to zero.

고려해야 하는 요소

- Symmetry in the weights matrix is generally required so that, in all cases:

$$w_{ij} = w_{ji}.$$

- Some methods for constructing the matrix do not guarantee such symmetry.
  - For example, in the k nearest neighbor approach, area A may have areas B, C, and D as its three nearest neighbors, while the three nearest neighbors of B are C, D, and E and do not include A.
  - In this case, w_{AB} does not equal w_{BA}.

고려해야 하는 요소

- To resolve this situation, we can enforce symmetry by setting:

$$\mathbf{W}_{\text{final}} = \frac{1}{2}(\mathbf{W} + \mathbf{W}^{\mathrm{T}})$$

so that each pairwise two-way relationship is the average of the two one-way relationships.

공간 가중 행렬의 중요성

- The map in the next slide shows reported cases of tuberculosis per 100,000 population for Auckland City, New Zealand, in 2001–2006.
  - These are not annual rates, but rates accumulated over the whole six-year period relative to the 2006 census population.
- Examination of the map suggests that there is a tendency for census areas in the southwest of the city (toward New Windsor) to have experienced higher rates of incidence of tuberculosis.
  - These areas form an arc from near Waterview to Onehunga.
- There is also a more isolated group of areas around Tamaki in the east, which also have higher incidence rates.

공간 가중 행렬의 중요성

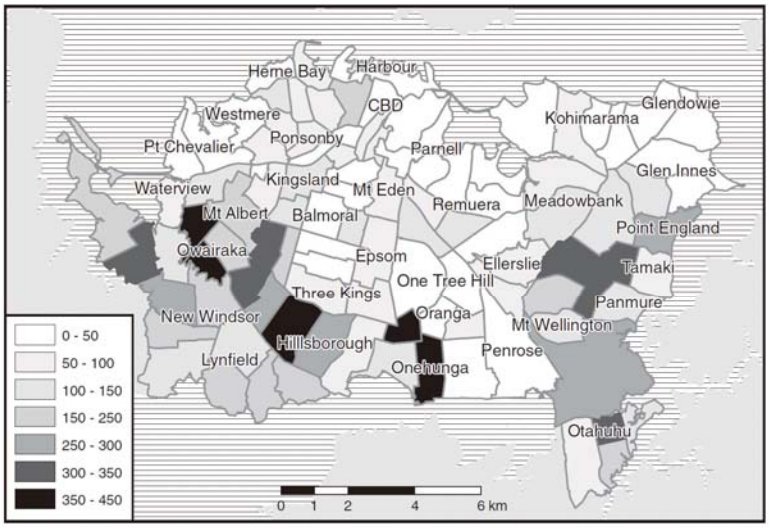


Figure 7.7 Reported cases of tuberculosis per 100,000 population, Auckland City, 2001–2006. The polygons are New Zealand Statistics census area units.

Source: O'Sullivan & Unwin (2010), p. 207

공간 가중 행렬의 중요성

- The Moran's I results for a number of different spatial weighting schemes are:

Spatial weighting scheme	Moran's I
Rook's adjacency	0.3830
Queen's adjacency	0.3941
d < 2,500 m	0.3510
k = 3 nearest neighbours	0.3780
k = 6 nearest neighbours	0.4014

이상적인 공간 가중 행렬

- The important points to appreciate are that:
  - A wide variety of spatial weights matrices are possible in any given situation, and
  - The choice of spatial weights for use in autocorrelation measurement is a key step in the analysis.
- Ideally, the spatial structure represented in $\mathbf{W}$ should correspond to some aspect of the problem that is meaningful in terms of the processes under consideration.

이상적인 공간 가중 행렬

- This is not always easy to arrange …
  - In the study of social processes in particular, census units or other administrative units are often used in the absence of any other convenient approach.
  - Developing a spatial weights matrix that relates to the posited underlying processes will also be difficult where those processes are not well understood.
- In such cases, it is advisable to work with simple adjacency-based approaches at least in the exploratory phase of the analysis.

The 7th KOSTAT-UNFPA
Summer Seminar on Population

09

국지적 측도

9. 국지적 측도

1) 국지적 통계와 GIS
2) 모란 산점도
3) 국지적 모란 지수
4) 전역적/국지적 G 통계량

국지적 통계의 개념

- Any descriptive statistic associated with a spatial data set whose value varies from place to place
  - In the broadest sense, any spatial data set is a collection of local statistics, in that the recorded attribute values are different at each location.
- Different from global statistics in that it is derived by considering a subset of the spatial data local at each location
  - The concept of a local statistic is widely deployed in spatial analysis, although it goes by different names in different contexts.

국지적 통계의 개념

Mean value of an attribute based on attribute values in the data set near the location of interest

- Localised Mean
  (Probably best known as ...)
- Focal Operation
- Smoothing Filter
- Spatial Interpolation

전역적 vs. 국지적 통계 비교

- An equivalent local measure can be calculated for most global measures.

Global statistics	Local statistics

• A single value which applies to the entire data set
  • The same pattern or process occurs over the entire geographic area
  • An average for the entire area
	• A value calculated for each observation unit
	  • Different patterns or processes may occur in different parts of the region
	  • A unique number for each location

국지적 통계와 GIS

- Unwin (1996) and Fotheringham (1997) explicitly highlighted the importance of local statistics.
  - Unwin, D. 1996. GIS, spatial analysis and spatial statistics. Progress in Human Geography 20, 540–551.
  - Fotheringham, A. S. 1997. Trends in quantitative methods I: stressing the local. Progress in Human Geography 21, 88–96.
- Why has the idea only taken off recently?
  1. Mapping capability provided by GIS tools
  2. Increased computing power
  3. Recognition of the importance of geographic variation in phenomena

국지적 통계와 GIS

1. Mapping capability provided by GIS tools
   - Many local statistics are a natural by-product of the calculation of summary global statistics.
   - The advent of readily available mapping tools has led to the exploration of the potential of local statistics as an analytical output in their own right.
2. Increased computing power
   - The statistical evaluation of local statistics is more challenging than the statistical assessment of related global measures.
   - Monte Carlo simulation approaches are often used to test statistical significance, but it requires substantial computing power.

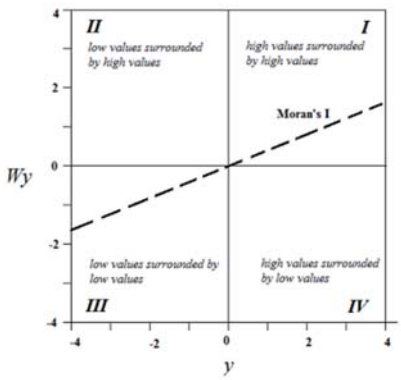
국지적 통계와 GIS

3. Recognition of the importance of geographic variation in phenomena (due to the widespread adoption and use of GIS tools and the accompanying increase in data availability)
   - As more data have become available, this has allowed studies both to expand their spatial range and to focus in at higher spatial resolution.
   - Both developments have prompted the realisation that the idea of a single global process or model being a realistic explanation is not always very plausible.

모란 산점도

- A scatter plot between x_i and the spatial lag of x_i formed by averaging all the values of x_i for the adjacent areas
  - Useful for identifying which type of spatial autocorrelation exists

35	24	31	58	21
36	53	27	32	28
28	40	40	27	39
56	57	34	37	45
58	41	20	56	22



https://www.researchgate.net/profile/Giorgian_Ionut_Gutoiu/publication/294870442/figure/fig1/AS:330092074029056@1455711491301/Figure-1-Anselin%27s-Moran-scatter-plot-interpretation-guide.ppm

국지적 모란 지수

- The local version of Moran's I is one of the most commonly used methods, and it is often called Anselin's LISA or just LISA.
- The local Moran statistic for areal unit i is:

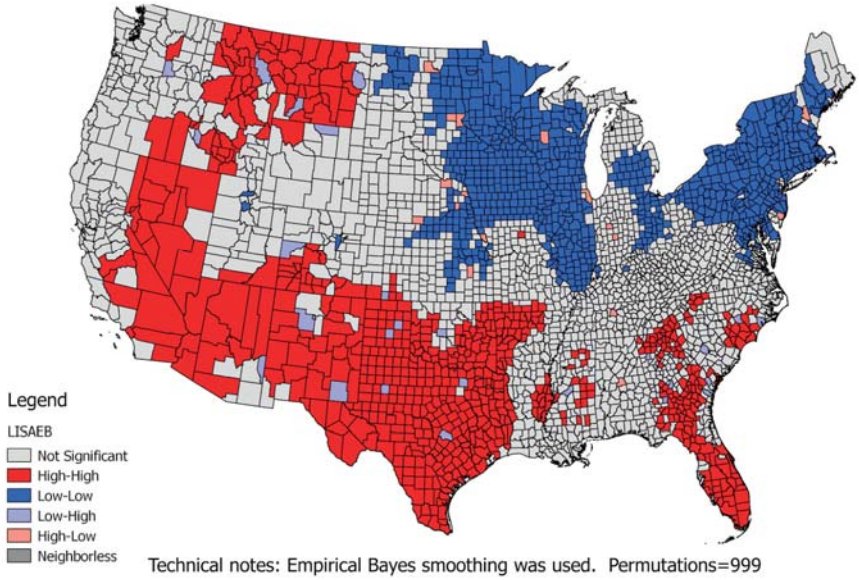
$$I_i = z_i \sum_j w_{ij} z_j$$

where z_i is the original variable x_i in standardised form (i.e., $z_i = \frac{x_i - \bar{x}}{SD_x}$), or it can be in deviation form (i.e., $z_i = x_i - \bar{x}$), and w_{ij} is the spatial weight.

국지적 모란 지수

- The statistic is calculated for each areal unit in the data
  - For each polygon, the index is calculated based on adjacent polygons with which (for example) it shares a border.
- The results can be displayed as a map:
  - Since a measure is available for each polygon, these can be mapped to indicate how spatial autocorrelation varies over the study region.
  - Since each index has an associated test statistic, we can also map which of the polygons has a statistically significant relationship with its neighbours, and show type of relationship.

Local Moran's I of Percent Uninsured under Age 65,by County



http://1.bp.blogspot.com/-mkBA_5098mc/U03gXjxh2XI/AAAAAAAAAg8/vJdn2vRQq68/s1600/Univariate+LISA.png

인접성의 정의

- Choices made in constructing localities prior to determining local statistics are a critical aspect of the analysis.
  - Local statistics may point to patterns of a particular kind when localities are constructed based on adjacency.
  - They may reveal completely different patterns when localities are constructed based on a distance criterion.
- How do we define localities then?
  - Where possible, examine a number of different weights matrix constructions
  - Choose the method that makes the most sense in substantive terms

인접성의 정의

- In general:
  - Simple spatial adjacency based on contiguity among a set of polygons is somehow the natural approach to constructing localities.
- If you interested in some phenomenon whose patterns are likely to be related to transport accessibility:
  - It is probably more relevant to connect locations via the transport network (e.g., adjacency on estimated distances over road).
- Such options have become much more readily explored using the capacity of GIS to relate spatial data in a wide range of ways.

전역적 G 통계량

- If you are interested in the characteristics of clustering (e.g., whether high values are clustered together, or low values …) …
- The global G statistic of overall spatial association is given as:

$$G = \frac{\sum_i \sum_j w_{ij} x_i x_j}{\sum_i \sum_j x_i x_j}$$

where x_i and x_j are attribute values for features i and j, and w_{ij} is the spatial weight between feature i and j.

  - The attribute values should be positive values.
  - If weights are binary (0 or 1), or always less than 1, the range for G will be between 0 and 1.

전역적 G 통계량

- The expected value of G is given as below—but why?

$$\mathrm{E}(G) = \frac{\sum_i \sum_j w_{ij}}{n(n-1)}$$

- Comparisons of G between different cities and regions might be pointless …
  - Temporal comparisons are probably okay.

국지적 G 통계량

- It is the proportion of all x values in the study area accounted for by the neighbours of location i.

$$G_i = \frac{\sum_j w_{ij} x_j}{\sum_j x_j}$$

  - G_i will be high where high values cluster.
  - G_i will be low where low values cluster.
- Interpreted relative to expected value if randomly distributed:

$$\mathrm{E}(G_i) = \frac{\sum_j w_{ij}}{n-1}$$

국지적 G 통계량

- Expected values and variances for the local G statistic are known.
  - Calculation of the statistic's variance is complex and is beyond the scope of this course (see Getis and Ord, 1992, p. 191 for details).
- A z-score can be determined for each location's G_i value and can be mapped as in the next slide.
  - Let's compare the map on the next slide to the one on p. 289.
    - The highest-incidence locations are not the ones with the highest associated G_i values.
    - The census area units neighboring high-incidence areas are highlighted (e.g., Mt. Wellington).

국지적 G 통계량

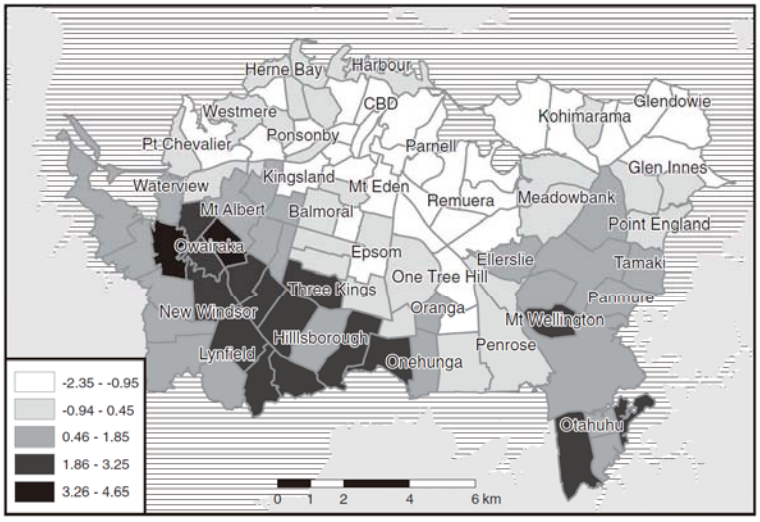


Figure 8.1 Map of the Auckland tuberculosis data z scores determined from the calculated G_i values.

Source: O'Sullivan & Unwin (2010), p. 202

통계적 유의성

- Care is required in making inferences from local statistics.
  - We interpret z-scores outside the range -1.96 to +1.96 as unusual cases and give particular attention to these parts of the map.
  - This is due to an assumption of normality, but is it valid in the case of local statistics?
- There are a number of possible problems:
  - The central limit theorem may not work well when the sample size is small ➔ If the localities under consideration are small, the statistic is being calculated based on small number of cases.
  - What if the localities under consideration are NOT small? ➔ The localities are no longer quite so local!

통계적 유의성

- There are other problems as well:
  - The data are evidently not well accounted for by a null model that assumes complete spatial randomness.
    - Because of the presence of spatial autocorrelation, it makes little sense to identify statistically unusual cases based on a null model that assumes complete spatial randomness!
  - Moreover, this is a situation where repeatedly applying a statistical test to the same data. ➔ What problem?
    - This is known as the multiple testing problem, which can be addressed by adjusting the probability threshold used to determine which results are considered statistically significant.

통계적 유의성

- To overcome these problems, a Monte Carlo simulation procedure is often used to produce pseudo-significance values.
  - Basically the same approach adopted in assessing many point pattern measures discussed in previous lectures
- In the context of local statistics, it is typically repeated using conditional permutation.
  - Each time the data are shuffled, the value at the location of interest is held constant (this is what makes the permutation conditional).
  - The calculations for the statistic in question are performed on the shuffled data, and the resulting value of the local statistic is determined ➔ Computationally intensive!

The 7th KOSTAT-UNFPA
Summer Seminar on Population

10

R과 RStudio 소개 및 설치

10. R과 RStudio 소개 및 설치

1) 국지적 통계와 GIS
2) 모란 산점도
3) 국지적 모란 지수
4) 전역적/국지적 G 통계량

통계프로그램 R

- 데이터 분석 및 시각화를 위한 프로그래밍 언어이자 오픈소스 소프트웨어로 무료로 다운로드 받아 사용할 수 있음
- 통계프로그램 S, S-Plus와 유사
  - 1990년대 뉴질랜드 오클랜드대학교 통계학과의 Ross Ihaka 교수와 Robert Gentleman 교수가 S 프로그램의 에뮬레이터 형태로 개발*
  - R과 S는 서로 다른 언어이나 S에서 작성된 많은 코드가 그대로 R에서도 실행이 가능함
- 현재 데이터 분석을 위해 세계에서 가장 많이 쓰이는 프로그램 중 하나
  - 대학, 연구소 등 학계 뿐만 아니라 산업계 전반에서 R의 사용이 빠르게 늘어나고 있음

통계프로그램 R

- 도메인 특화 언어(domain-specific language, 또는 줄여서 DSL)
  - 데이터 분석 및 시각화라는 특정 영역에서의 활용에 특화되어 있음
  - 반면 C, Java, Python과 같은 언어를 범용 언어(general-purpose programming language)라 함

Rank	Language	Type	Score
1	Python		100.0
2	Java		96.3
3	C		94.4
4	C++		87.5
5	R		81.5
6	JavaScript		79.4
7	C#		74.5
8	Matlab		70.6
9	Swift		69.1
10	Go		68.0

https://spectrum.ieee.org/computing/software/the-top-programming-languages-2019

R의 장점과 특징

- 기존 프로그램과의 유사성
  - R은 완전히 새로운 언어(또는 소프트웨어)는 아니며, R에서 사용되는 구문(syntax)과 데이터 저장구조는 S, LISP 등의 기존 프로그래밍 언어에서 차용되어 왔음
  - 따라서 이러한 프로그래밍 언어에 이미 익숙한 사용자의 경우 비교적 쉽게 R을 배울 수 있음
- 데이터 분석 및 시각화를 위한 다양한 내장 함수의 제공
  - 선형 및 비선형 회귀분석, 군집분석, 주성분분석, 시계열분석 등과 같이 일반적으로 많이 쓰이는 통계기법이 기본적으로 구현되어 있어 사용자가 쉽게 사용할 수 있음
  - 이와 같이 기본적으로 제공되는 기능 외에도 전세계의 수많은 사용자들이 직접 개발, 공유하는 패키지를 통해 손쉽게 확장이 가능

R의 장점과 특징

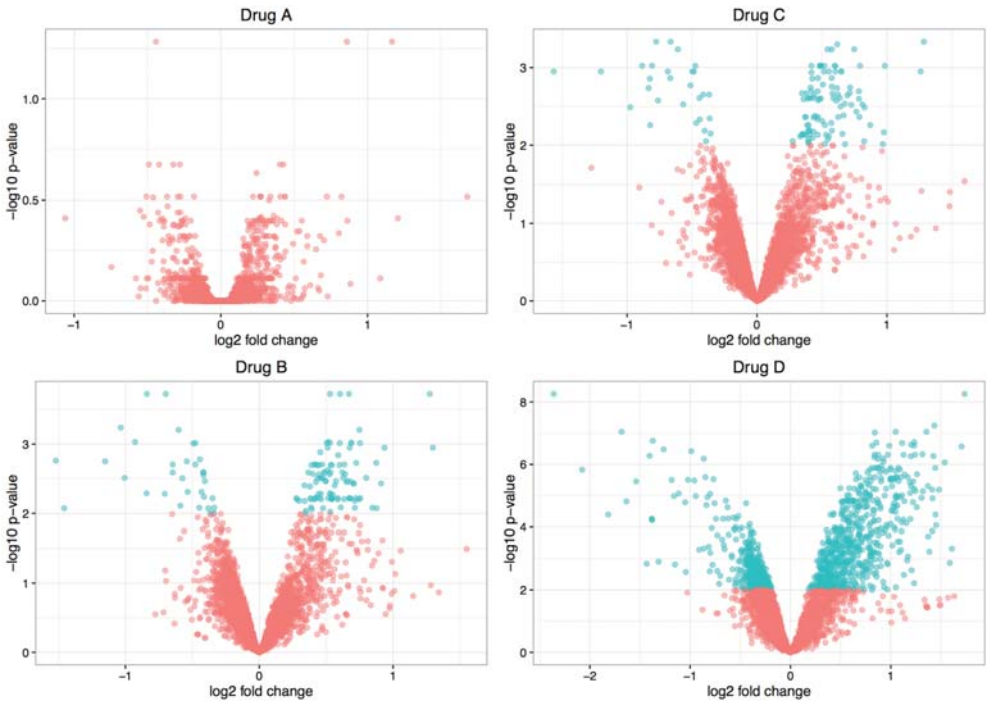
- 오픈소스(무료) 프로그램
  - 프로그램의 무료 배포와 소스 공개를 통해 넓은 사용자층을 확보하고 이들이 개발에 참여할 수 있는 기회를 제공
  - 각각의 사용자가 개발한 추가 기능(패키지)는 CRAN(Comprehensive R Archive Network, http://cran.r-project.org/)을 통해 다른 사용자와 공유가 가능
  - 2020년 3월 14일 기준, CRAN을 통해 15,367개의 패키지가 공유되고 있으며(https://cran.r-project.org/web/packages/), 이는 R이 지금과 같이 다양한 기능을 갖고 폭넓은 분야에서 활용될 수 있는 토대가 됨

R의 장점과 특징

- 효율적인 대용량 데이터 분석 처리
  - 분석에서 컴퓨터 자원을 많이 소요하는 부분의 경우 C++, 포트란과 같은, 보다 효율적으로 컴퓨터 자원을 제어할 수 있는 프로그램에서 계산하여 다시 R로 가져올 수 있음
  - 빅데이터의 저장, 분석과 시각화를 위한 다양한 패키지 제공
  - 예를 들어 100,000 x 100,000 크기의 데이터를 저장, 처리하기 위해 데이터를 파일 형식으로 보관하는 패키지를 사용할 수 있음

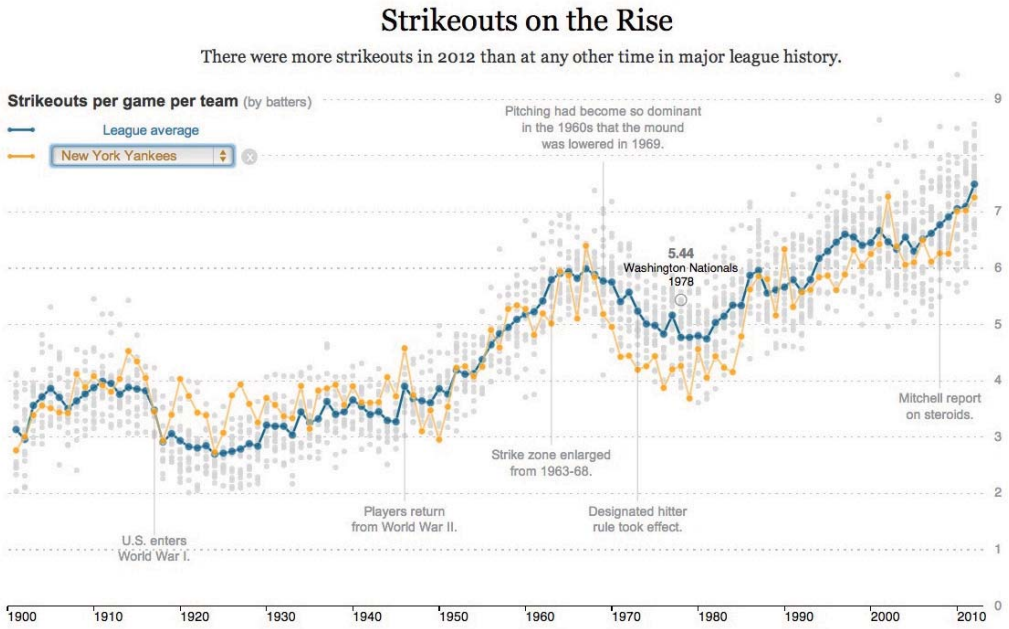
R의 장점과 특징

- 뛰어난 시각화 기능
  - 수학 기호와 수식 등을 포함된 전문적인(publication-ready) 그래프 생성이 가능
  - 시각화에 대한 전문적인 지식이 없는 사용자도 기본 설정만을 사용해서 효과적인 그래프를 만들 수 있도록 함
  - 동시에 그래프의 각 구성요소를 사용자가 세밀하게 조율할 수 있도록 허용함으로써 전문가가 필요에 맞게 그래프를 수정할 수도 있음



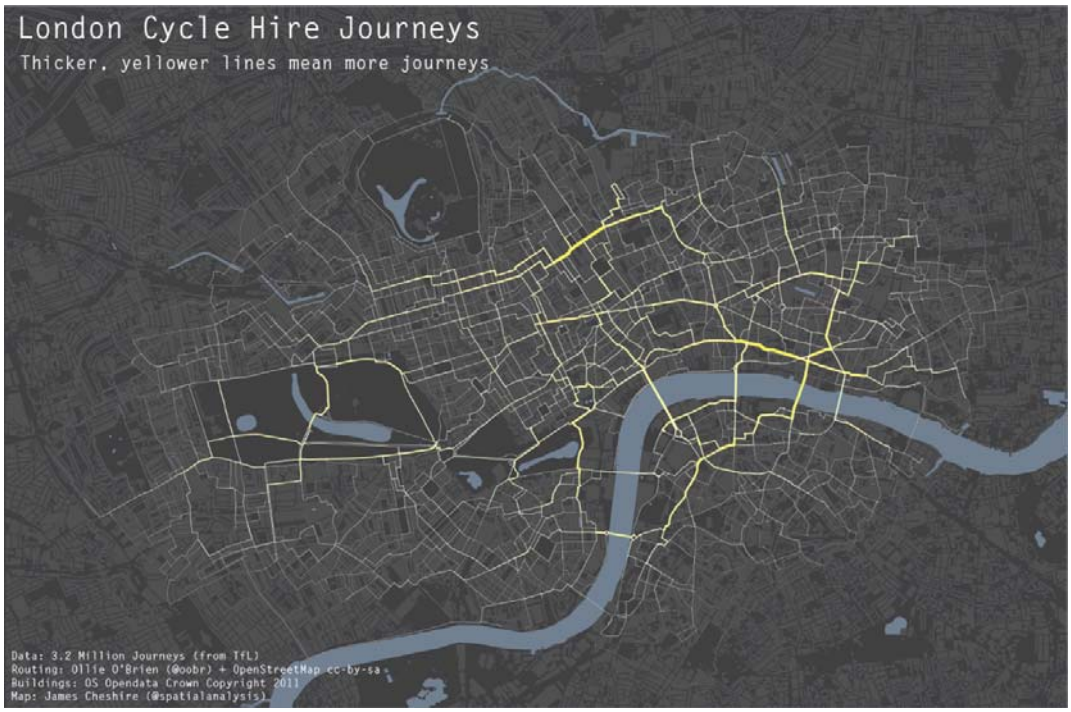
http://rforbiochemists.blogspot.kr/2016/03/gene-expression-analysis-and_7.html

10. R과 RStudio



http://www.datacommunitydc.org/blog/2013/05/data-visualization-rcharts

10. R과 RStudio



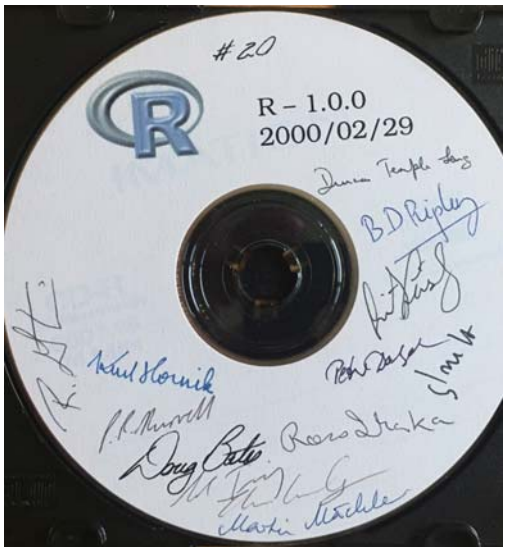
http://spatialanalysis.co.uk/wp-content/uploads/2012/02/bike_ggplot.png

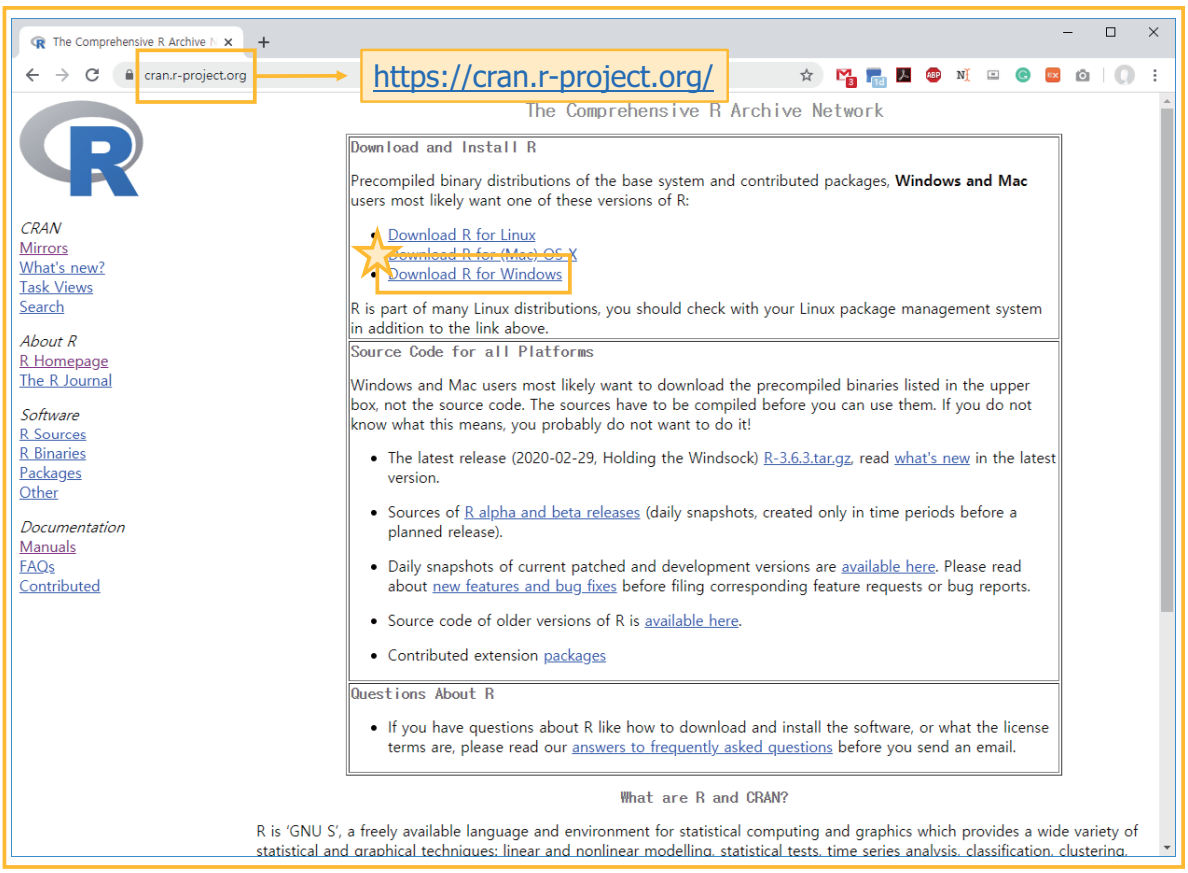
R의 단점과 한계

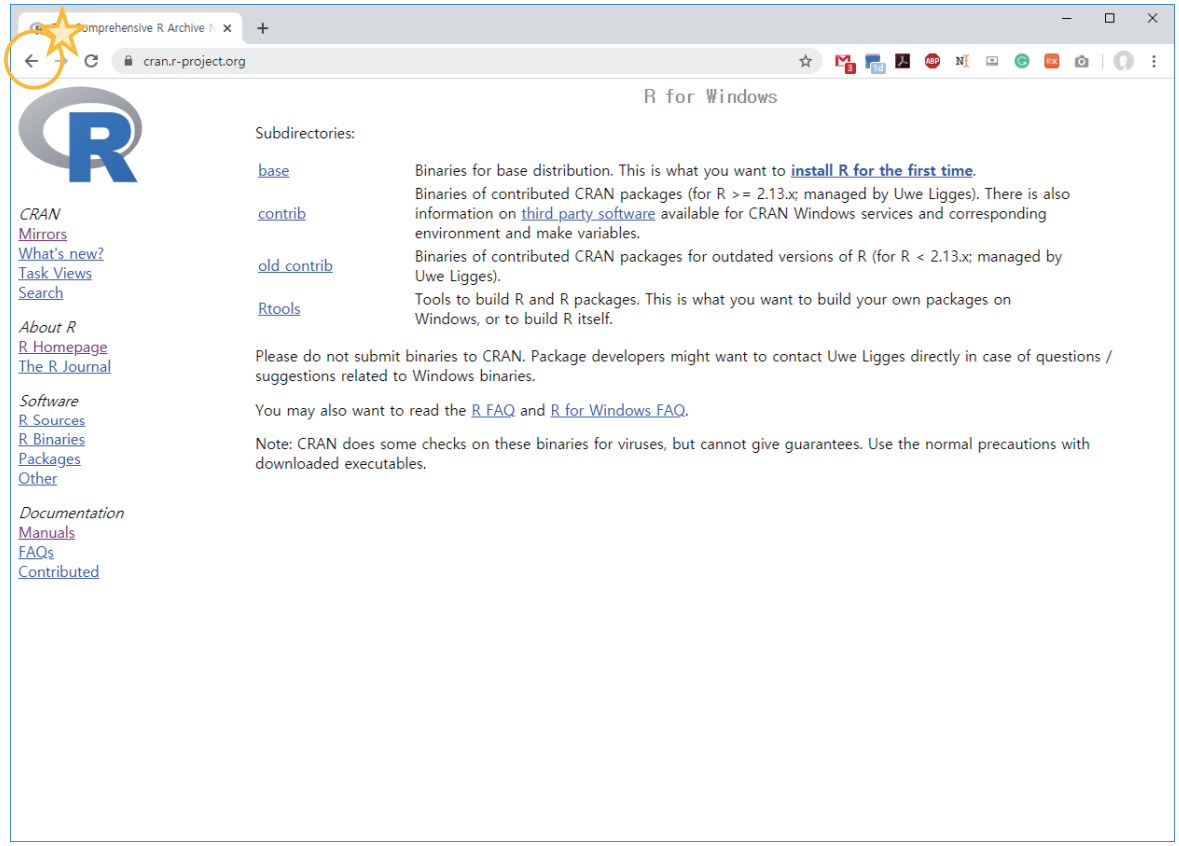
- 사용성 측면에서 상대적으로 높은 진입장벽
  - SPSS, Minitab과 같은 프로그램과 달리 R은 그래픽 기반 인터페이스(GUI)가 아닌 명령어 기반 인터페이스를 가지고 있음
  - 스크립트를 작성하는 과정에서 사소한 실수가 있는 경우, 이를 찾아내고 수정하는데(디버깅) 많은 시간이 소요될 수 있음
  - 연습을 통해 코드 작성에 익숙해지면 이후에는 빠르게 배울 수 있음
- 너무 많은 패키지 …
  - 패키지의 수가 늘어남에 따라 좋은 패키지를 선택하는 것이 어려워질 수 있음
  - 사용자 기여 패키지의 품질 제어 문제

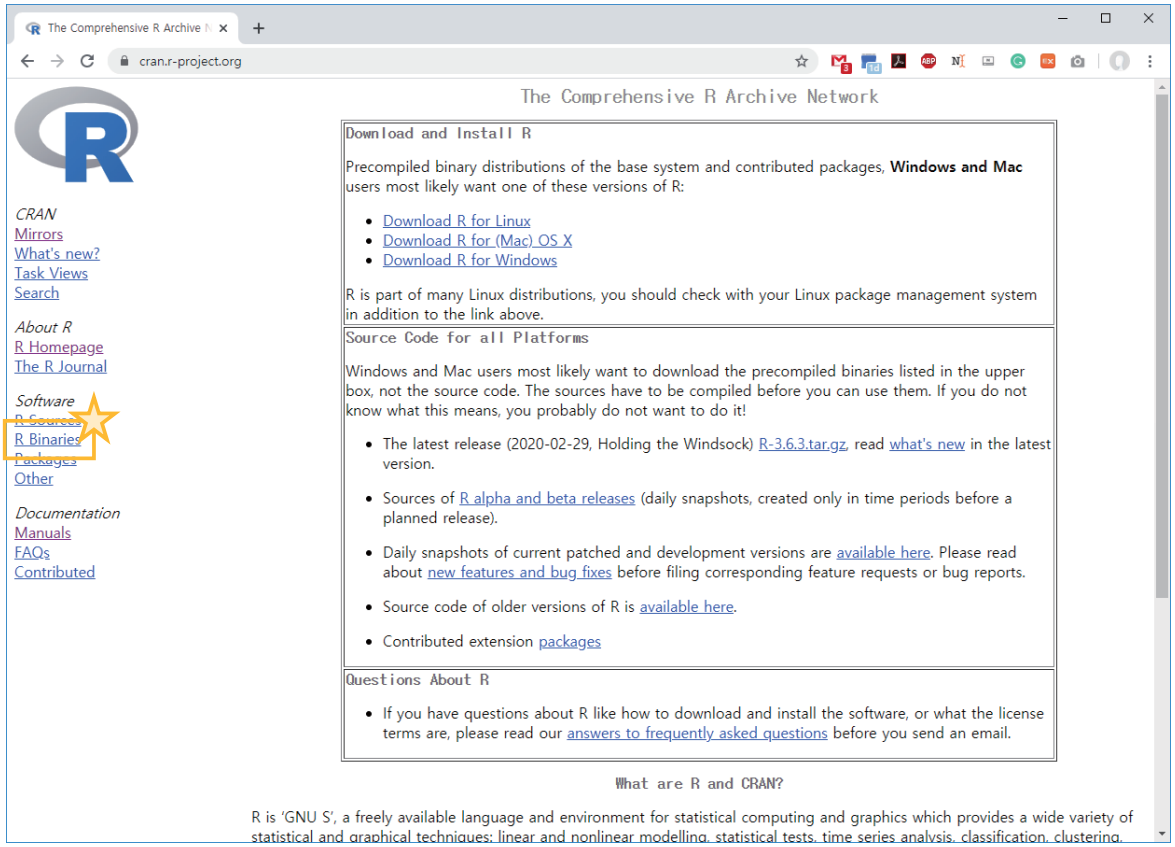
R 설치 및 다운로드

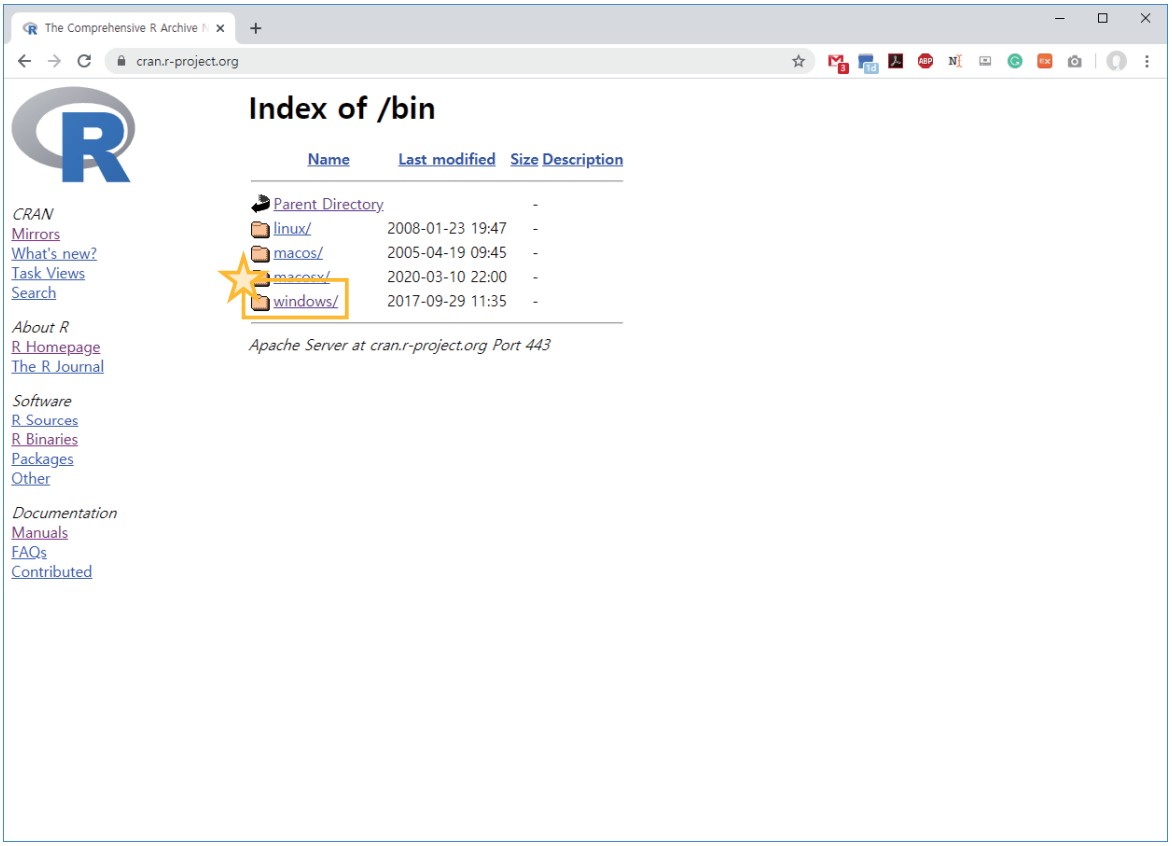
- 20여년 전에는 오른쪽 그림과 같이 CD 형태로 배포됐지만 …
- (당연하게도) 지금은 인터넷을 통해 다운로드가 가능함:
  - CRAN: https://cran.r-project.org/mirrors.html
- R의 설치와 다운로드는 사실 매우 간단하기 때문에 특별한 설명이 필요치 않음

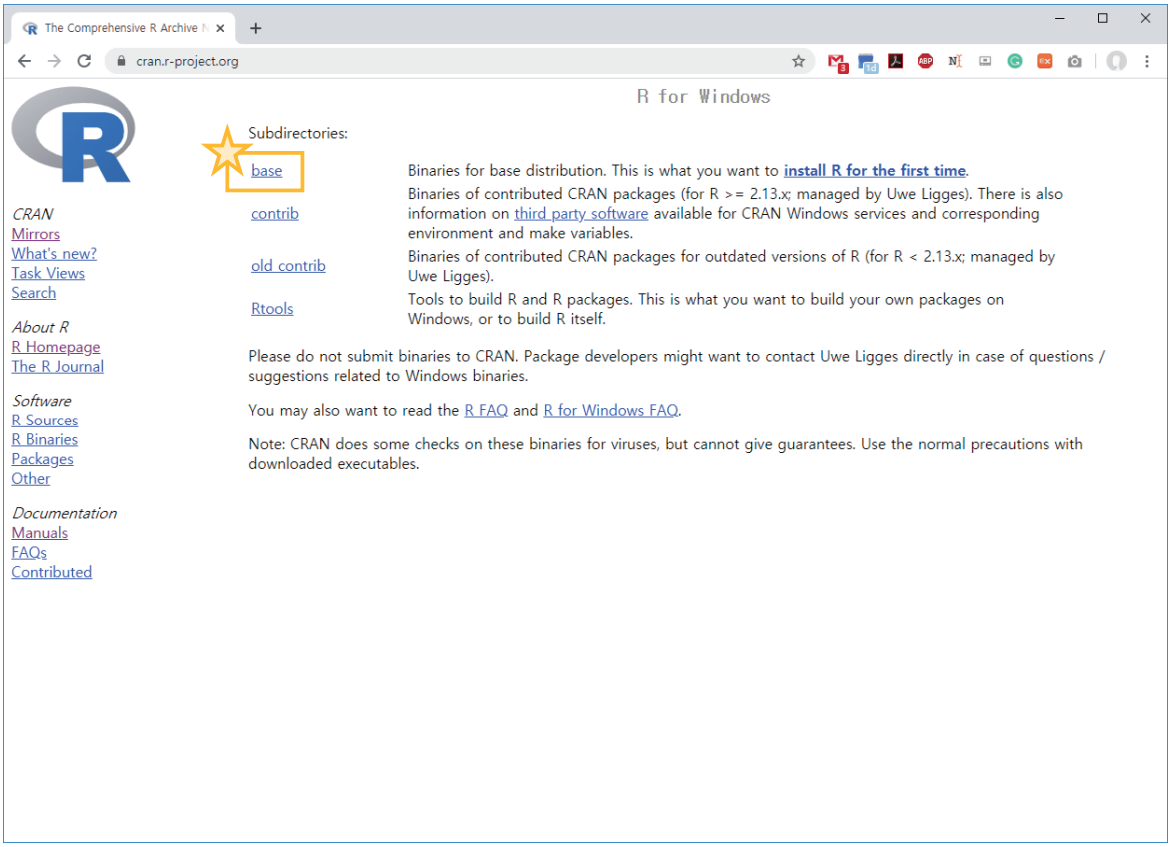


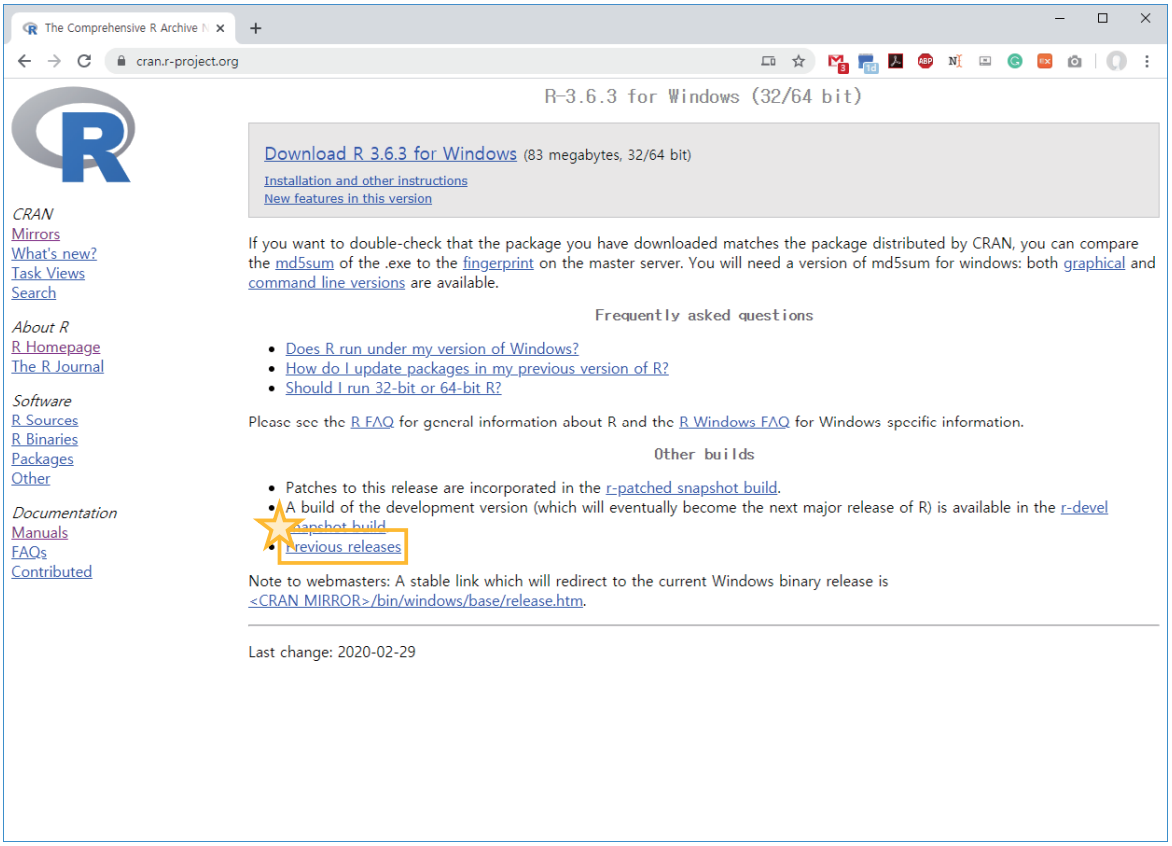

The Comprehensive R Archive
cran.r-project.org
https://cran.r-project.org/
The Comprehensive R Archive Network
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed
Download and Install R
Precompiled binary distributions of the base system and contributed packages, Windows and Mac users most likely want one of these versions of R:
Download R for Linux
Download R for (Mac) OS X
Download R for Windows
R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.
Source Code for all Platforms
Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!
The latest release (2020-02-29, Holding the Windsock) R-3.6.3.tar.gz, read what's new in the latest version.
Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
Source code of older versions of R is available here.
Contributed extension packages
Questions About R
If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.
What are R and CRAN?
R is 'GNU S', a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering,

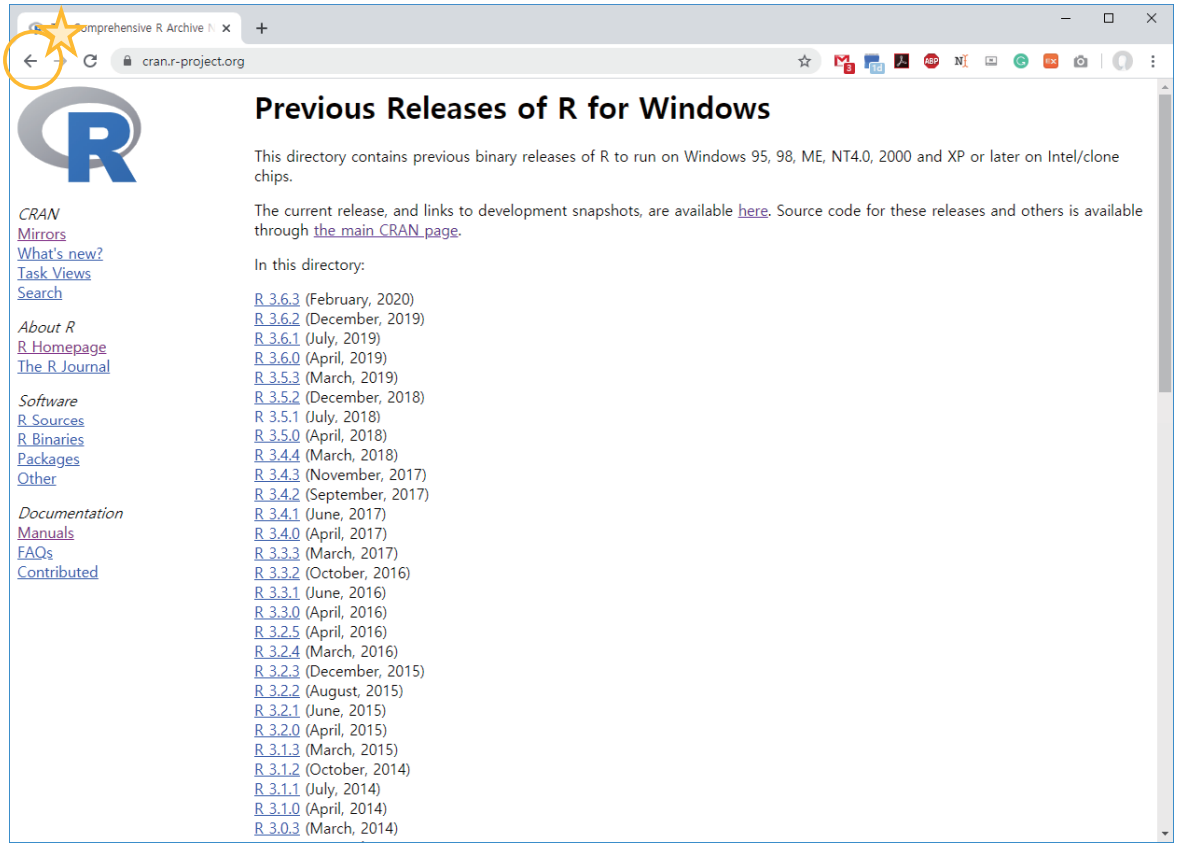

cran.r-project.org
R for Windows
Subdirectories:
base
Binaries for base distribution. This is what you want to install R for the first time.
contrib
Binaries of contributed CRAN packages (for R >= 2.13.x; managed by Uwe Ligges). There is also information on third party software available for CRAN Windows services and corresponding environment and make variables.
old contrib
Binaries of contributed CRAN packages for outdated versions of R (for R < 2.13.x; managed by Uwe Ligges).
Rtools
Tools to build R and R packages. This is what you want to build your own packages on Windows, or to build R itself.
Please do not submit binaries to CRAN. Package developers might want to contact Uwe Ligges directly in case of questions / suggestions related to Windows binaries.
You may also want to read the R FAQ and R for Windows FAQ.
Note: CRAN does some checks on these binaries for viruses, but cannot give guarantees. Use the normal precautions with downloaded executables.
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed

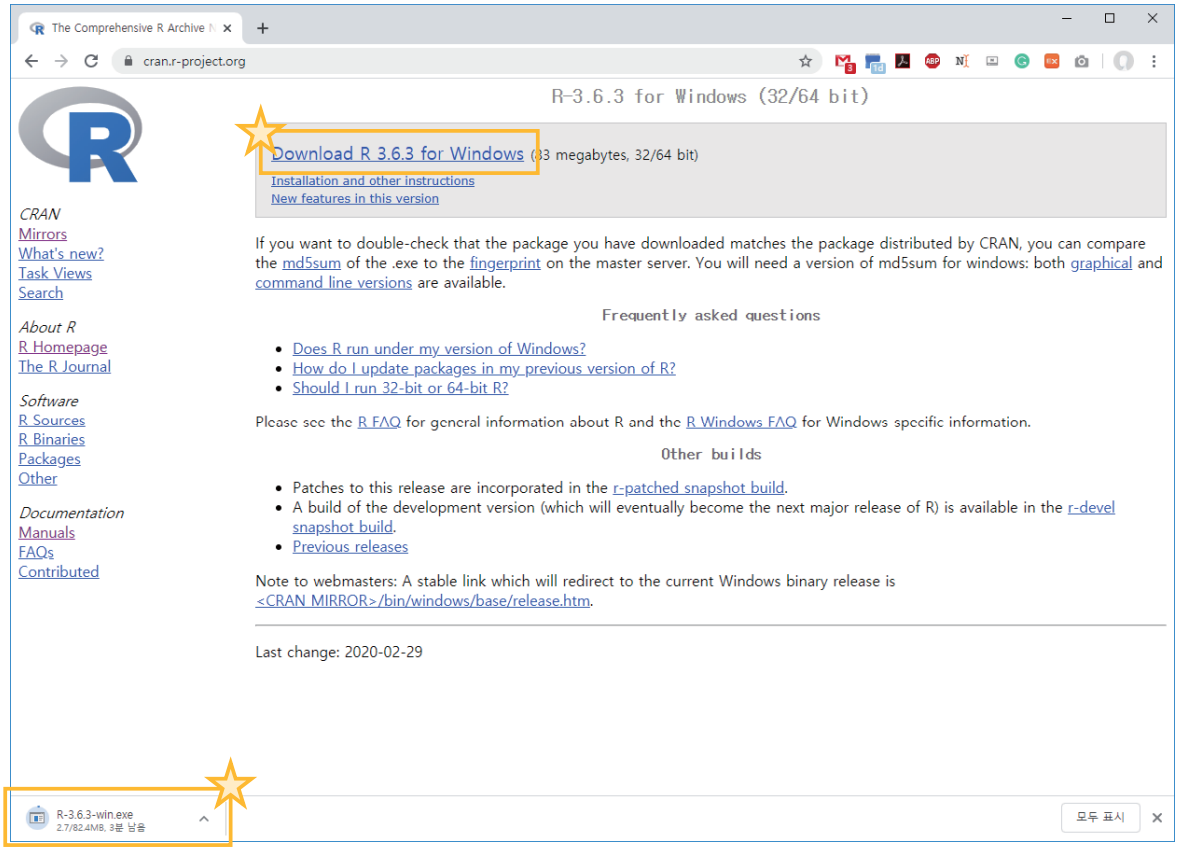

The Comprehensive R Archive
cran.r-project.org
The Comprehensive R Archive Network
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed
Download and Install R
Precompiled binary distributions of the base system and contributed packages, Windows and Mac users most likely want one of these versions of R:
Download R for Linux
Download R for (Mac) OS X
Download R for Windows
R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.
Source Code for all Platforms
Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!
The latest release (2020-02-29, Holding the Windsock) R-3.6.3.tar.gz, read what's new in the latest version.
Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
Source code of older versions of R is available here.
Contributed extension packages
Questions About R
If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.
What are R and CRAN?
R is 'GNU S', a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering,

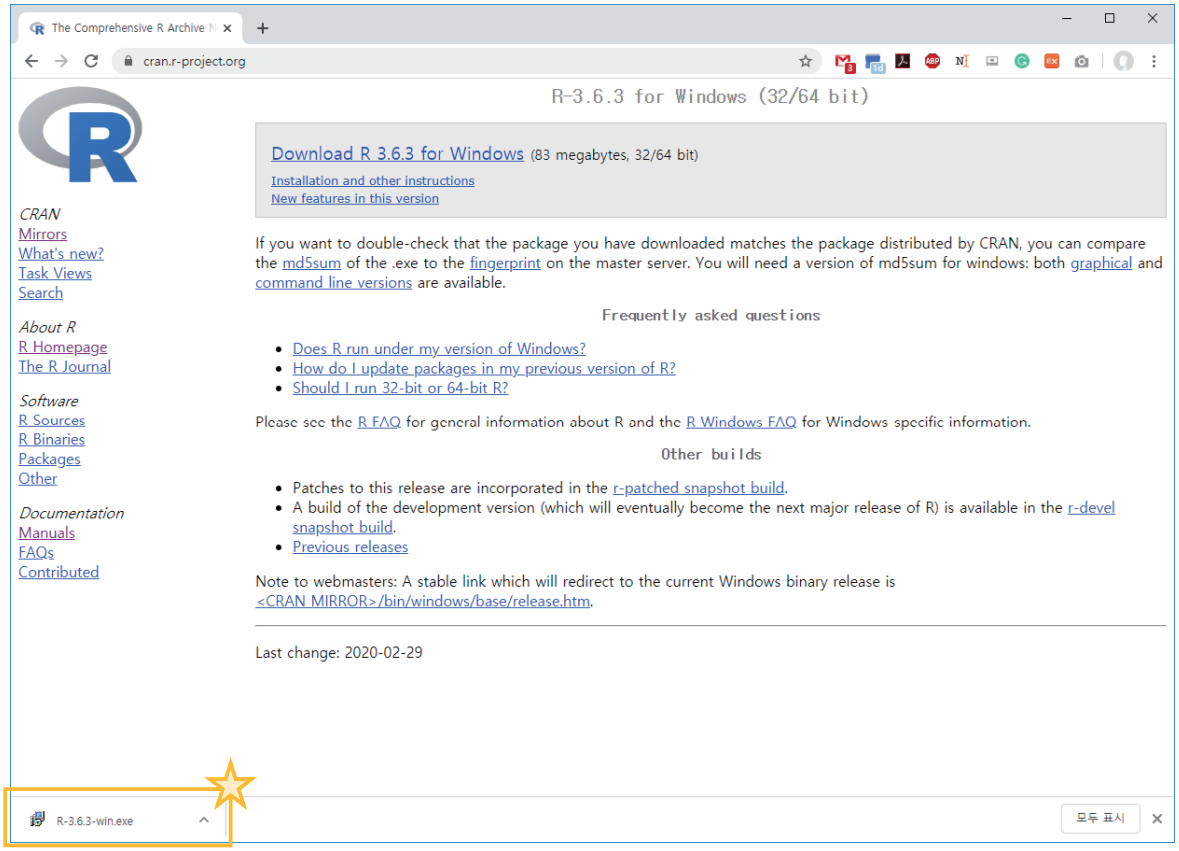

The Comprehensive R Archive
cran.r-project.org
Index of /bin
Name
Last modified
Size
Description
Parent Directory -
linux/ 2008-01-23 19:47 -
macos/ 2005-04-19 09:45 -
macosx/ 2020-03-10 22:00 -
windows/ 2017-09-29 11:35 -
Apache Server at cran.r-project.org Port 443
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed

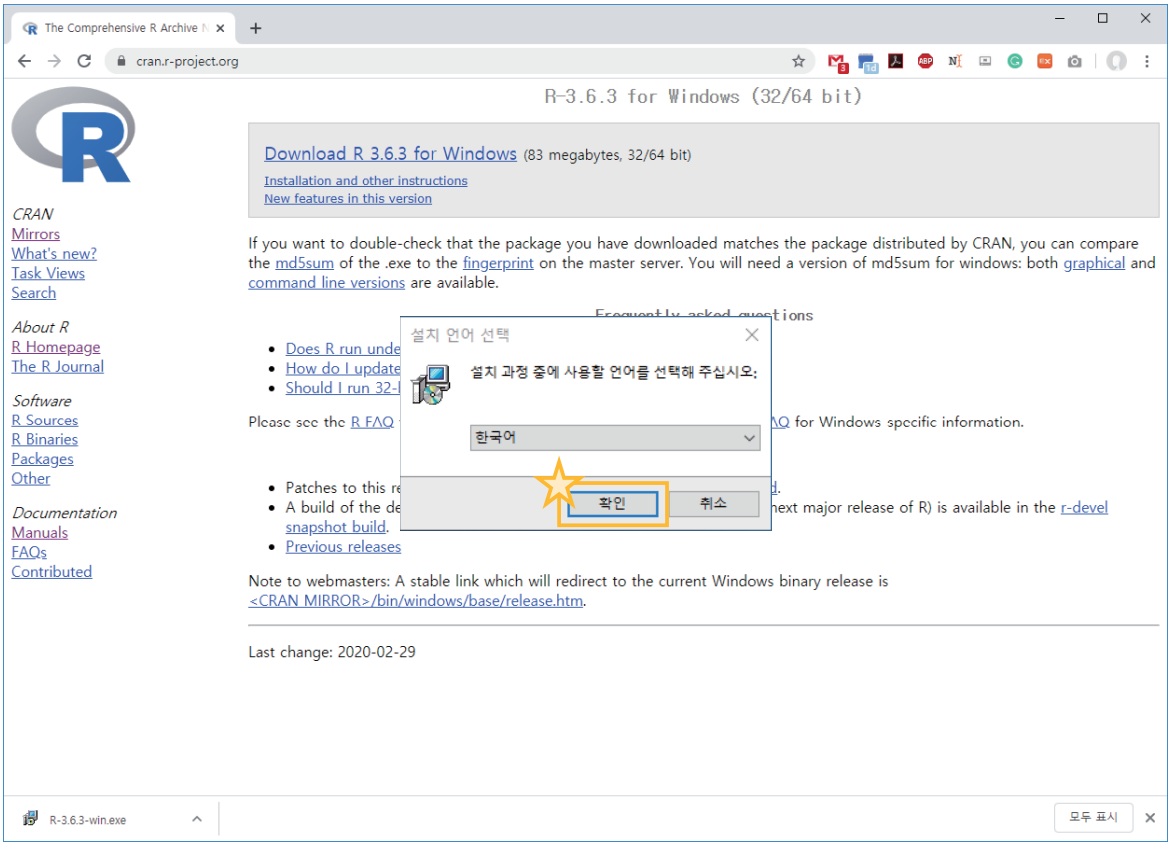

The Comprehensive R Archive
cran.r-project.org
R for Windows
Subdirectories:
base
Binaries for base distribution. This is what you want to install R for the first time.
contrib
Binaries of contributed CRAN packages (for R >= 2.13.x; managed by Uwe Ligges). There is also information on third party software available for CRAN Windows services and corresponding environment and make variables.
old contrib
Binaries of contributed CRAN packages for outdated versions of R (for R < 2.13.x; managed by Uwe Ligges).
Rtools
Tools to build R and R packages. This is what you want to build your own packages on Windows, or to build R itself.
Please do not submit binaries to CRAN. Package developers might want to contact Uwe Ligges directly in case of questions / suggestions related to Windows binaries.
You may also want to read the R FAQ and R for Windows FAQ.
Note: CRAN does some checks on these binaries for viruses, but cannot give guarantees. Use the normal precautions with downloaded executables.
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed

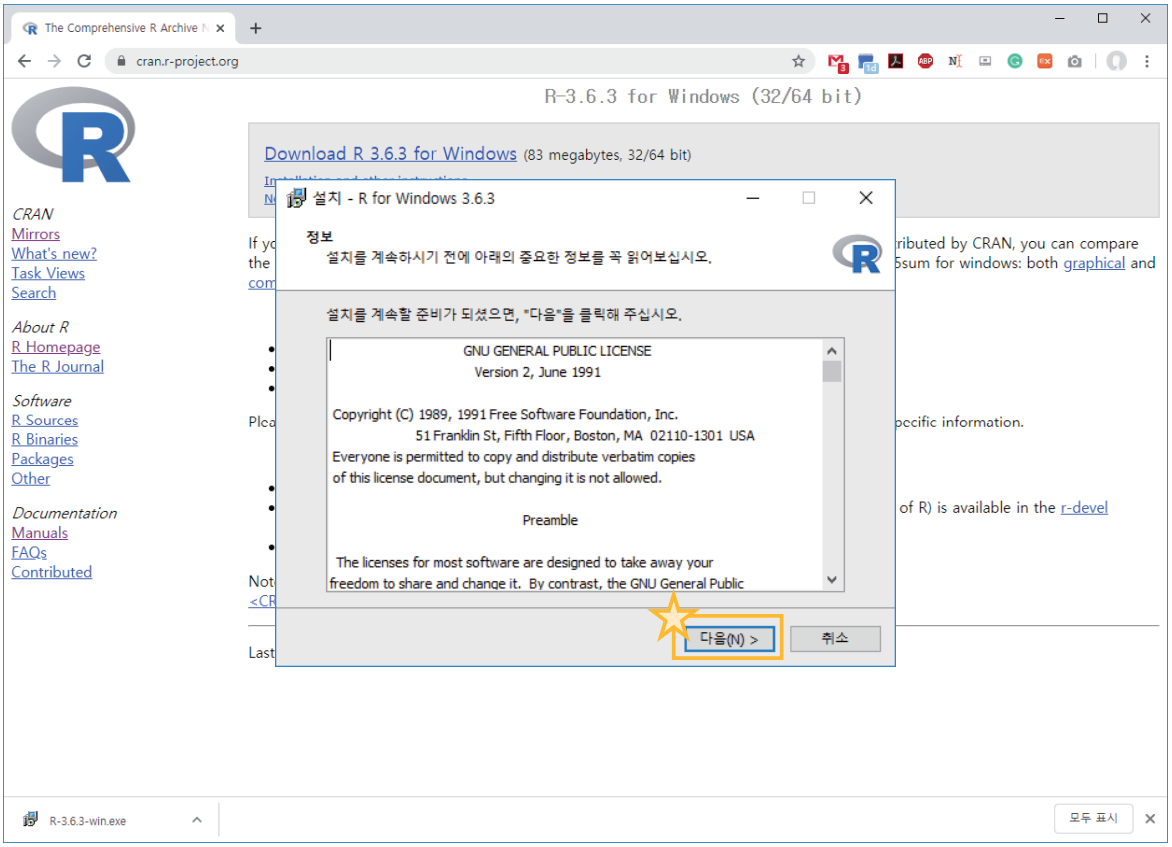

The Comprehensive R Archive
cran.r-project.org
R-3.6.3 for Windows (32/64 bit)
Download R 3.6.3 for Windows (83 megabytes, 32/64 bit)
Installation and other instructions
New features in this version
If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the md5sum of the .exe to the fingerprint on the master server. You will need a version of md5sum for windows: both graphical and command line versions are available.
Frequently asked questions
Does R run under my version of Windows?
How do I update packages in my previous version of R?
Should I run 32-bit or 64-bit R?
Please see the R FAQ for general information about R and the R Windows FAQ for Windows specific information.
Other builds
Patches to this release are incorporated in the r-patched snapshot build.
A build of the development version (which will eventually become the next major release of R) is available in the r-devel snapshot build.
Previous releases
Note to webmasters: A stable link which will redirect to the current Windows binary release is <CRAN MIRROR>/bin/windows/base/release.htm.
Last change: 2020-02-29
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed

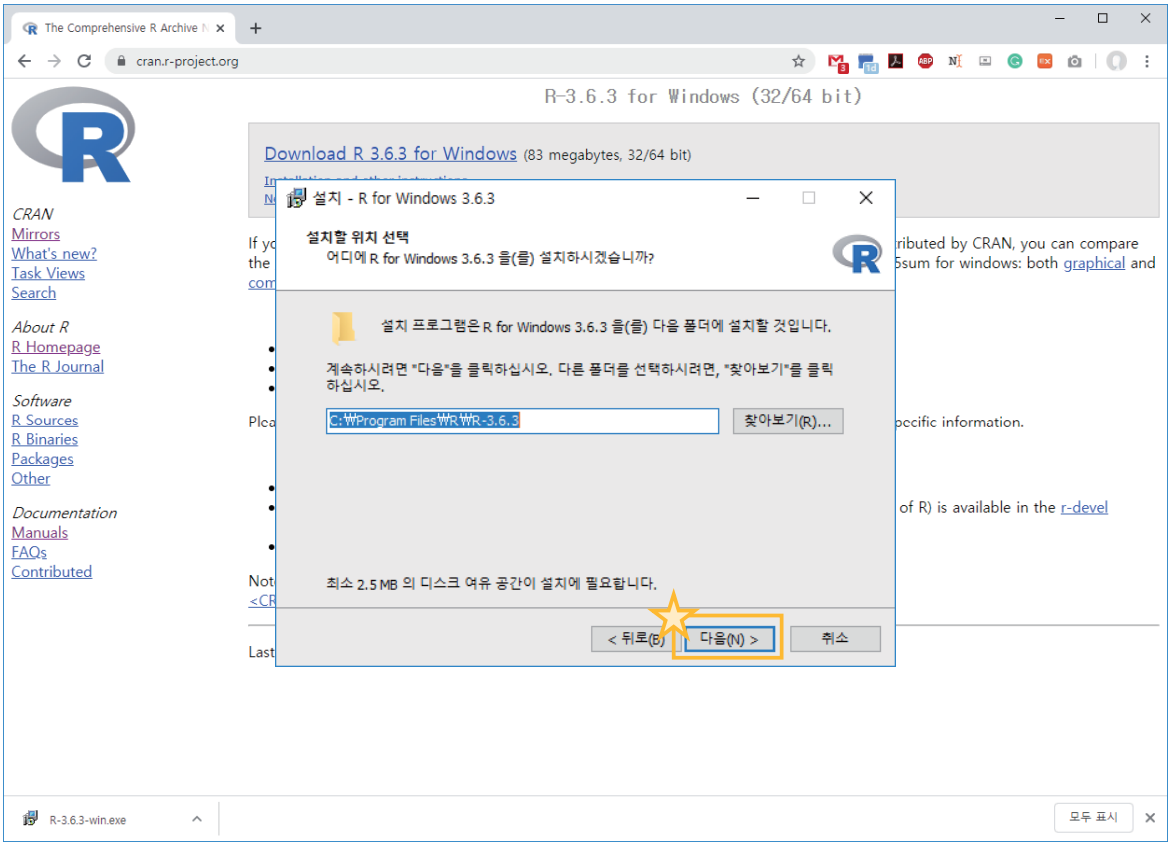

cran.r-project.org
Previous Releases of R for Windows
This directory contains previous binary releases of R to run on Windows 95, 98, ME, NT4.0, 2000 and XP or later on Intel/clone chips.
The current release, and links to development snapshots, are available here. Source code for these releases and others is available through the main CRAN page.
In this directory:
R 3.6.3 (February, 2020)
R 3.6.2 (December, 2019)
R 3.6.1 (July, 2019)
R 3.6.0 (April, 2019)
R 3.5.3 (March, 2019)
R 3.5.2 (December, 2018)
R 3.5.1 (July, 2018)
R 3.5.0 (April, 2018)
R 3.4.4 (March, 2018)
R 3.4.3 (November, 2017)
R 3.4.2 (September, 2017)
R 3.4.1 (June, 2017)
R 3.4.0 (April, 2017)
R 3.3.3 (March, 2017)
R 3.3.2 (October, 2016)
R 3.3.1 (June, 2016)
R 3.3.0 (April, 2016)
R 3.2.5 (April, 2016)
R 3.2.4 (March, 2016)
R 3.2.3 (December, 2015)
R 3.2.2 (August, 2015)
R 3.2.1 (June, 2015)
R 3.2.0 (April, 2015)
R 3.1.3 (March, 2015)
R 3.1.2 (October, 2014)
R 3.1.1 (July, 2014)
R 3.1.0 (April, 2014)
R 3.0.3 (March, 2014)
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed

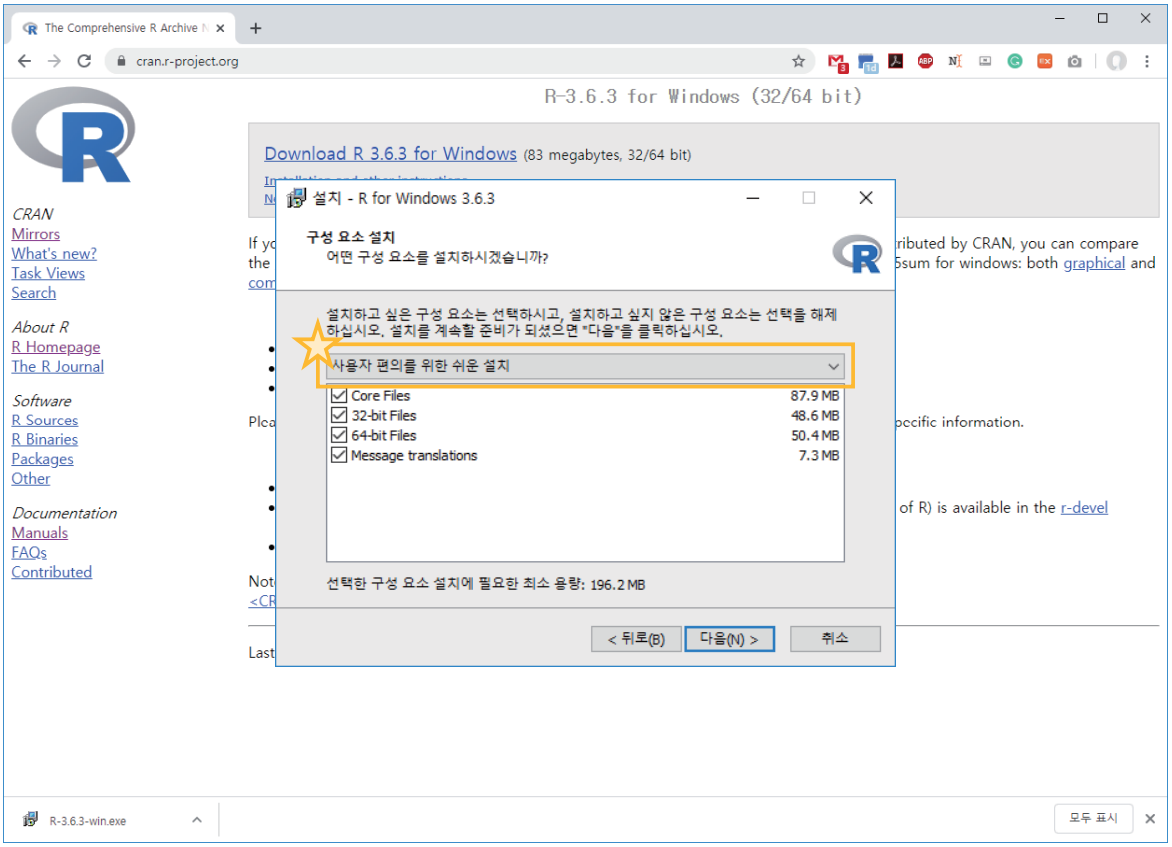

The Comprehensive R Archive
cran.r-project.org
R-3.6.3 for Windows (32/64 bit)
Download R 3.6.3 for Windows (83 megabytes, 32/64 bit)
Installation and other instructions
New features in this version
If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the md5sum of the .exe to the fingerprint on the master server. You will need a version of md5sum for windows: both graphical and command line versions are available.
Frequently asked questions
Does R run under my version of Windows?
How do I update packages in my previous version of R?
Should I run 32-bit or 64-bit R?
Please see the R FAQ for general information about R and the R Windows FAQ for Windows specific information.
Other builds
Patches to this release are incorporated in the r-patched snapshot build.
A build of the development version (which will eventually become the next major release of R) is available in the r-devel snapshot build.
Previous releases
Note to webmasters: A stable link which will redirect to the current Windows binary release is <CRAN MIRROR>/bin/windows/base/release.htm.
Last change: 2020-02-29
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed
R-3.6.3-win.exe
2.7/82.4MB, 3분 남음
모두 표시

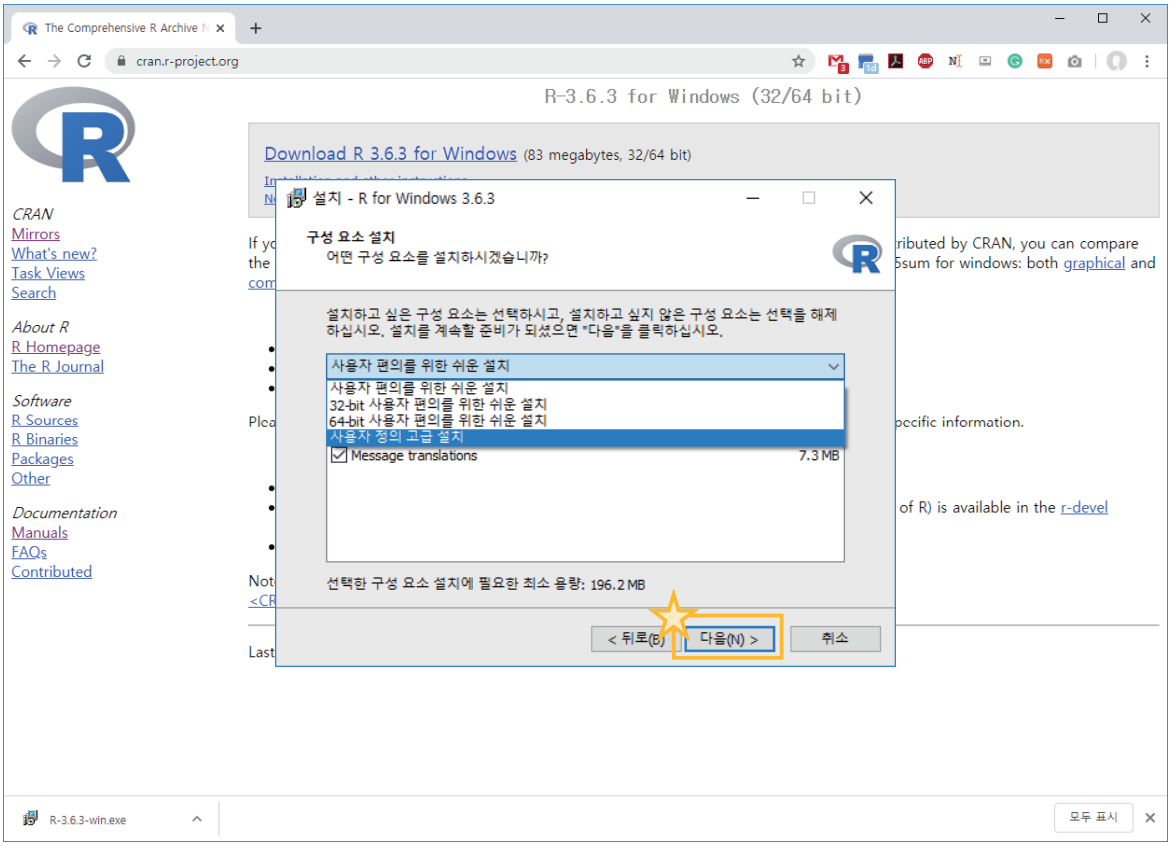

The Comprehensive R Archive
cran.r-project.org
R-3.6.3 for Windows (32/64 bit)
Download R 3.6.3 for Windows (83 megabytes, 32/64 bit)
Installation and other instructions
New features in this version
If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the md5sum of the .exe to the fingerprint on the master server. You will need a version of md5sum for windows: both graphical and command line versions are available.
Frequently asked questions
Does R run under my version of Windows?
How do I update packages in my previous version of R?
Should I run 32-bit or 64-bit R?
Please see the R FAQ for general information about R and the R Windows FAQ for Windows specific information.
Other builds
Patches to this release are incorporated in the r-patched snapshot build.
A build of the development version (which will eventually become the next major release of R) is available in the r-devel snapshot build.
Previous releases
Note to webmasters: A stable link which will redirect to the current Windows binary release is <CRAN MIRROR>/bin/windows/base/release.htm.
Last change: 2020-02-29
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed
R-3.6.3-win.exe
모두 표시

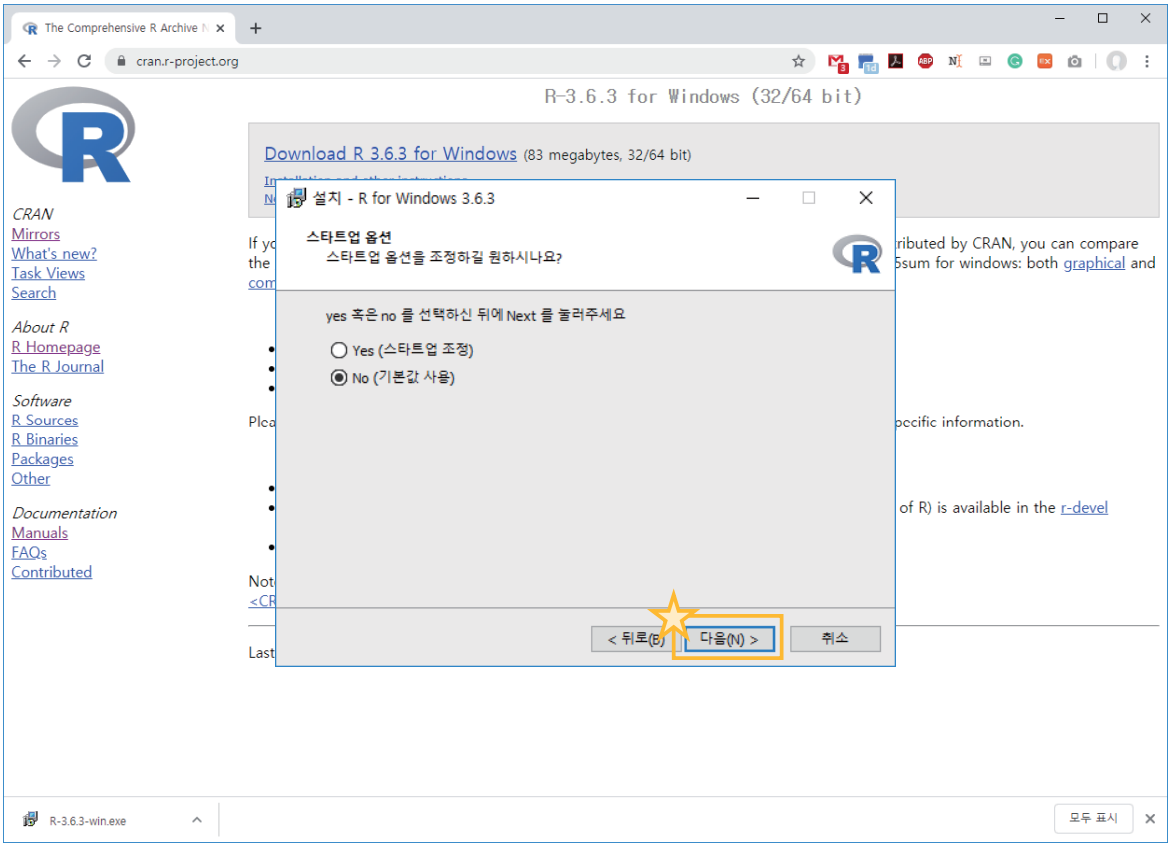

The Comprehensive R Archive
cran.r-project.org
R-3.6.3 for Windows (32/64 bit)
Download R 3.6.3 for Windows (83 megabytes, 32/64 bit)
Installation and other instructions
New features in this version
If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the md5sum of the .exe to the fingerprint on the master server. You will need a version of md5sum for windows: both graphical and command line versions are available.
설치 언어 선택
설치 과정 중에 사용할 언어를 선택해 주십시오:
한국어
확인
취소
Note to webmasters: A stable link which will redirect to the current Windows binary release is <CRAN MIRROR>/bin/windows/base/release.htm.
Last change: 2020-02-29
R-3.6.3-win.exe
모두 표시

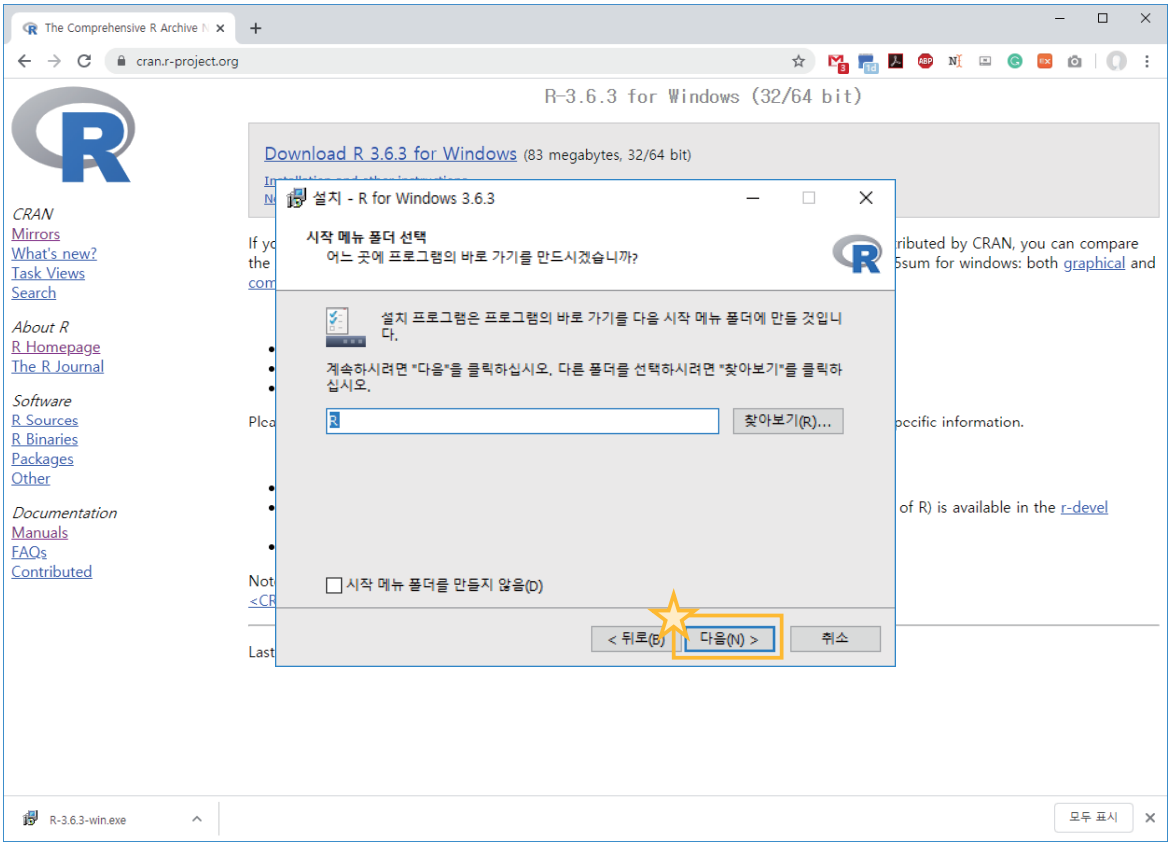

The Comprehensive R Archive
cran.r-project.org
R-3.6.3 for Windows (32/64 bit)
Download R 3.6.3 for Windows (83 megabytes, 32/64 bit)
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed
설치 - R for Windows 3.6.3
정보
설치를 계속하시기 전에 아래의 중요한 정보를 꼭 읽어보십시오.
설치를 계속할 준비가 되셨으면, "다음"을 클릭해 주십시오.
GNU GENERAL PUBLIC LICENSE
Version 2, June 1991
Copyright (C) 1989, 1991 Free Software Foundation, Inc.
51 Franklin St, Fifth Floor, Boston, MA 02110-1301 USA
Everyone is permitted to copy and distribute verbatim copies
of this license document, but changing it is not allowed.
Preamble
The licenses for most software are designed to take away your
freedom to share and change it. By contrast, the GNU General Public
다음(N) >
취소
R-3.6.3-win.exe
모두 표시

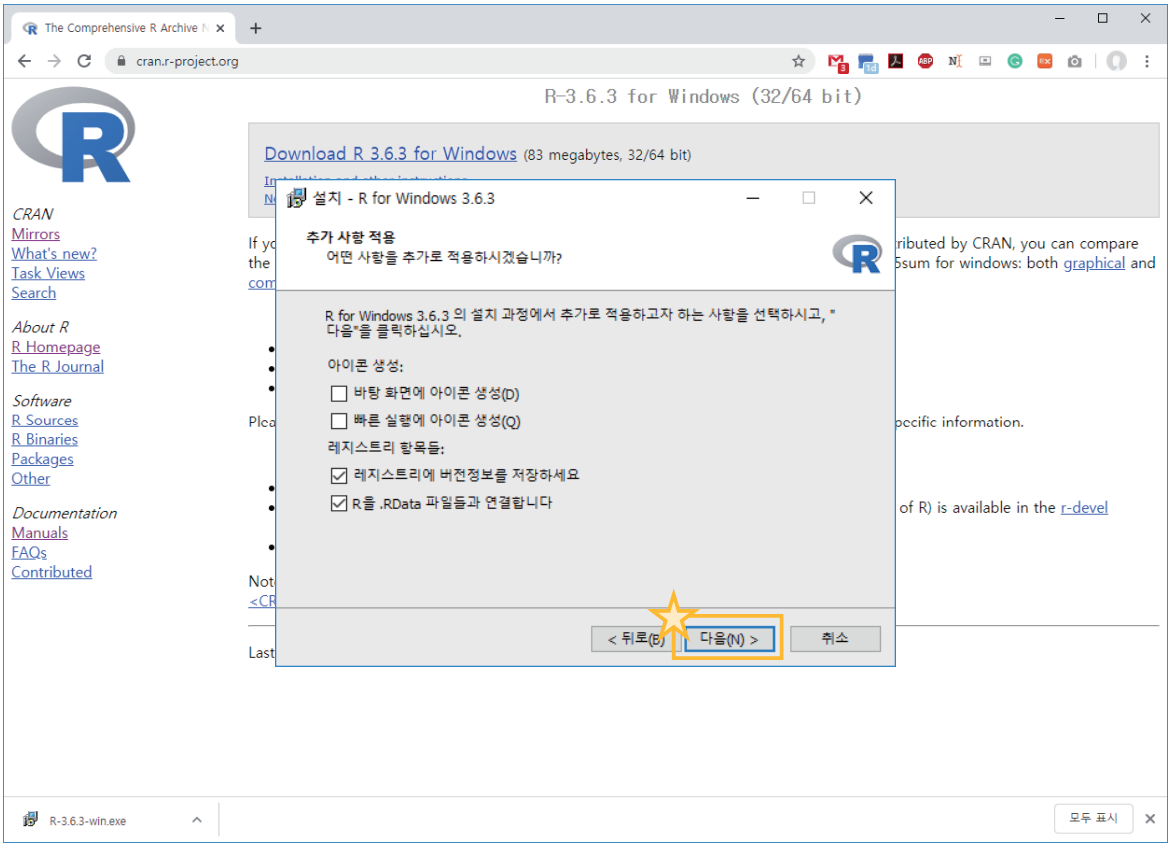

The Comprehensive R Archive
cran.r-project.org
R-3.6.3 for Windows (32/64 bit)
Download R 3.6.3 for Windows (83 megabytes, 32/64 bit)
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed
설치 - R for Windows 3.6.3
설치할 위치 선택
어디에 R for Windows 3.6.3 을(를) 설치하시겠습니까?
설치 프로그램은 R for Windows 3.6.3 을(를) 다음 폴더에 설치할 것입니다.
계속하시려면 "다음"을 클릭하십시오. 다른 폴더를 선택하시려면, "찾아보기"를 클릭하십시오.
C:₩Program Files₩R₩R-3.6.3
찾아보기(R)...
최소 2.5 MB 의 디스크 여유 공간이 설치에 필요합니다.
< 뒤로(B)
다음(N) >
취소
R-3.6.3-win.exe
모두 표시

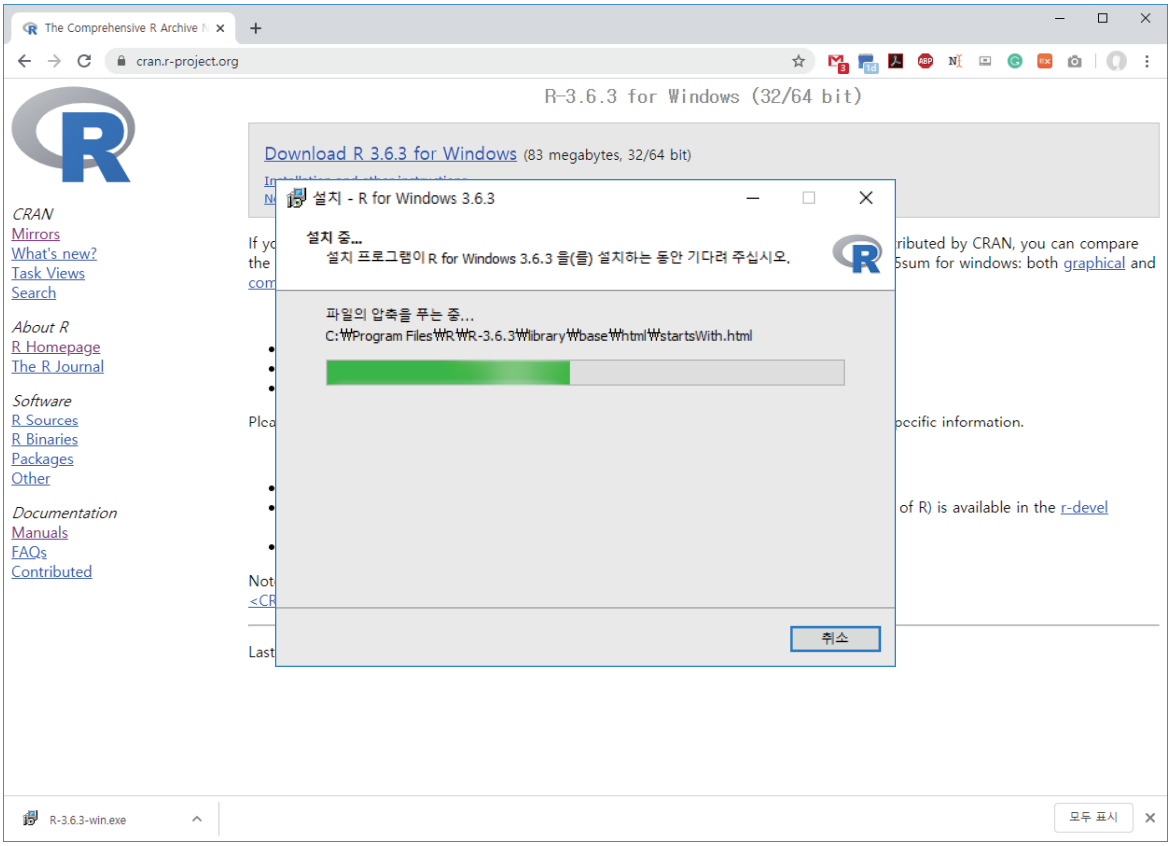

The Comprehensive R Archive
cran.r-project.org
R-3.6.3 for Windows (32/64 bit)
Download R 3.6.3 for Windows (83 megabytes, 32/64 bit)
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed
설치 - R for Windows 3.6.3
구성 요소 설치
어떤 구성 요소를 설치하시겠습니까?
설치하고 싶은 구성 요소는 선택하시고, 설치하고 싶지 않은 구성 요소는 선택을 해제하십시오. 설치를 계속할 준비가 되셨으면 "다음"을 클릭하십시오.
사용자 편의를 위한 쉬운 설치
Core Files 87.9 MB
32-bit Files 48.6 MB
64-bit Files 50.4 MB
Message translations 7.3 MB
선택한 구성 요소 설치에 필요한 최소 용량: 196.2 MB
< 뒤로(B)
다음(N) >
취소
R-3.6.3-win.exe
모두 표시

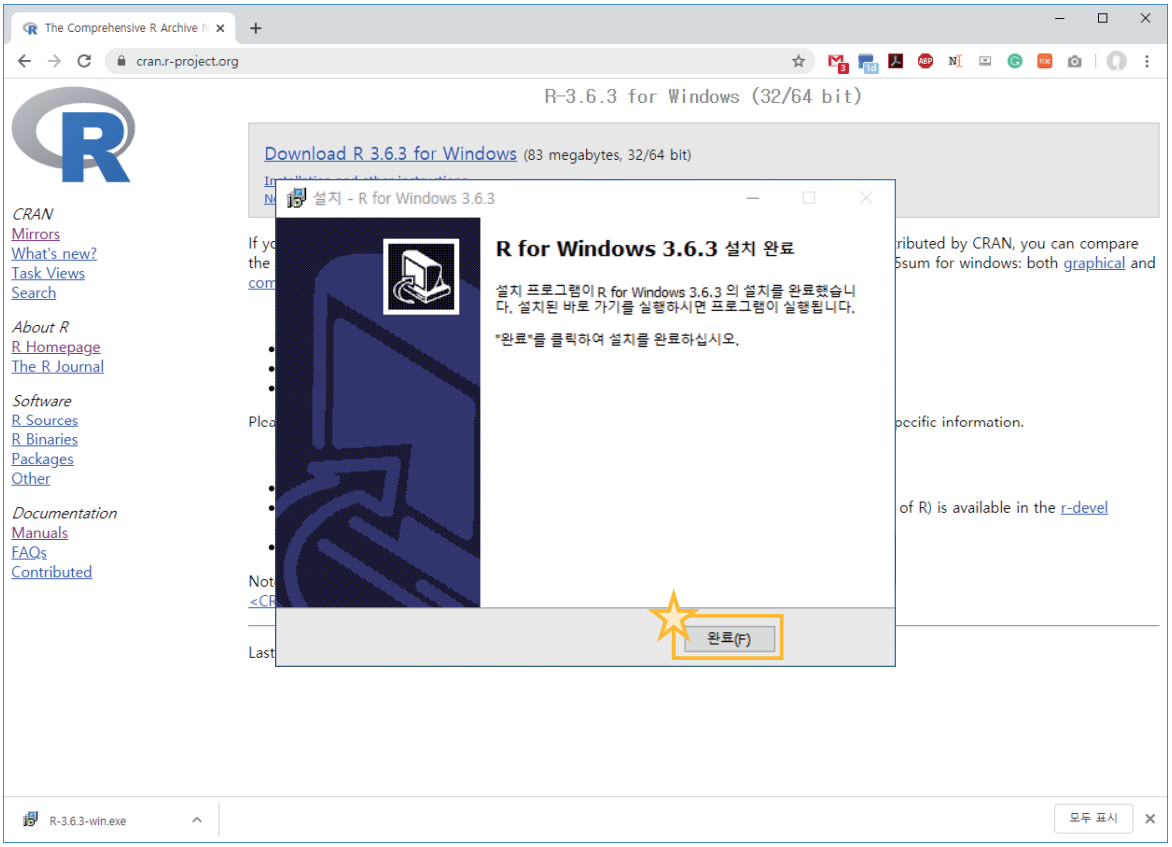

The Comprehensive R Archive
cran.r-project.org
R-3.6.3 for Windows (32/64 bit)
Download R 3.6.3 for Windows (83 megabytes, 32/64 bit)
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed
설치 - R for Windows 3.6.3
구성 요소 설치
어떤 구성 요소를 설치하시겠습니까?
설치하고 싶은 구성 요소는 선택하시고, 설치하고 싶지 않은 구성 요소는 선택을 해제하십시오. 설치를 계속할 준비가 되셨으면 "다음"을 클릭하십시오.
사용자 편의를 위한 쉬운 설치
사용자 편의를 위한 쉬운 설치
32-bit 사용자 편의를 위한 쉬운 설치
64-bit 사용자 편의를 위한 쉬운 설치
사용자 정의 고급 설치
Message translations 7.3 MB
선택한 구성 요소 설치에 필요한 최소 용량: 196.2 MB
< 뒤로(B)
다음(N) >
취소
R-3.6.3-win.exe
모두 표시


The Comprehensive R Archive
cran.r-project.org
R-3.6.3 for Windows (32/64 bit)
Download R 3.6.3 for Windows (83 megabytes, 32/64 bit)
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed
설치 - R for Windows 3.6.3
스타트업 옵션
스타트업 옵션을 조정하길 원하시나요?
yes 혹은 no 를 선택하신 뒤에 Next 를 눌러주세요
Yes (스타트업 조정)
No (기본값 사용)
< 뒤로(B)
다음(N) >
취소
R-3.6.3-win.exe
모두 표시


The Comprehensive R Archive
cran.r-project.org
R-3.6.3 for Windows (32/64 bit)
Download R 3.6.3 for Windows (83 megabytes, 32/64 bit)
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed
설치 - R for Windows 3.6.3
시작 메뉴 폴더 선택
어느 곳에 프로그램의 바로 가기를 만드시겠습니까?
설치 프로그램은 프로그램의 바로 가기를 다음 시작 메뉴 폴더에 만들 것입니다.
계속하시려면 "다음"을 클릭하십시오. 다른 폴더를 선택하시려면 "찾아보기"를 클릭하십시오.
찾아보기(R)...
시작 메뉴 폴더를 만들지 않음(D)
< 뒤로(B)
다음(N) >
취소
R-3.6.3-win.exe
모두 표시

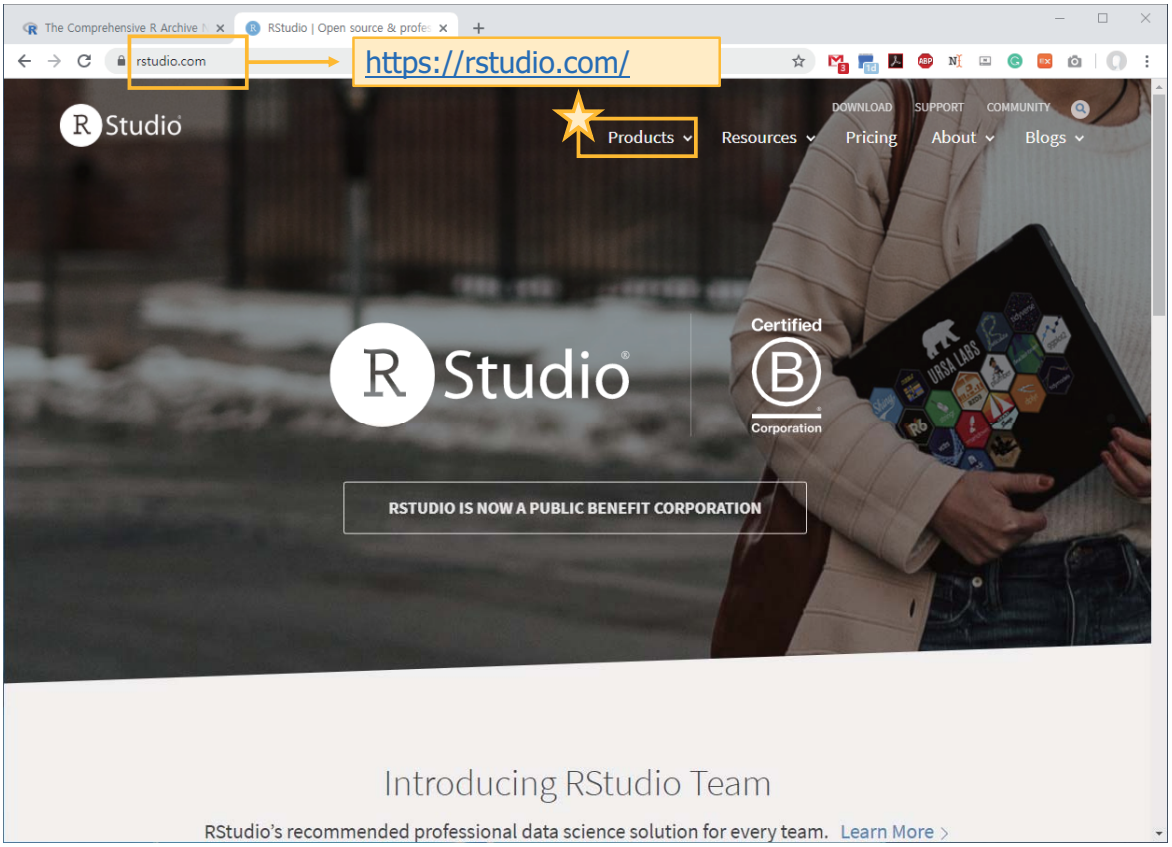

The Comprehensive R Archive
cran.r-project.org
R-3.6.3 for Windows (32/64 bit)
Download R 3.6.3 for Windows (83 megabytes, 32/64 bit)
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed
설치 - R for Windows 3.6.3
추가 사항 적용
어떤 사항을 추가로 적용하시겠습니까?
R for Windows 3.6.3 의 설치 과정에서 추가로 적용하고자 하는 사항을 선택하시고, "다음"을 클릭하십시오.
아이콘 생성:
바탕 화면에 아이콘 생성(D)
빠른 실행에 아이콘 생성(Q)
레지스트리 항목들:
레지스트리에 버전정보를 저장하세요
R을 .RData 파일들과 연결합니다
< 뒤로(B)
다음(N) >
취소
R-3.6.3-win.exe
모두 표시

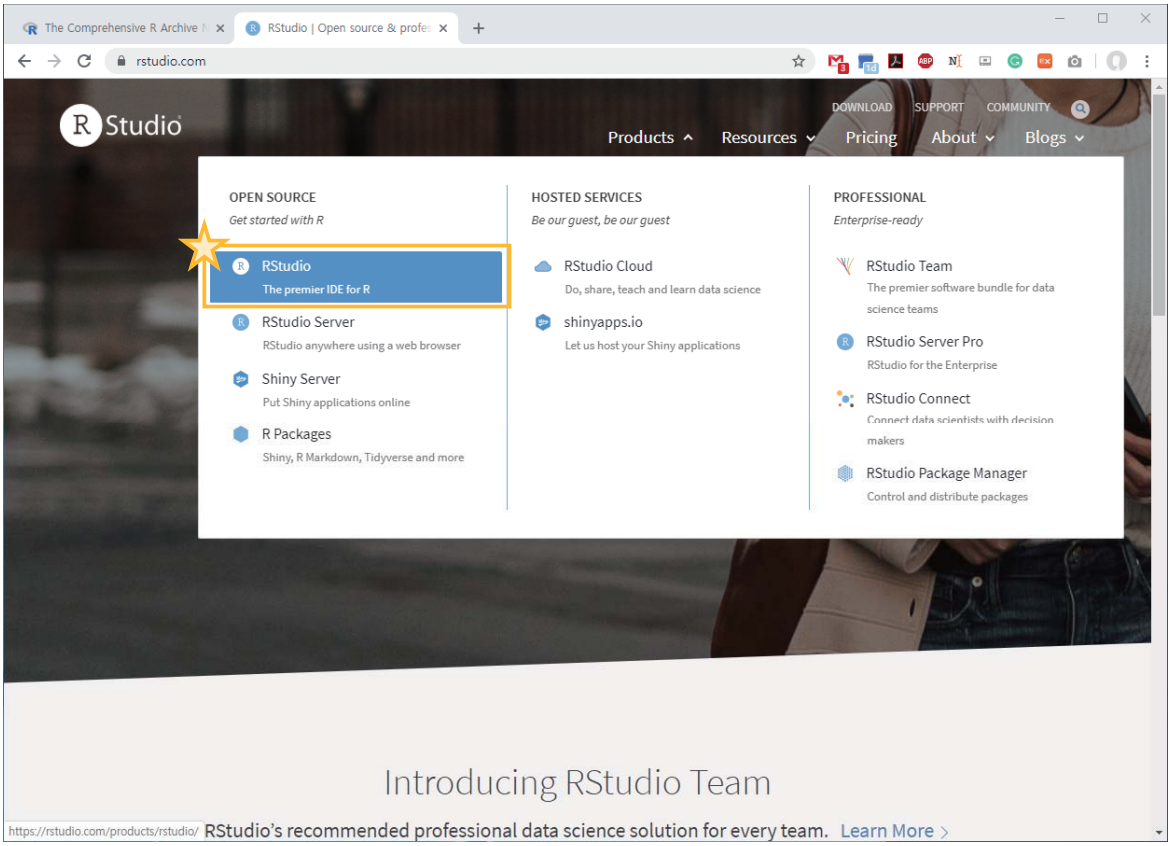

The Comprehensive R Archive
cran.r-project.org
R-3.6.3 for Windows (32/64 bit)
Download R 3.6.3 for Windows (83 megabytes, 32/64 bit)
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed
설치 - R for Windows 3.6.3
설치 중...
설치 프로그램이 R for Windows 3.6.3 을(를) 설치하는 동안 기다려 주십시오.
파일의 압축을 푸는 중...
C:₩Program Files₩R₩R-3.6.3₩library₩base₩html₩startsWith.html
취소
R-3.6.3-win.exe
모두 표시

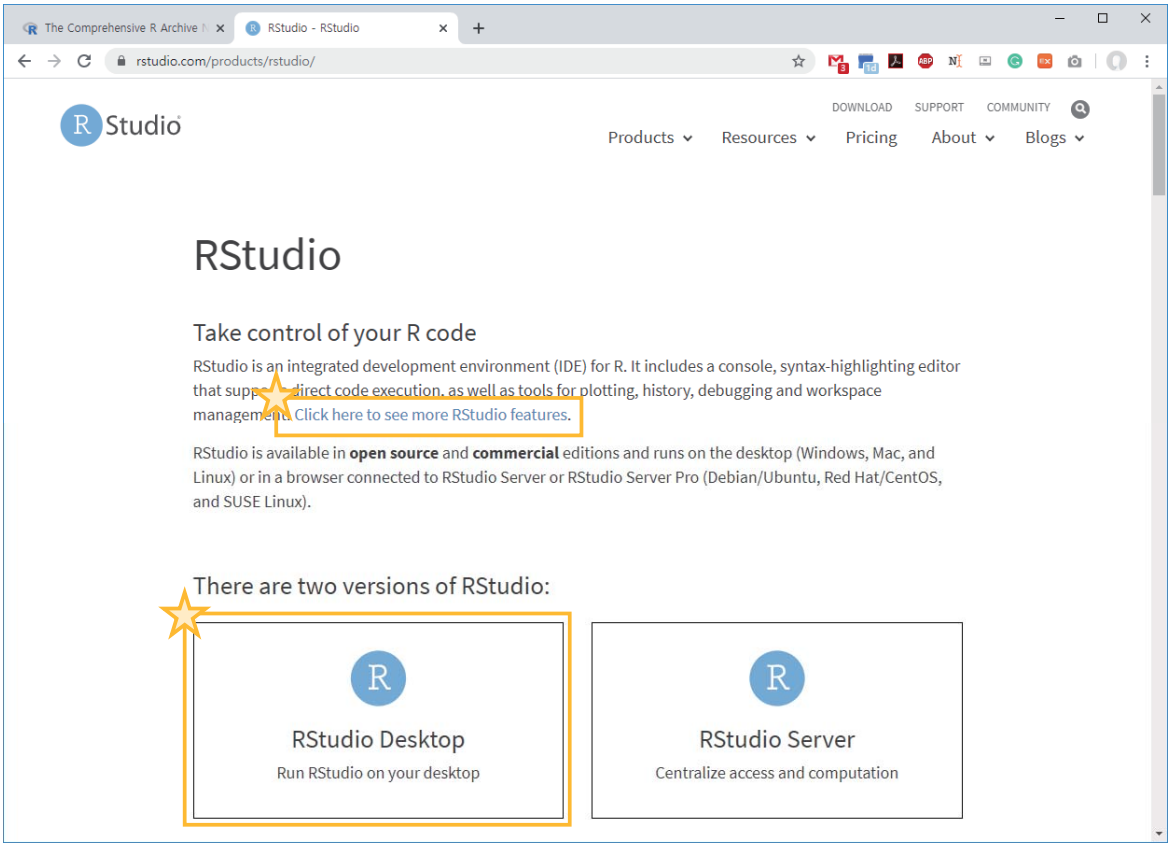


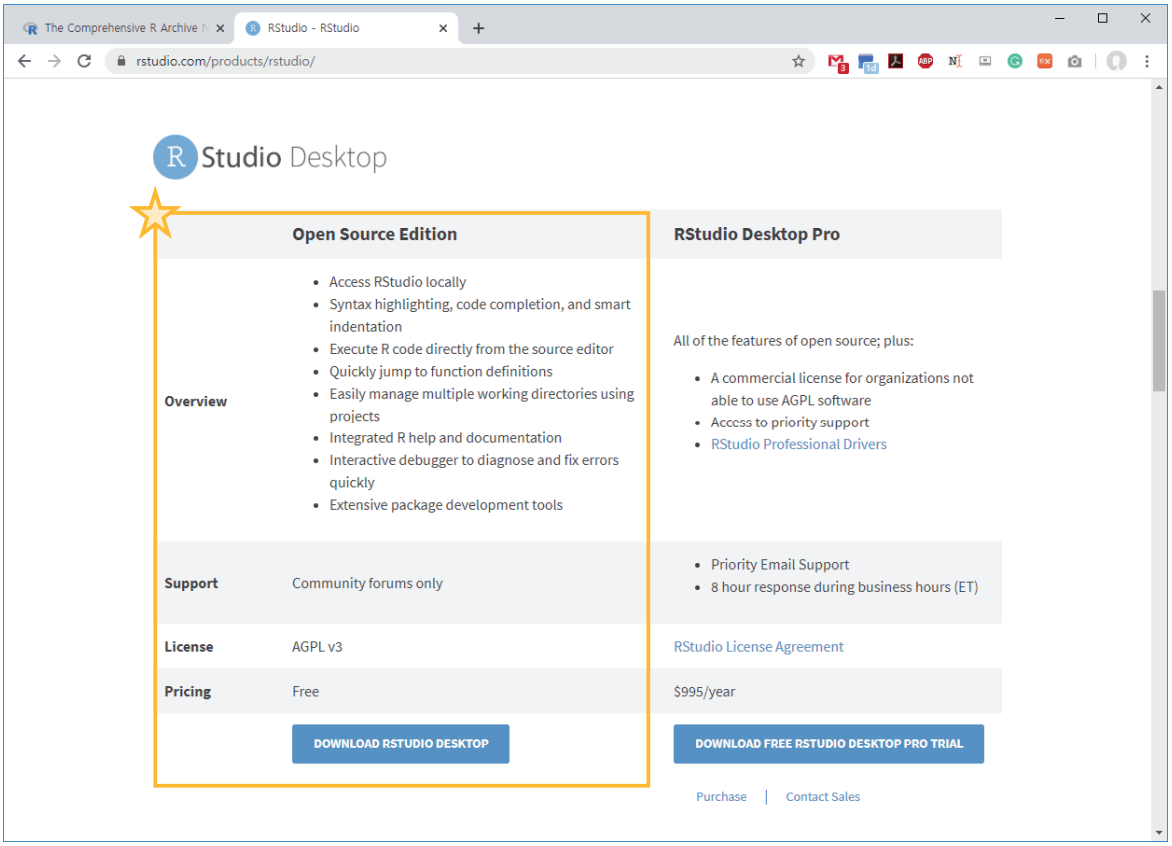
RStudio 설치 및 다운로드

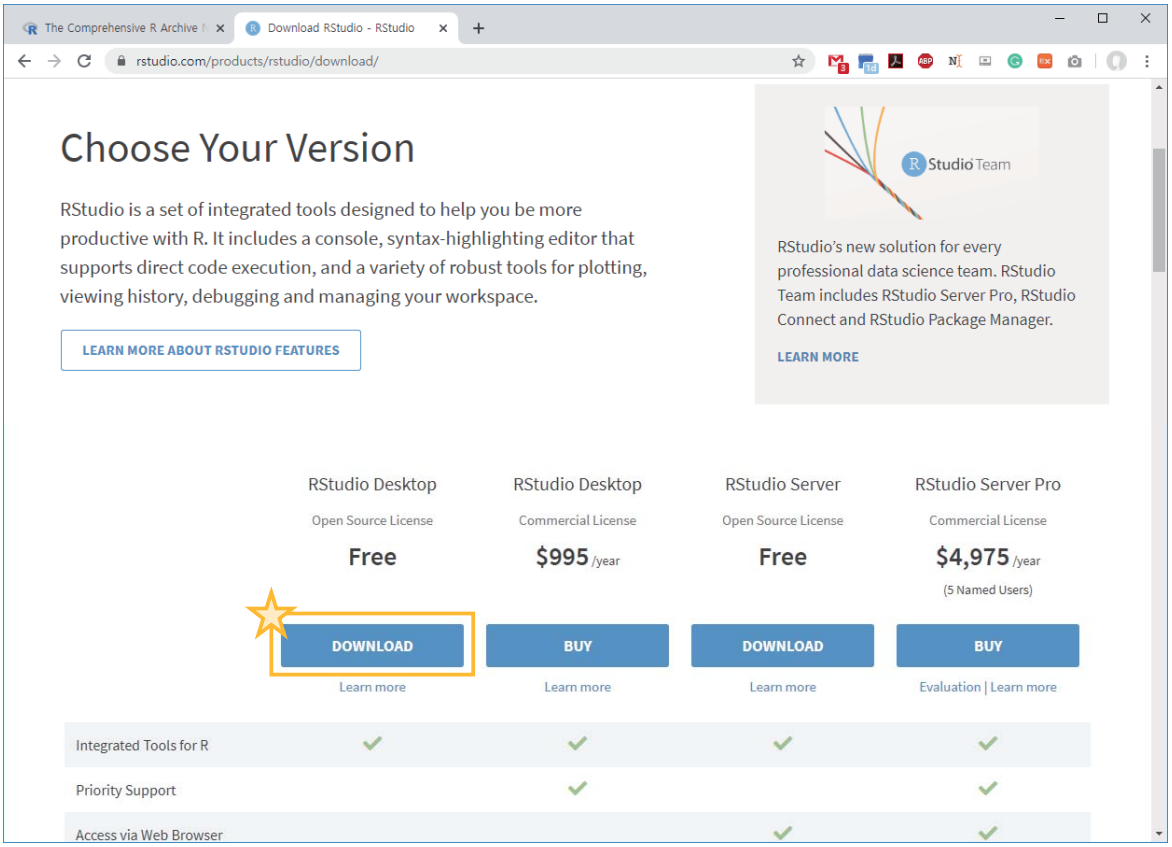
- R 프로그램의 사용을 도와주는 통합 개발 환경(integrated development environment; IDE)을 제공
  - 코드를 입력하고 실행할 수 있는 콘솔창과 저장된 코드를 불러오거나 편집할 수 있는 코드 편집기, 작업 환경, 코드 실행 내역, 그래프 등을 확인할 수 있는 다양한 창(panes)으로 구성되어 있음
  - 사용 목적 및 대상에 따라 판매용 라이선스와 오픈소스 라이선스 중 선택이 가능함(우리 수업에서는 오픈소스 버전을 사용)
  - 다음에서 다운로드 가능: http://www.rstudio.com/
- R의 통합 개발 환경 소프트웨어는 이 외에도 여러 가지가 있음
  - Tinn-R, Emacs, Microsoft Visual Studio 등도 R을 위한 통합 개발 환경을 제공함

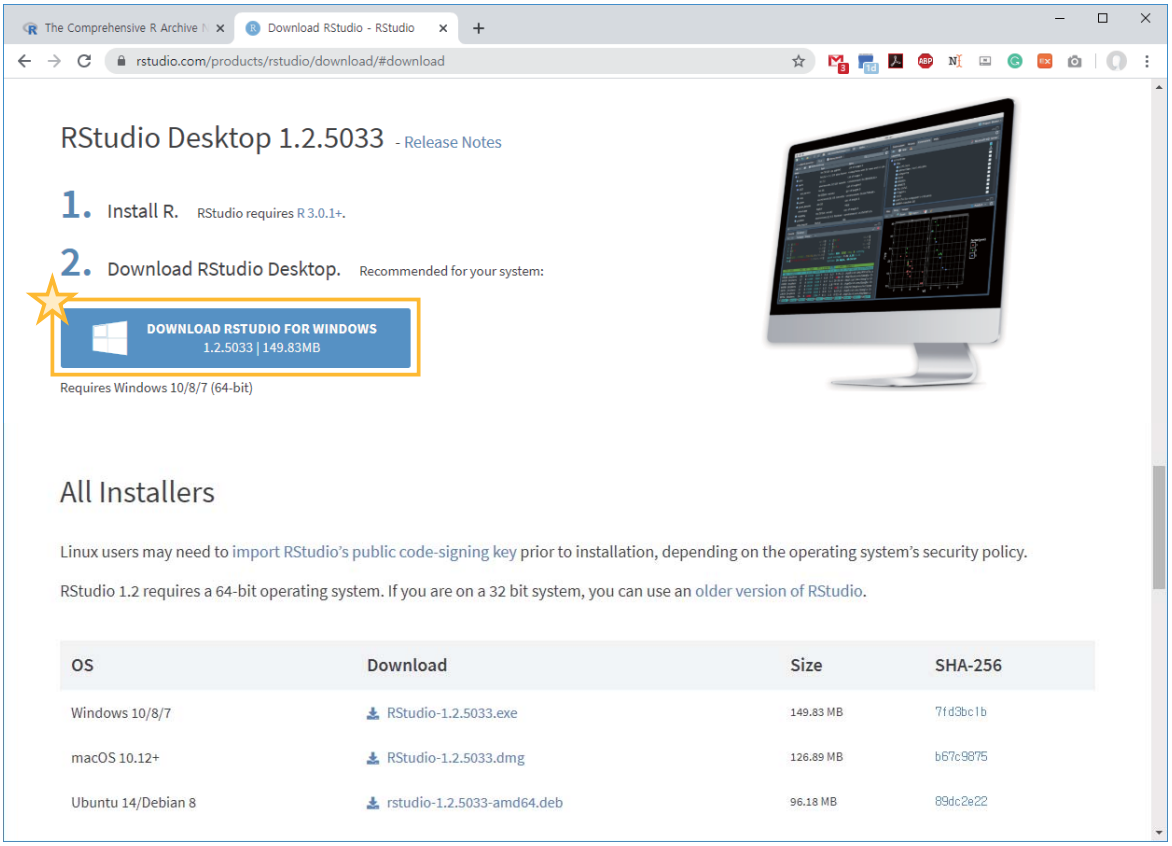

The Comprehensive R Archive
RStudio | Open source & profes
rstudio.com
https://rstudio.com/
RStudio
DOWNLOAD
SUPPORT
COMMUNITY
Products
Resources
Pricing
About
Blogs
RStudio
Certified
B
Corporation
RSTUDIO IS NOW A PUBLIC BENEFIT CORPORATION
Introducing RStudio Team
RStudio's recommended professional data science solution for every team. Learn More

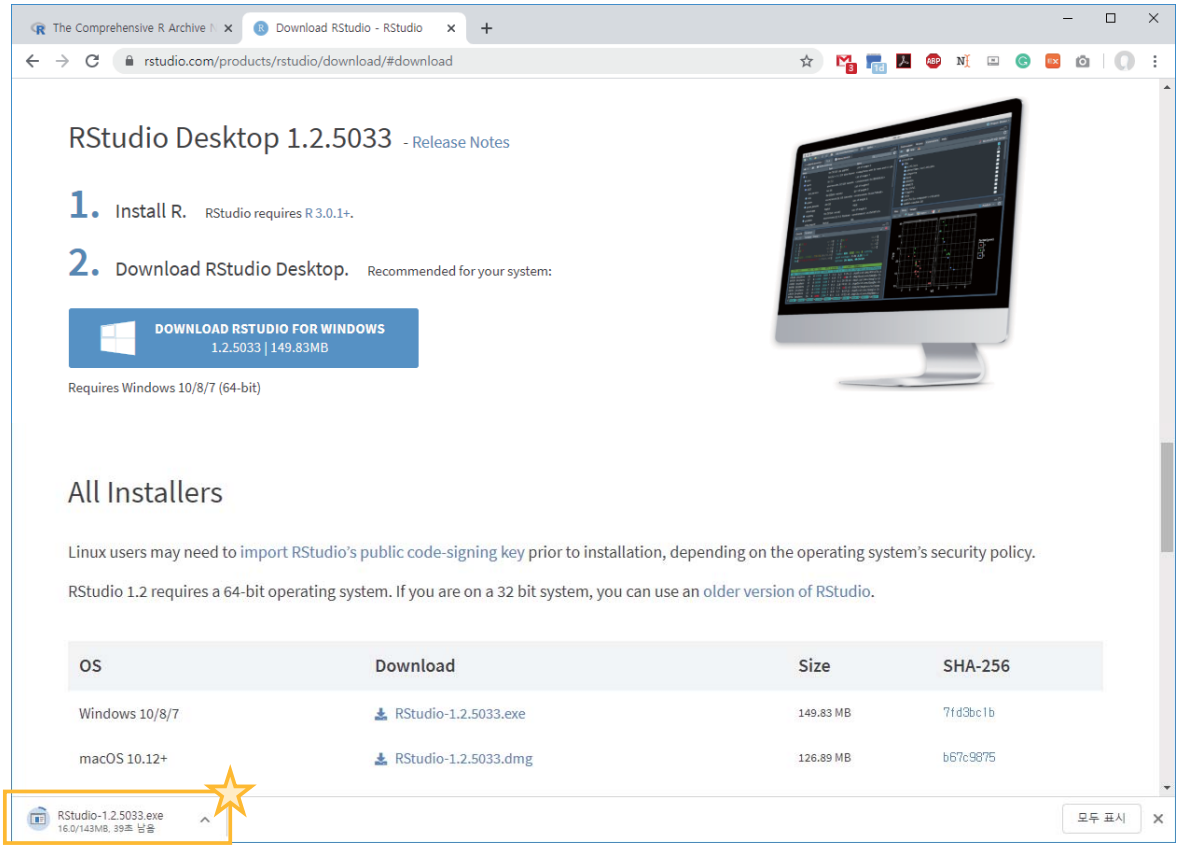

The Comprehensive R Archive
RStudio | Open source & profes
rstudio.com
RStudio
DOWNLOAD
SUPPORT
COMMUNITY
Products
Resources
Pricing
About
Blogs
OPEN SOURCE
Get started with R
RStudio
The premier IDE for R
RStudio Server
RStudio anywhere using a web browser
Shiny Server
Put Shiny applications online
R Packages
Shiny, R Markdown, Tidyverse and more
HOSTED SERVICES
Be our guest, be our guest
RStudio Cloud
Do, share, teach and learn data science
shinyapps.io
Let us host your Shiny applications
PROFESSIONAL
Enterprise-ready
RStudio Team
The premier software bundle for data science teams
RStudio Server Pro
RStudio for the Enterprise
RStudio Connect
Connect data scientists with decision makers
RStudio Package Manager
Control and distribute packages
Introducing RStudio Team
https://rstudio.com/products/rstudio/
RStudio's recommended professional data science solution for every team. Learn More

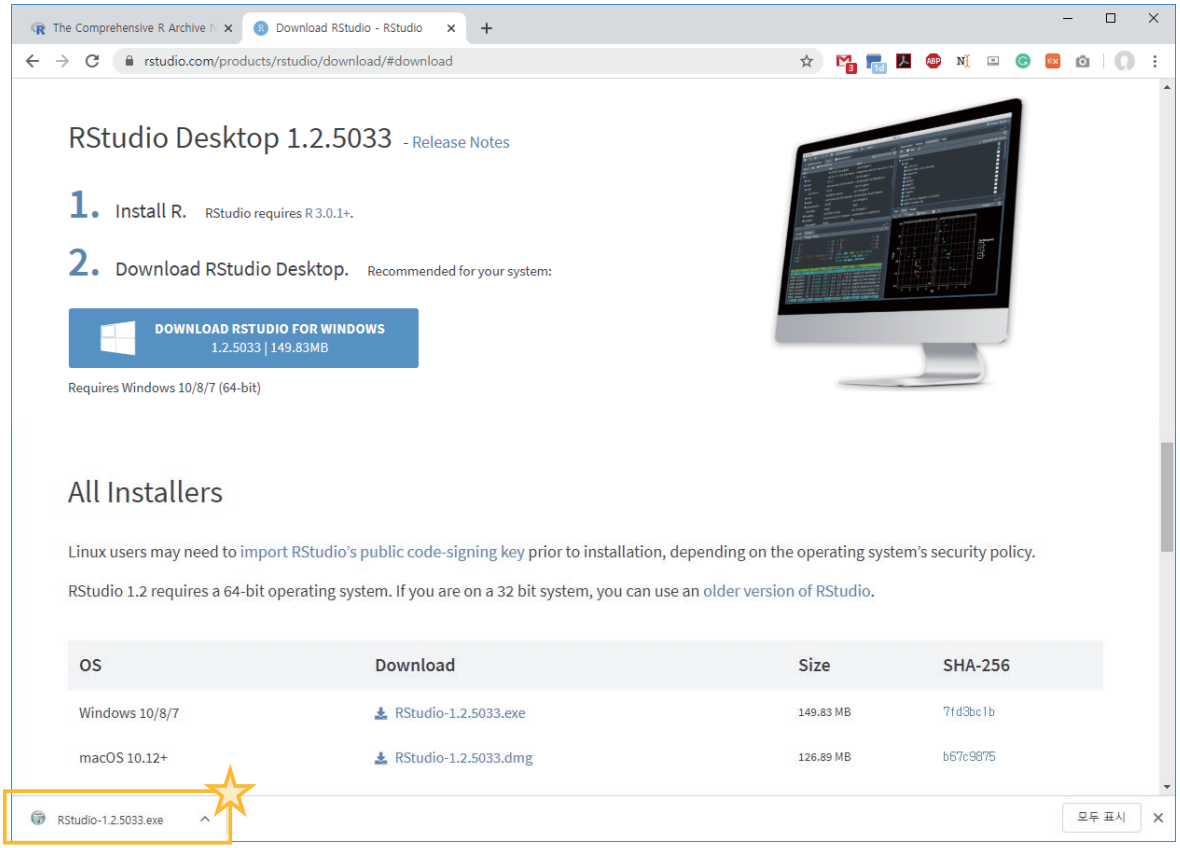

The Comprehensive R Archive
RStudio - RStudio
rstudio.com/products/rstudio/
RStudio
DOWNLOAD
SUPPORT
COMMUNITY
Products
Resources
Pricing
About
Blogs
RStudio
Take control of your R code
RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management. Click here to see more RStudio features.
RStudio is available in open source and commercial editions and runs on the desktop (Windows, Mac, and Linux) or in a browser connected to RStudio Server or RStudio Server Pro (Debian/Ubuntu, Red Hat/CentOS, and SUSE Linux).
There are two versions of RStudio:
RStudio Desktop
Run RStudio on your desktop
RStudio Server
Centralize access and computation

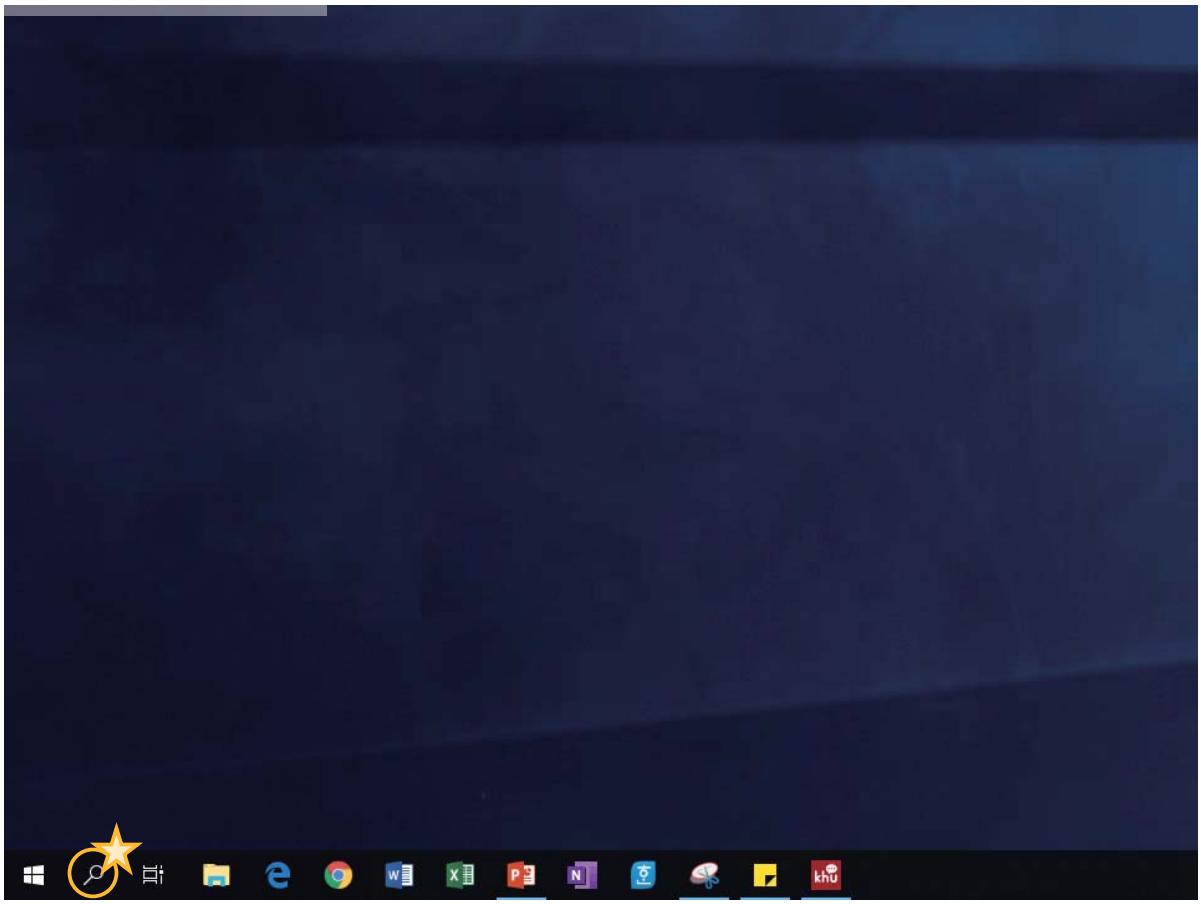

The Comprehensive R Archive
RStudio - RStudio
rstudio.com/products/rstudio/
RStudio Desktop
Open Source Edition
RStudio Desktop Pro
Overview
Access RStudio locally
Syntax highlighting, code completion, and smart indentation
Execute R code directly from the source editor
Quickly jump to function definitions
Easily manage multiple working directories using projects
Integrated R help and documentation
Interactive debugger to diagnose and fix errors quickly
Extensive package development tools
All of the features of open source; plus:
A commercial license for organizations not able to use AGPL software
Access to priority support
RStudio Professional Drivers
Support
Community forums only
Priority Email Support
8 hour response during business hours (ET)
License
AGPL v3
RStudio License Agreement
Pricing
Free
$995/year
DOWNLOAD RSTUDIO DESKTOP
DOWNLOAD FREE RSTUDIO DESKTOP PRO TRIAL
Purchase | Contact Sales

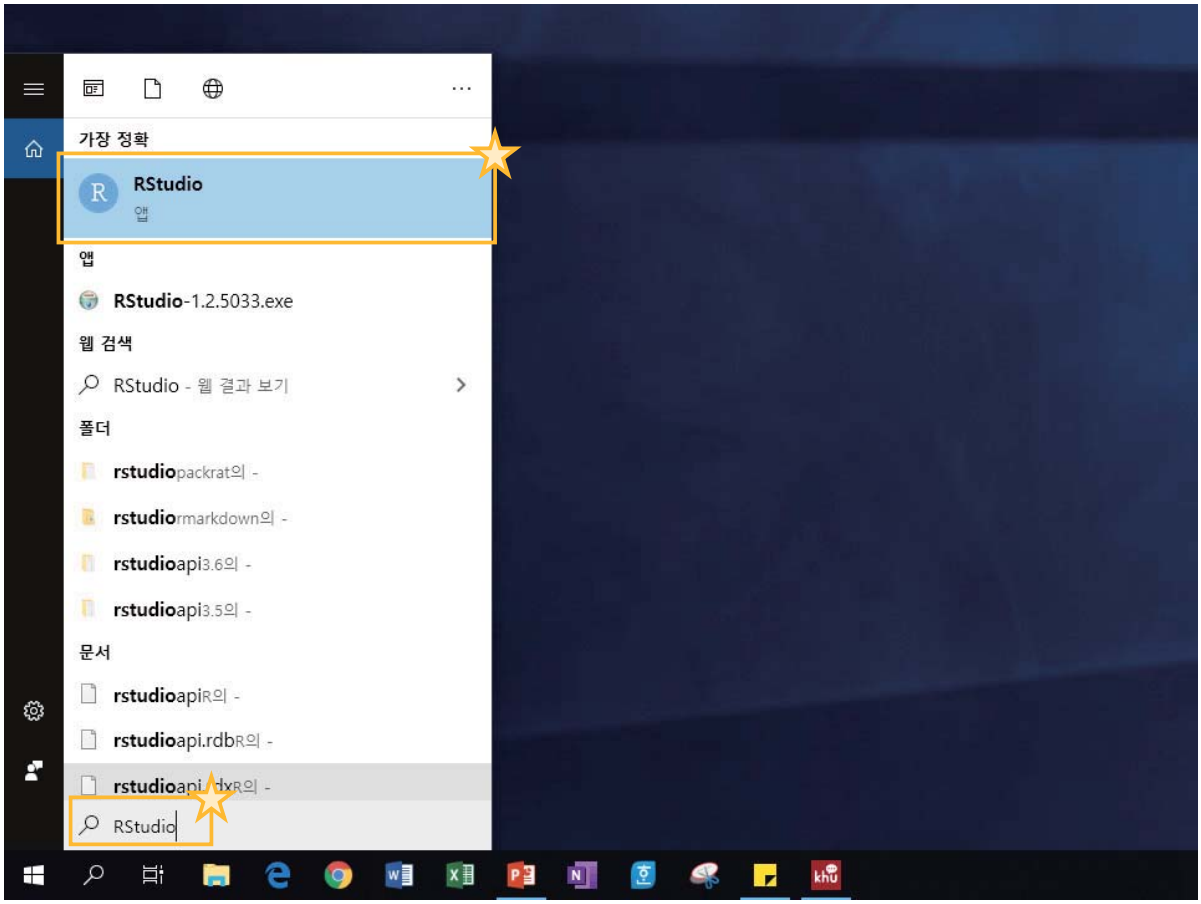

The Comprehensive R Archive
Download RStudio - RStudio
rstudio.com/products/rstudio/download/
Choose Your Version
RStudio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct code execution, and a variety of robust tools for plotting, viewing history, debugging and managing your workspace.
LEARN MORE ABOUT RSTUDIO FEATURES
Studio Team
RStudio's new solution for every professional data science team. RStudio Team includes RStudio Server Pro, RStudio Connect and RStudio Package Manager.
LEARN MORE
RStudio Desktop
Open Source License
Free
DOWNLOAD
Learn more
RStudio Desktop
Commercial License
$995 /year
BUY
Learn more
RStudio Server
Open Source License
Free
DOWNLOAD
Learn more
RStudio Server Pro
Commercial License
$4,975 /year
(5 Named Users)
BUY
Evaluation | Learn more
Integrated Tools for R
Priority Support
Access via Web Browser


The Comprehensive R Archive
Download RStudio - RStudio
rstudio.com/products/rstudio/download/#download
RStudio Desktop 1.2.5033 - Release Notes
1. Install R. RStudio requires R 3.0.1+.
2. Download RStudio Desktop. Recommended for your system:
DOWNLOAD RSTUDIO FOR WINDOWS
1.2.5033 | 149.83MB
Requires Windows 10/8/7 (64-bit)
All Installers
Linux users may need to import RStudio's public code-signing key prior to installation, depending on the operating system's security policy.
RStudio 1.2 requires a 64-bit operating system. If you are on a 32 bit system, you can use an older version of RStudio.
OS
Download
Size
SHA-256
Windows 10/8/7
RStudio-1.2.5033.exe
149.83 MB
7fd3bc1b
macOS 10.12+
RStudio-1.2.5033.dmg
126.89 MB
b67c9875
Ubuntu 14/Debian 8
rstudio-1.2.5033-amd64.deb
96.18 MB
89dc2e22

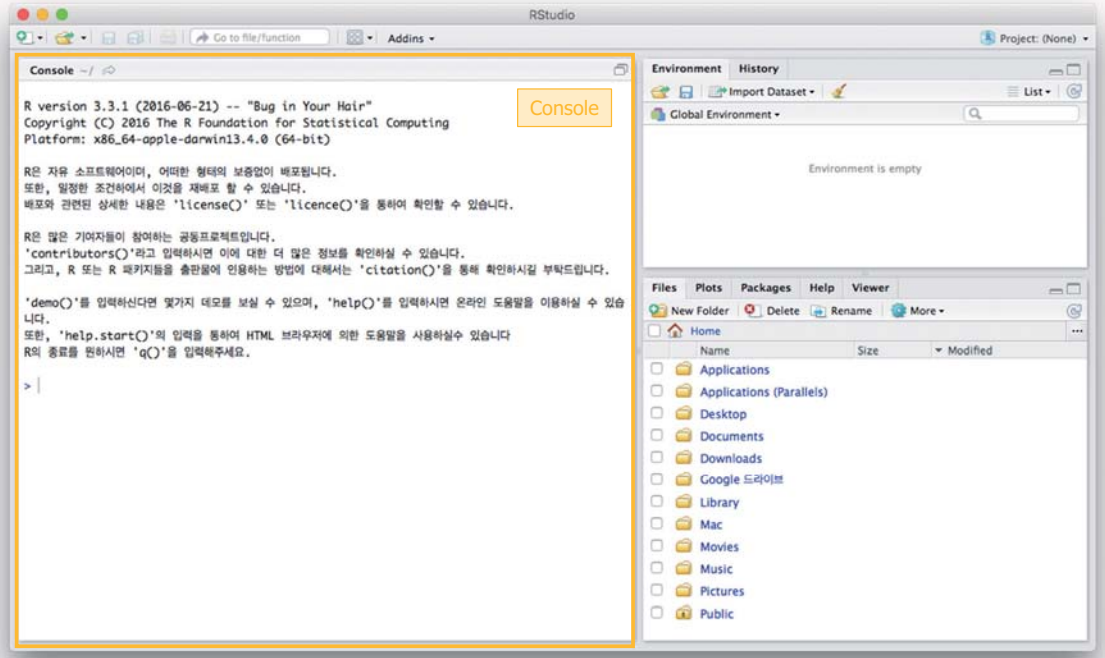

The Comprehensive R Archive
Download RStudio - RStudio
rstudio.com/products/rstudio/download/#download
RStudio Desktop 1.2.5033 - Release Notes
1. Install R. RStudio requires R 3.0.1+.
2. Download RStudio Desktop. Recommended for your system:
DOWNLOAD RSTUDIO FOR WINDOWS
1.2.5033 | 149.83MB
Requires Windows 10/8/7 (64-bit)
All Installers
Linux users may need to import RStudio's public code-signing key prior to installation, depending on the operating system's security policy.
RStudio 1.2 requires a 64-bit operating system. If you are on a 32 bit system, you can use an older version of RStudio.
OS
Download
Size
SHA-256
Windows 10/8/7
RStudio-1.2.5033.exe
149.83 MB
7fd3bc1b
macOS 10.12+
RStudio-1.2.5033.dmg
126.89 MB
b67c9875
RStudio-1.2.5033.exe
모두 표시


The Comprehensive R Archive
Download RStudio - RStudio
rstudio.com/products/rstudio/download/#download
RStudio Desktop 1.2.5033 - Release Notes
1. Install R. RStudio requires R 3.0.1+.
2. Download RStudio Desktop. Recommended for your system:
DOWNLOAD RSTUDIO FOR WINDOWS
1.2.5033 | 149.83MB
Requires Windows 10/8/7 (64-bit)
All Installers
Linux users may need to import RStudio's public code-signing key prior to installation, depending on the operating system's security policy.
RStudio 1.2 requires a 64-bit operating system. If you are on a 32 bit system, you can use an older version of RStudio.
OS
Download
Size
SHA-256
Windows 10/8/7
RStudio-1.2.5033.exe
149.83 MB
7fd3bc1b
macOS 10.12+
RStudio-1.2.5033.dmg
126.89 MB
b67c9875
RStudio-1.2.5033.exe
모두 표시

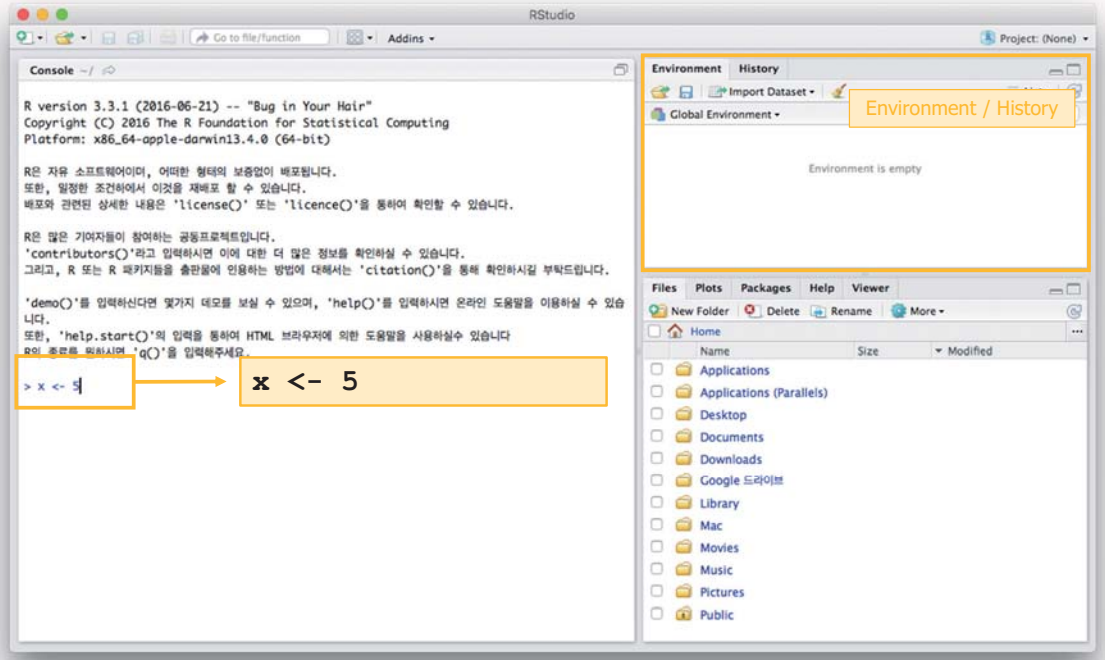



가장 정확
RStudio
앱
앱
RStudio-1.2.5033.exe
웹 검색
RStudio - 웹 결과 보기
폴더
rstudiopackrat의 -
rstudiormarkdown의 -
rstudioapi3.6의 -
rstudioapi3.5의 -
문서
rstudioapiR의 -
rstudioapi.rdbR의 -
RStudio

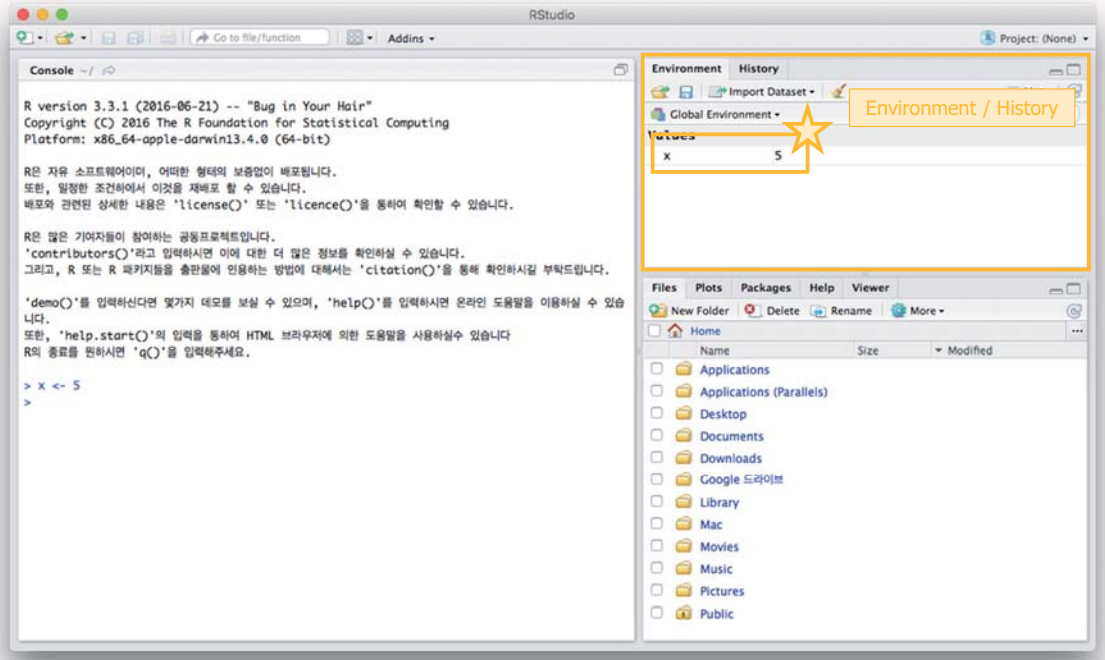
10. R과 RStudio



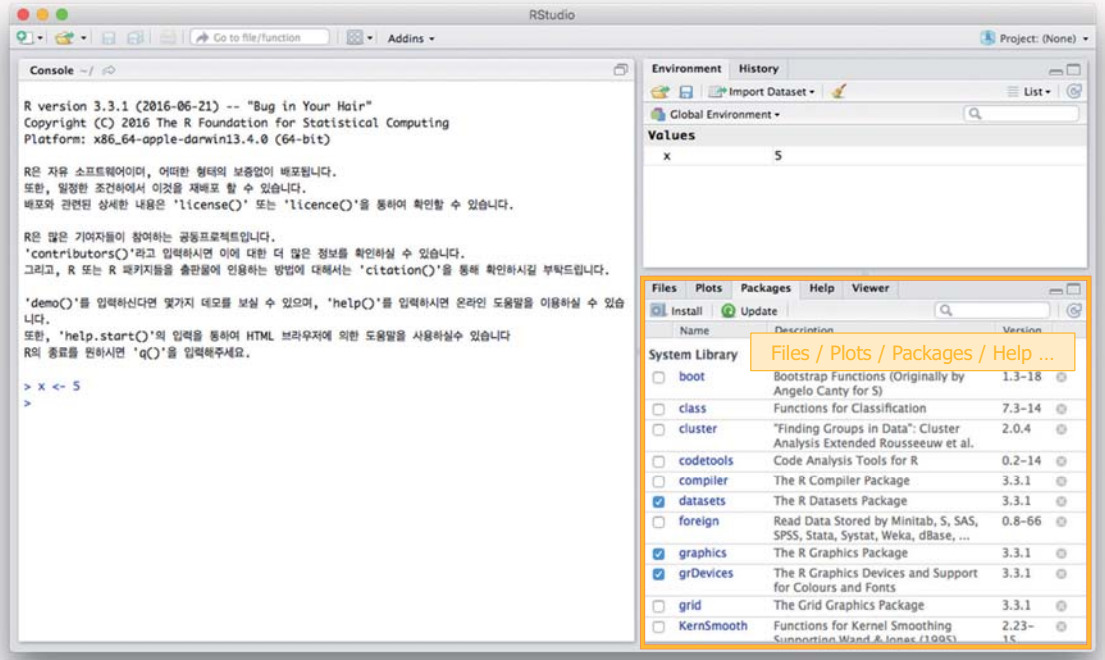
10. R과 RStudio



10. R과 RStudio



10. R과 RStudio



참고문헌

- The R foundation, "What is R?," *The R Project for Statistical Computing*.
  - http://www.r-project.org/about.html (accessed September 4, 2016).
- Ihaka, R. 1998. *R: Past and Future History*.
  - http://cran.r-project.org/doc/html/interface98-paper/paper.html
- Wikipedia contributors, "R (programming language)," *Wikipedia, The Free Encyclopedia.*
  - http://en.wikipedia.org/w/index.php?title=R_(programming_language)&oldid=737192067 (accessed September 4, 2016).

The 7th KOSTAT-UNFPA
Summer Seminar on Population

11

데이터 입력과 객체, 클래스

11. 데이터 입력과 객체, 클래스

1) 데이터 입력 및 수정
2) 형(形) 변환
3) 객체와 클래스의 개념

예제 데이터

- 아래의 표는 기초통계 과목에서 무작위로 선택된 일곱 명 학생의 중간고사와 기말고사 점수를 보여준다.

학생	중간고사	기말고사
A	30	36
B	26	13
C	24	32
D	24	22
E	20	19
F	30	35
G	22	30

데이터 입력하기

- R에서 모든 데이터는 객체(object) 형태로 저장됨
  - 예를 들어 아래와 같이 "5" 라는 숫자를 "a" 라는 저장공간에 저장한 다면 여기서 "a" 가 바로 하나의 객체가 되는 것
  - 이렇게 저장된 객체를 벡터(vector) 객체라 함

```
> a <- 5
```

- 저장된 내용은 객체의 이름을 사용해서 다시 확인이 가능함
  - 객체의 이름이 기억나지 않을 때는 `ls()` 명령어(함수)를 사용

```
> a
[1] 5
> ls()
[1] "a"
```


데이터 입력하기

- 여러 개의 데이터 값(예: 모든 학생들의 중간고사 성적)을 R에서 한 번에 입력할 수는 없을까?

학생	중간고사	기말고사
A	30	36
B	26	13
C	24	32
D	24	22
E	20	19
F	30	35
G	22	30

한 번에?

데이터 입력하기

- 하나의 벡터(vector) 객체에 여러 개의 값을 저장하기 위해 c() 함수를 사용할 수 있음
  - c(값1, 값2, 값3 …)과 같은 방법으로 사용

```
> midterm <- c(30, 26, 24, 24, 20, 30, 22)
> final = c(36, 13, 32, 22, 19, 35, 30)
> midterm
[1] 30 26 24 24 20 30 22
> final
[1] 36 13 32 22 19 35 30
> midterm + final
[1] 66 39 56 46 39 65 52
```


11. 데이터 입력

데이터 입력하기

- 하나의 벡터(vector) 객체에 여러 개의 값을 저장하기 위해 c() 함수를 사용할 수 있음
  - c(값1, 값2, 값3 …)과 같은 방법으로 사용

```
> midterm <- c(30, 26, 24, 24, 20, 30, 22)
> final = c(36, 13, 32, 22, 19, 35, 30)
> midterm
[1] 30 26 24 24 20 30 22
> final
[1] 36 13 32 22 19 35 30
> midterm + final
[1] 66 39 56 46 39 65 52
```

데이터 입력하기

- 벡터(vector) 객체에 들어있는 데이터 값의 수(원소의 수)를 확인하고자 할 때는 `length()` 함수를 사용함
  - `length(객체명)` 형식으로 사용

```
> x <- c(1, 2, 3, 4, 5)
> length(x)
[1] 5
> y <- c(10, 9, 8)
> length(y)
[1] 3
```


데이터 수정하기

- 입력된 데이터에서 일부를 수정하고자 하는 경우:
  - 중간고사 점수에서 다섯 번째 학생의 점수를 수정

```
> midterm[5] <- 30
> midterm
[1] 30 26 24 24 30 30 22
```

  - 기말고사 점수에서 두 번째와 여섯 번째 학생의 점수를 수정

```
> final[c(2, 6)] <- c(23, 25)
> final
[1] 36 23 32 22 19 25 30
```

연산자 : 의 사용

- 1씩 증가 또는 감소하는 간단한 수열은 아래와 같이 : 연산자를 사용하여 생성할 수도 있음

```
> 1:10
[1] 1 2 3 4 5 6 7 8 9 10
> 10:1
[1] 10 9 8 7 6 5 4 3 2 1
```

- 연산자의 앞뒤에 위치하는 값은 반드시 정수일 필요는 없음

```
> 1.5:7.5
[1] 1.5 2.5 3.5 4.5 5.5 6.5 7.5
> 1.5:7
[1] 1.5 2.5 3.5 4.5 5.5 6.5
```


수열의 생성

- 만약 1이 아닌 다른 숫자로 증감하는 수열을 만들고 싶다면?
- R 명령어(함수) 중 하나인 `seq()`를 사용하면 다양한 종류의 수열 객체를 만들 수 있음
  - `seq()` 함수에는 `from`, `to`, `by` , `length.out`과 같은 다양한 인자가 조합되어 사용됨

```
> x.seq <- seq(0, 1, by = .1)
> x.seq
[1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
> y.seq <- seq(0, 1, length.out = 11)
> y.seq
[1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
```

수열의 생성

- `seq()`와 같이 함수에 여러 개의 인자(arguments)가 들어갈 때:
  - 함수에서 정의된 순서대로 인자를 입력하는 경우에는 인자 이름을 생략할 수 있음
  - 만약 순서가 바뀌거나 중간에 들어가는 인자를 생략하는 경우에는 인자의 이름을 정확히 지정해주어야 함

```
> seq(0, 1, .1)
[1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
> seq(0, 1, 11)
[1] 0
```

- 함수에서 정의된 인자의 순서는 도움말을 통해 확인 가능함:

```
> help(seq)
```


반복되는 값의 입력

- 만약 동일한 숫자가 반복되는 데이터가 필요하다면 `rep()` 함수가 유용하게 사용됨
  - `rep()` 함수에 들어가는 첫 번째 값은 반복될 숫자를 의미하며, 쉼표로 구분되는 두 번째 값은 반복 횟수를 말함
  - 각각에 하나 이상의 숫자를 사용할 수도 있음

```
> rep(1:4, 3)
[1] 1 2 3 4 1 2 3 4 1 2 3 4
> rep(1:4, c(2, 3, 2, 3))
[1] 1 1 2 2 2 3 3 4 4 4
> rep(1:4, each = 3)
[1] 1 1 1 2 2 2 3 3 3 4 4 4
```

반복되는 값의 입력

- 벡터 객체는 숫자가 아닌, 문자와 같은 다른 유형의 값으로 만들어 질 수도 있음
  - 그러나 하나의 객체에는 모두 같은 유형의 값만 들어가야 함(숫자 벡터, 문자 벡터 등)
- R에서 사용가능한 대표적 데이터 유형

유형 이름	종류	예
logical	참/거짓	TRUE, FALSE
numeric	숫자(실수)	0, 1, -2, 0.4, -3.9, 65535
complex	복소수	3+5i, 10+0i
character	문자	봄, 가을, 겨울

숫자가 아닌 데이터의 입력

- 문자열은 반드시 따옴표 안에 넣어서 저장
  - 따옴표가 없는 문자열은 R에서 객체의 이름으로 인식하게 됨

```
> b <- "Hello"
> e <- Hello
Error: object 'Hello' not found
```

- 참/거짓과 값은 논리값은 문자열과 달리 따옴표 없이 입력하며, 반드시 대문자로 적어야 함

```
> lv <- c(TRUE, FALSE, TRUE, TRUE)
> wv <- c(true, false, true, true)
Error: object 'true' not found
```

형(形) 변환

- 서로 다른 유형(type)의 값을 하나의 벡터에 입력하면, 형(形) 변환이 이루어짐

```
> num <- c(1, 2, 3, 4, 5)
> num
[1] 1 2 3 4 5
> mode(num)
[1] "numeric"
> num[4] <- "Four"
> num
[1] "1"    "2"    "3"    "Four" "5"
> mode(num)
[1] "character"
```

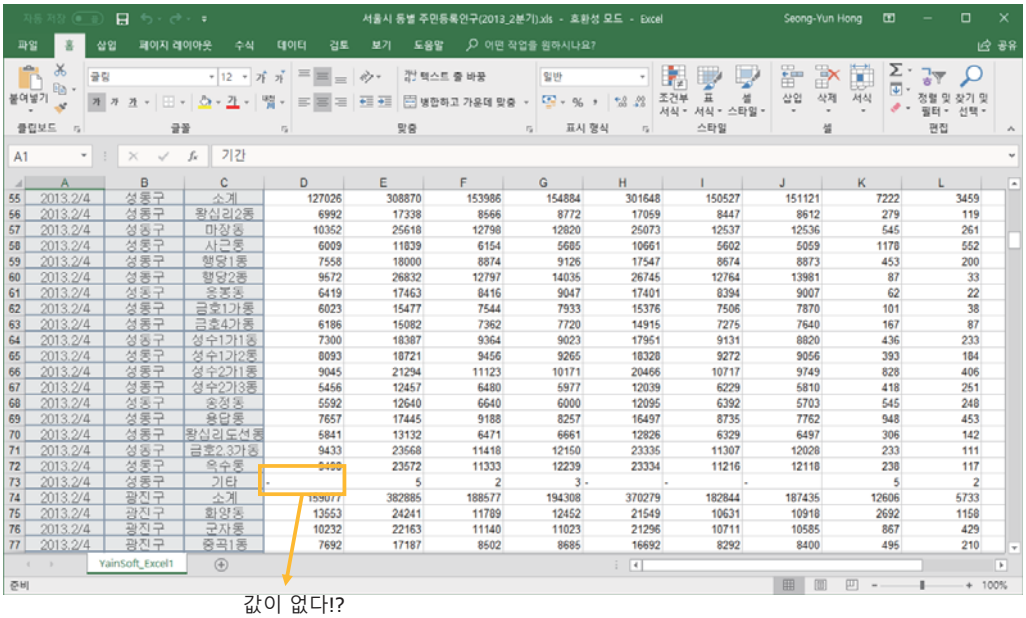

형(形) 변환

- 앞의 예와 같이 숫자 유형의 벡터에 문자를 넣으면 값 전체가 문자열로 변환(coerce)됨
  - `mode()` 함수를 사용하면 객체의 유형을 확인할 수 있음
- 형(形) 변환은 다음과 같은 순서로 이루어짐

Logical ➔ Numeric ➔ Complex ➔ Character

  - 논리값 `TRUE`는 숫자 1로, `FALSE`는 0으로 변환될 수 있음
  - 실수(實數) 5는 복소수(複素數) 5+0i 형태로 입력될 수 있음
- 벡터 유형에 관계 없이 `NA`, `Inf`, `-Inf`, `NaN`와 같이 특별히 약속된 값들은 형 변환을 일으키지 않음

결측값의 입력



	A	B	C	D	E	F	G	H	I	J	K	L
55	2013.2/4	성동구	소계	127026	308870	153986	154884	301648	150527	151121	7222	3459
56	2013.2/4	성동구	왕십리2동	6992	17338	8566	8772	17059	8447	8612	279	119
57	2013.2/4	성동구	마장동	10352	25618	12798	12820	25073	12537	12536	545	261
58	2013.2/4	성동구	사근동	6009	11839	6154	5685	10661	5602	5059	1178	552
59	2013.2/4	성동구	행당1동	7558	18000	8874	9126	17547	8674	8873	453	200
60	2013.2/4	성동구	행당2동	9572	26832	12797	14035	26745	12764	13981	87	33
61	2013.2/4	성동구	응봉동	6419	17463	8416	9047	17401	8394	9007	62	22
62	2013.2/4	성동구	금호1가동	6023	15477	7544	7933	15376	7506	7870	101	38
63	2013.2/4	성동구	금호4가동	6186	15082	7362	7720	14915	7275	7640	167	87
64	2013.2/4	성동구	성수1가1동	7300	18387	9364	9023	17951	9131	8820	436	233
65	2013.2/4	성동구	성수1가2동	8093	18721	9456	9265	18328	9272	9056	393	184
66	2013.2/4	성동구	성수2가1동	9045	21294	11123	10171	20466	10717	9749	828	406
67	2013.2/4	성동구	성수2가3동	5456	12457	6480	5977	12039	6229	5810	418	251
68	2013.2/4	성동구	송정동	5592	12640	6640	6000	12095	6392	5703	545	248
69	2013.2/4	성동구	용답동	7657	17445	9188	8257	16497	8735	7762	948	453
70	2013.2/4	성동구	왕십리도선동	5841	13132	6471	6661	12826	6329	6497	306	142
71	2013.2/4	성동구	금호2.3가동	9433	23568	11418	12150	23335	11307	12028	233	111
72	2013.2/4	성동구	옥수동	2400	23572	11333	12239	23334	11216	12118	238	117
73	2013.2/4	성동구	기타	-	5	2	3	-	-	-	5	2
74	2013.2/4	광진구	소계	159077	382885	188577	194308	370279	182844	187435	12606	5733
75	2013.2/4	광진구	화양동	13553	24241	11789	12452	21549	10631	10918	2692	1158
76	2013.2/4	광진구	군자동	10232	22163	11140	11023	21296	10711	10585	867	429
77	2013.2/4	광진구	중곡1동	7692	17187	8502	8685	16692	8292	8400	495	210

값이 없다!?

결측값의 입력

- R에서 NA는 Not Available의 약자로 결측값을 의미하는 용도로 많이 사용됨

```
> x <- c(34, 25, 27, 43, 29, 30)
> x
[1] 34 25 27 43 29 30
> mode(x)
[1] "numeric"
> x[2] <- NA
> x
[1] 34 NA 27 43 29 30
> mode(x)
[1] "numeric"
```

객체와 클래스

- R에서 존재하는 모든 것들은 객체로 볼 수 있음!

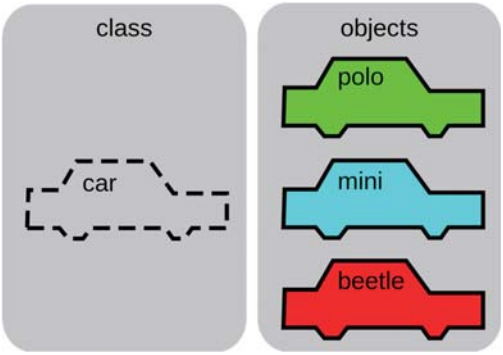
John Chambers
The creator of the S programming language

To understand computations in R, two slogans are helpful:
1) Everything that exists is an object.
2) Everything that happens is a function call.

Read this for more detailed discussion:
https://stackoverflow.com/questions/34376318/whats-the-real-meaning-about-everything-that-exists-is-an-object-in-r

객체와 클래스

- 객체와 클래스
  - 모든 객체에는 클래스(class)라고 하는 속성이 있음
  - 클래스에 따라 데이터가 저장되는 구조가 달라짐
  - 오늘 수업에서는 벡터 클래스의 객체만을 다루었으나, 앞으로 수업에서 행렬, 리스트, 데이터프레임 등 다양한 클래스를 다루게 됨
  - R에는 공간데이터를 위한 클래스도 있음



https://upload.Wikimedia.org/wikipedia/commons/thumb/6/62/CPT-OOP-objects_and_classes.svg/1280px-CPT-OOP-objects_and_classes.svg.png

더 들어가기에 앞서 ...

- 작업내용의 저장
  - R 코드(또는 스크립트)의 저장 → *.R
  - 객체의 저장 → *.RData
- 작업내용을 불러오기
  - 저장된 코드를 불러오기 → `source()`
  - 저장된 객체를 불러오기 → `load()`
- 작업폴더 설정
  - 작업폴더의 설정 → `setwd()`
  - 작업폴더의 확인 → `getwd()`

더 들어가기에 앞서 ...

- 도움말 기능
  - `help()` 또는 `?` 사용
- R 코드 작성시 Google's R Style Guide 내용을 준수하고, 충분한 설명(주석)을 포함하는 것이 중요
  - R에서 주석은 `#` 기호 사용

The 7th KOSTAT-UNFPA
Summer Seminar on Population

12

파일 데이터 열기

12. 파일 데이터 열기

1) CSV 파일 열기
2) 데이터 탐색과 추출
3) 요인과 문자열
4) 데이터 저장

정형 데이터의 구조

- 일반적으로 통계 분석에 쓰이는 데이터는 대부분 아래와 같은 정형적 틀을 가지고 있음(정형 데이터)
  - 각 열(column)은 변수를 나타내는 경우가 많으며, 이 경우 각 행(row)은 개개의 관찰 대상을 의미함

변수

관찰 대상

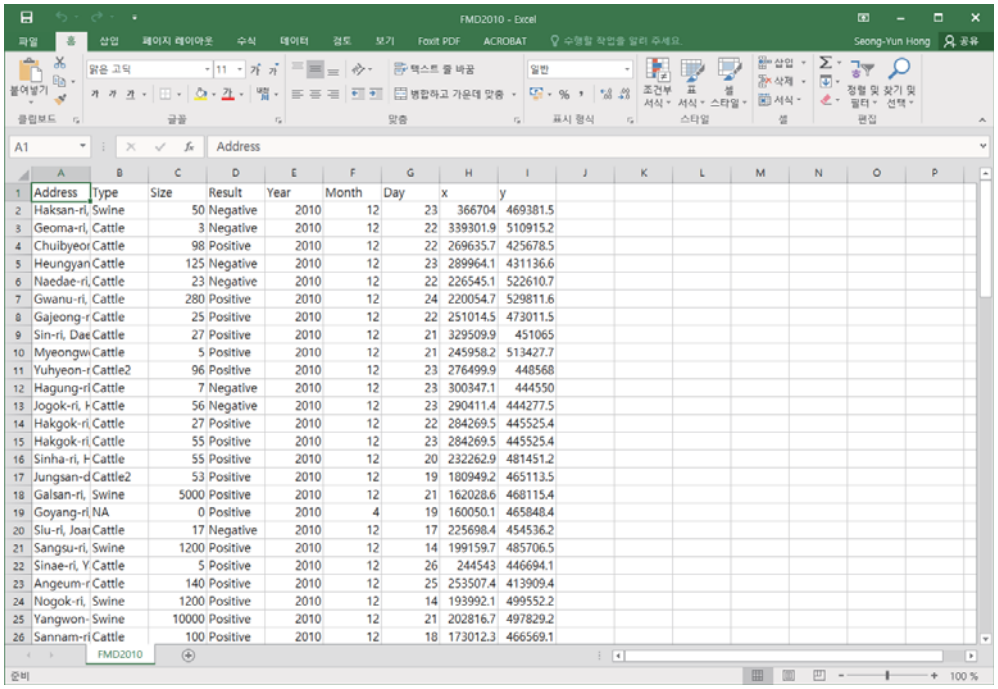
ID	Year	GPA	…		
20120…	4	3.5			
…					

- 이러한 데이터는 일반적으로 엑셀 파일과 같은 형식으로 저장되어 있음

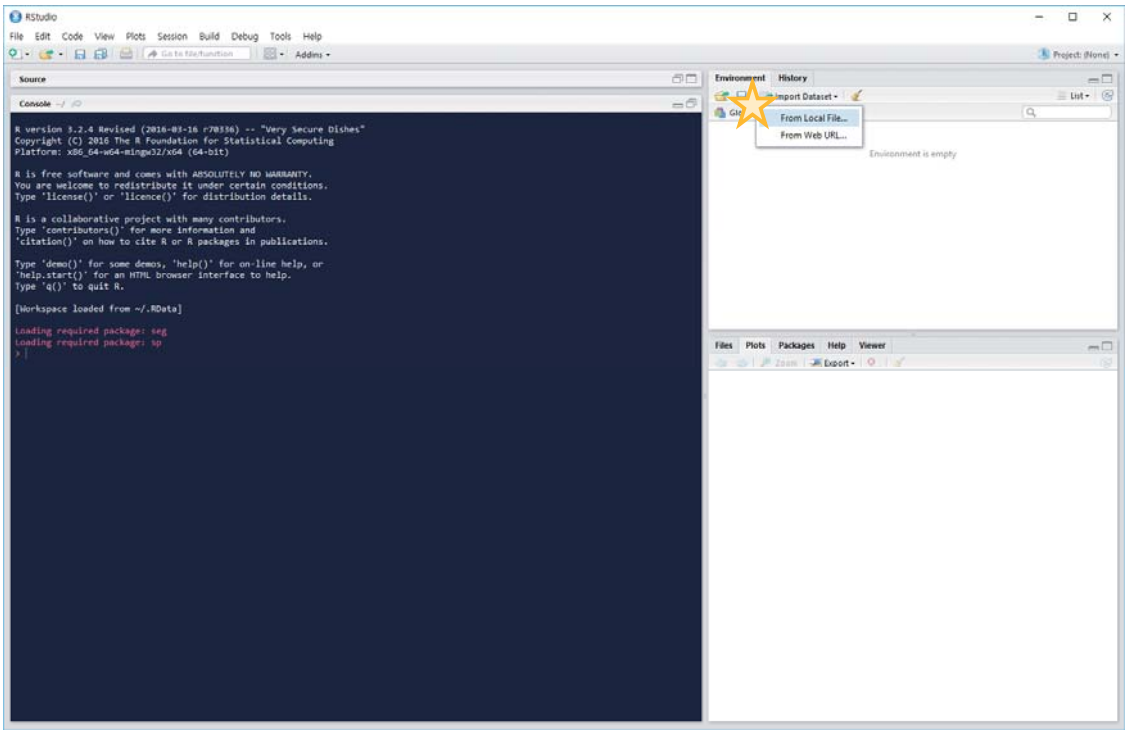
구제역 발생 데이터

- 국내 구제역 발생지역 데이터
  - 2010년 4월 8일부터 12월 26일까지의 구제역 발생 내역
  - 당시 농림수산식품부에 공개된 데이터를 가공, 편집한 자료로 CSV 형식으로 저장되어 있음
  - CSV(comma-separated values) 파일은 각 열(column)을 쉼표로 구분한 텍스트 파일로 대부분의 프로그램에서 읽고 쓸 수 있음
- R에서 CSV 파일에 기록된 데이터를 사용하기 위해서는 해당 파일을 R로 불러와 객체 형태로 저장하는 것이 필요함
  - `read.table()` 또는 `read.csv()`과 같은 함수를 사용하여 데이터를 가져올 수 있음

```
fmd <- read.csv(file.choose(), header = TRUE)
```

Address	Type	Size	Result	Year	Month	Day	x	y
Haksan-ri,	Swine	50	Negative	2010	12	23	366704	469381.5
Geoma-ri,	Cattle	3	Negative	2010	12	22	339301.9	510915.2
Chuibyeor	Cattle	98	Positive	2010	12	22	269635.7	425678.5
Heungyan	Cattle	125	Negative	2010	12	23	289964.1	431136.6
Naedae-ri,	Cattle	23	Negative	2010	12	22	226545.1	522610.7
Gwanu-ri,	Cattle	280	Positive	2010	12	24	220054.7	529811.6
Gajeong-r	Cattle	25	Positive	2010	12	22	251014.5	473011.5
Sin-ri, Dae	Cattle	27	Positive	2010	12	21	329509.9	451065
Myeongw	Cattle	5	Positive	2010	12	21	245958.2	513427.7
Yuhyeon-r	Cattle2	96	Positive	2010	12	23	276499.9	448568
Hagung-ri	Cattle	7	Negative	2010	12	23	300347.1	444550
Jogok-ri, H	Cattle	56	Negative	2010	12	23	290411.4	444277.5
Hakgok-ri	Cattle	27	Positive	2010	12	22	284269.5	445525.4
Hakgok-ri	Cattle	55	Positive	2010	12	23	284269.5	445525.4
Sinha-ri, H	Cattle	55	Positive	2010	12	20	232262.9	481451.2
Jungsan-d	Cattle2	53	Positive	2010	12	19	180949.2	465113.5
Galsan-ri,	Swine	5000	Positive	2010	12	21	162028.6	468115.4
Goyang-ri	NA	0	Positive	2010	4	19	160050.1	465848.4
Siu-ri, Joa	Cattle	17	Negative	2010	12	17	225698.4	454536.2
Sangsu-ri,	Swine	1200	Positive	2010	12	14	199159.7	485706.5
Sinae-ri, Y	Cattle	5	Positive	2010	12	26	244543	446694.1
Angeum-r	Cattle	140	Positive	2010	12	25	253507.4	413909.4
Nogok-ri,	Swine	1200	Positive	2010	12	14	193992.1	499552.2
Yangwon-	Swine	10000	Positive	2010	12	21	202816.7	497829.2
Sannam-ri	Cattle	100	Positive	2010	12	18	173012.3	466569.1



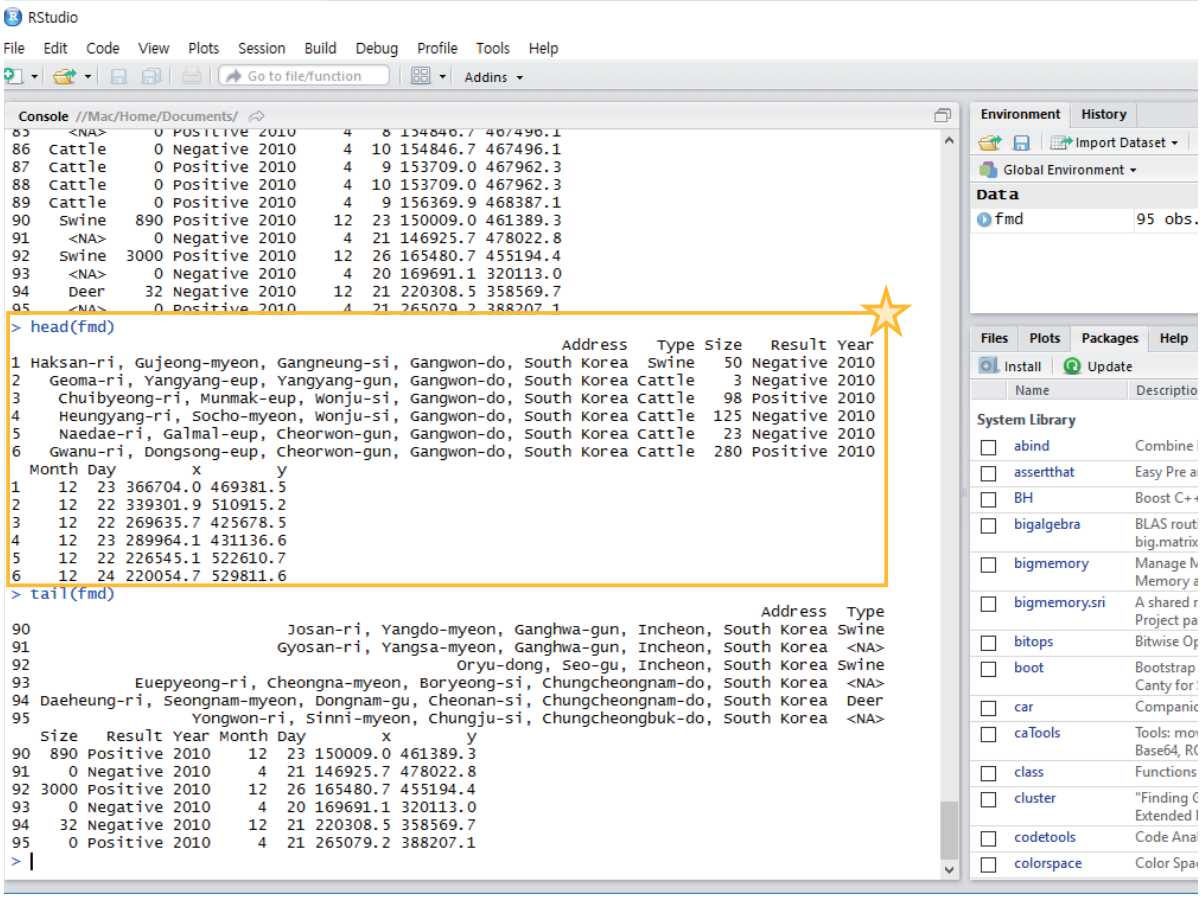
데이터 살펴보기

- R로 가져온 데이터의 내용을 확인하기 위해서는 지난 시간에 배운 바와 같이 객체의 이름을 직접 입력하면 됨:

```
> fmd
...
```

  - 결과는 다음 페이지에!

- 데이터의 크기가 큰 경우(한 화면에 나타내기 어려운 경우)에는 부분적으로 나눠서 살펴보는 것이 효과적일 수 있음
  - `head()` 또는 `tail()` 함수를 사용하는 방법
  - 특정 조건을 만족하는 데이터를 확인하기 위해 `[]` 또는 `subset()` 함수를 사용하는 방법



12. 파일 데이터 열기

특정 위치의 원소 선택

- 지난 실습 시간을 통해 벡터 객체에서 특정 위치의 원소를 선택하는 방법에 대해 살펴봤음:

```
> x <- c(10, 20, 30, 40, 50)
> x[2]
[1] 20
> x[c(1, 3, 5)]
[1] 10 30 50
> x[4:5]
[1] 40 50
> x[]
[1] 10 20 30 40 50
```

특정 위치의 원소 선택

- 사용자가 지정한 위치가 존재하지 않는 경우는?
  - 원소의 위치값(index)은 1부터 시작하며, 0을 입력하는 경우에는 아무것도 반환되지 않음(2020년 8월 기준)

```
> x[0]
numeric(0)
```

  - 벡터의 길이보다 큰 원소 번호를 사용하는 경우에는 아래 예제와 같이 NA가 반환됨

```
> x[c(3, 6)]
[1] 30 NA
> x[1:(length(x)+3)]
[1] 10 20 30 40 50 NA NA NA
```


특정 위치의 원소 선택

- 이와 유사한 방법으로 벡터 객체에서 특정 위치의 원소를 제외한 나머지 원소들만 출력할 수도 있음:
  - 위치값 앞에 빼기(–) 부호를 사용함

```
> x[-2]
[1] 10 30 40 50
> x[-c(1, 3, 5)]
[1] 20 40
> x[-(4:5)]
[1] 10 20 30
> x[-c(3, 6)]
[1] 10 20 40 50
```

특정 위치의 원소 선택

- 위치값(index)이 아닌 참/거짓의 논리값을 사용해 원소 중 일부만 선별해 출력할 수도 있음
  - 벡터 길이와 동일한 길이의 논리값을 `[]` 안에 입력하면, `TRUE`에 해당하는 원소는 출력되고, `FALSE`에 해당하는 원소는 출력되지 않음

```
> x <- c(10, 20, 30, 40)
> x[c(FALSE, TRUE, FALSE, TRUE)]
[1] 20 40
> x[c(TRUE, TRUE, TRUE, TRUE)]
[1] 10 20 30 40
> x[c(FALSE, FALSE, FALSE, FALSE)]
numeric(0)
```


재활용 규칙

- 만약 `[]` 안에 입력한 논리값 벡터의 길이가 데이터 벡터의 길이보다 짧은 경우에는 어떻게 될까?

```
> x[c(FALSE, TRUE)]
[1] 20 40
```

어떻게 이런 결과가 나오는 걸까?

재활용 규칙(Recycling Rule)

- 서로 다른 길이를 가진 두 개 이상의 벡터를 함께 사용할 때, 길이가 짧은 벡터를 자동으로 연장해 길이가 긴 벡터와 맞추기 위한 규칙
- 짧은 길이의 벡터에서 원소를 반복(재활용)함으로써 길이를 연장함

재활용 규칙

- 원소를 선택하는 경우 뿐만 아니라, R에서 사칙연산과 같은 다양한 부분에 적용되고 있음:

```
> c(1, 2, 3, 4) + c(1, 2)
[1] 2 4 4 6
```

- 위의 예는 아래와 같이 재활용 규칙이 적용된 후 연산이 이루어진 것임:

$$\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \\ 1 \\ 2 \end{bmatrix} \Rightarrow \begin{bmatrix} 2 \\ 4 \\ 4 \\ 6 \end{bmatrix}$$

Recycling rule applied!

재활용 규칙

- 많은 R 함수에서 암묵적으로(implicitly) 재활용 규칙을 적용하여 사용자 편의성을 높이고 있음:

```
> round(1:4 * pi, 4)
[1] 3.1416 6.2832 9.4248 12.5664
> round(pi, 1:4)
[1] 3.1000 3.1400 3.1420 3.1416
```

- 함수에 따라 재활용 규칙이 적용되지 않는 경우도 있음:

```
> rep(1:5, 5)
[1] 1 2 3 4 5 1 2 3 4 5 1 2 3 4 5 1 2 3 4 5 1 2 3 4 5
> rep(1:5, rep(5, 5))
[1] 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 4 4 4 4 4 5 5 5 5 5
```

조건을 만족하는 원소 선택

- 논리값을 직접 입력하는 대신, 논리값을 생성하는 조건식을 사용하여 벡터에서 원소를 선택하는 것도 가능함!

```
> x[3 < 4]
[1] 10 20 30 40
> x[3 > 4]
numeric(0)
```

- 비교 연산에도 아래와 같이 재활용 규칙이 적용됨:

```
> x[c(1, 2, 3, 4) < 3]
[1] 10 20
```


비교 연산자와 논리 연산자

- R에서 사용할 수 있는 비교 연산자와 논리 연산자는 다음과 같다:

<	작다
<=	작거나 같다
==	같다
!=	같지 않다
>	크다
>=	크거나 같다
&	논리 연산자(그리고)
|	논리 연산자(또는)
!	논리 연산자(부정)

비교 연산자와 논리 연산자

- 비교 연산자와 논리 연산자를 활용하면 복잡한 조건에 맞는 원소를 선택할 수 있음:

```
> x[x < 30]
[1] 10 20
> x[(x >= 10 & x < 40)]
[1] 10 20 30
> x[!(x >= 10 & x < 40)]
[1] 40
```

- 그렇다면 앞서 가져온 구제역 발생 데이터처럼 행과 열로 구성된 데이터에서는 어떻게 원하는 부분만 추출할 수 있을까?

2차원 데이터의 탐색

- 전체 데이터에서 특정한 열(column), 또는 행(row)의 내용을 확인하고 싶다면:

```
> fmd[,2]
 [1] Swine   Cattle  Cattle  Cattle  Cattle  ...
[12] Cattle  Cattle  Cattle  Cattle  Cattle2 ...
[23] Swine   Swine   Cattle  <NA>    Cattle2 ...
...
```

- 객체이름[행 번호, 열 번호] 의 형식으로 사용함
  - 첫 번째, 세 번째 열의 데이터를 보고 싶다면?
  - 두 번째 행부터 여섯 번째 행까지의 데이터를 보고 싶다면?
  - 네 번째 열, 다섯 번째 행의 데이터 값은?

2차원 데이터의 탐색

- 각 열의 변수가 이름을 가지고 있다면 해당 이름을 사용해서 특정 부분을 추출할 수도 있음:

```
> fmd$Type
 [1] Swine   Cattle  Cattle  Cattle  Cattle  ...
[12] Cattle  Cattle  Cattle  Cattle  Cattle2 ...
[23] Swine   Swine   Cattle  <NA>    Cattle2 ...
...
```

- 열 이름은 `names()` 함수를 통해 조회 가능

```
> names(fmd)
[1] "Address" "Type"    "Size"    "Result"  "Year"
[6] "Month"   "Day"     "x"       "y"
```


2차원 데이터의 탐색

- 전체 데이터에서 특정 조건을 만족하는 데이터를 보고 싶다면:

```
> subset(fmd, Type == "Swine")
...
```

데이터 객체 이름

조건식

  – 직접 실행해서 결과를 확인합시다!

- 만약 추출된 데이터를 단순히 화면에 보여주는 것이 아니라, 별도의 객체로 저장하고 싶다면:

```
> fmd.swine <- subset(fmd, Type == "Swine")
```

클래스의 확인

- 지난 시간에 다룬 벡터 데이터와 달리 구제역 발생 데이터는 행과 열로 구성된 2차원 데이터임
  - 구제역 발생 데이터의 클래스는 무엇일까?

```
> class(fmd.df)
[1] "data.frame"
```

  - 구제역 발생 데이터에서 각 열은 하나의 변수를 나타내며, 변수에 따라 클래스가 다를 수 있음

```
> class(fmd.df[,9])
[1] "numeric"
> class(fmd.df$Address)
[1] "factor"
```

character가 아닌 factor?

요인과 명목 변수, 서열 변수

- R에서 `factor`, 즉 요인 유형의 벡터는 범주형 변수를 저장하는 용도로 주로 사용함
  - 범주형 변수란 변수가 취할 수 있는 값이 몇 가지 경우의 수(범주)로 이루어져 있는 변수로 아래와 같은 예가 있음:
    - 눈동자 색: 검은색, 갈색, 푸른색 등
    - 거주하는 지역(도시): 서울, 인천, 부산 등
    - 성별: 남성, 여성
    - 만족도: 매우 만족, 만족, 보통, 불만족, 매우 불만족
- 요인 유형의 벡터는 파일로 저장된 데이터를 불러오는 과정에서 자동으로 생성되기도 하고, `factor()` 또는 `ordered()`와 같은 함수를 사용해 직접 만들 수도 있음

요인과 명목 변수, 서열 변수

- 앞서 사용한 `read.csv()` 함수는 기본적으로 글자가 포함된 열을 요인 유형의 벡터 객체로 불러옴
  - 글자가 들어있는 열을 문자열 유형의 벡터 객체로 불러오려면 아래와 같이 `stringsAsFactors = FALSE`로 설정해 주어야 함

```
> class(fmd$Address)
[1] "factor"
> fmd2 <- read.csv(file.choose(), header = TRUE,
+ stringsAsFactors = FALSE)
> class(fmd2$Address)
[1] "character"
```


요인의 생성

- R에 이미 존재하는 객체를 요인으로 변경하거나, 요인 객체를 직접 생성하기 위해서는 `factor()` 함수를 사용
  - 아래 예의 눈동자 색과 같이 항목(범주)들 간에 정해진 순서가 없는 경우(예: 성별, 종교 등) 사용하게 된다.

```
> eyes <- c("hazel", "brown", "black", "brown")
> eyecol <- factor(eyes)
> eyes
[1] "hazel" "brown" "black" "brown"
> eyecol
[1] hazel brown black brown
Levels: black brown hazel
```

요인의 생성

- 요인 객체에는 사전에 정의된 데이터 항목만 포함될 수 있음
  - 항목을 정의하기 위해서는 `levels` 옵션을 사용

```
> eyecol <- factor(eyes, levels = c("black", "brown",
                                    "hazel", "blue"))
> eyecol
[1] hazel brown black brown
Levels: black brown hazel blue
> eyecol2 <- factor(eyes, levels = c("black", "brown"))
> eyecol2
[1] <NA>  brown black brown
Levels: black brown
```


요인의 생성

- 별도로 항목을 정의하지 않은 경우, 입력 데이터에 있는 모든 값을 사용하여 R이 자동으로 항목을 생성하게 됨
- 벡터에 저장된 데이터 값을 나타낼 때, 화면 하단에 해당 벡터가 갖고 있는 항목이 같이 표기됨
  - 데이터 항목만을 확인하고 싶다면 `levels()` 함수를 사용

```
> eyecol
[1] hazel brown black brown
Levels: black brown hazel blue
> levels(eyecol)
[1] "black" "brown" "hazel" "blue"
```

요인의 생성

- 요인 객체에 저장된 항목들 간 순서가 있는 경우에는 `ordered()` 함수를 사용해 별도로 서열을 지정해줄 수 있음
  - 소득수준, 리커트 척도 등을 저장하는데 유용하게 사용됨
  - `ordered()` 함수를 사용할 때에는 각 항목(levels)을 반드시 순서에 맞춰서 입력해야 함

```
> pain <- ordered(c("low", "medium", "medium",
                    "high"),
                  levels = c("low", "medium", "high"))
```

  - 이러한 항목들 간의 순서는 CSV 파일을 불러올 때 자동으로 생성될 수가 없기 때문에, 필요한 경우 위와 같이 직접 지정해주어야 함

요인의 탐색

- 요인 객체에서 항목들 간 순서가 있는 경우와 없는 경우는 데이터 항목이 표기되는 방식에서 차이가 있음
  - 순서가 있는 경우(서열 변수의 경우) 각 항목이 < 기호로 구분됨

```
> eyecol
[1] hazel brown black brown
Levels: black brown hazel blue
> pain
[1] low medium medium high
Levels: low < medium < high
```

  - 필요에 따라 `is.factor()`나 `is.ordered()`와 같은 별도의 함수를 사용하여 확인할 수도 있음*

요인의 탐색

- 요인 객체에서도 지금까지 살펴본 일반적인 벡터와 유사한 방식으로 특정한 값을 선택할 수 있음:

```
> eyecol[1:3]
[1] hazel brown black
Levels: black brown hazel blue
> eyecol[pain > "medium"]
[1] brown
Levels: black brown hazel blue
> pain[eyecol == "brown"]
[1] medium high
Levels: low < medium < high
```


요인 벡터와 문자열 벡터

- 문자열 벡터와 동일하게 비교 연산자 중 같다(==), 같지 않다(!=)와 같은 연산자를 사용할 수 있음
- 항목들 간 순서가 정해진 경우에는 숫자 유형의 벡터처럼 작다(<, <=), 크다(>, >=)와 같은 연산자를 모두 사용할 수 있음

```
> eyecol == "black"
[1] FALSE FALSE  TRUE FALSE
> eyecol < "black"
[1] NA NA NA NA
Warning message:
In Ops.factor(eyecol, "blue") :
  < not meaningful for factors
```

요인 벡터와 문자열 벡터

- 문자열 벡터와 요인 벡터는 유사하게 보이지만, 엄밀히 다른 유형의 데이터이기 때문에 사용에 주의가 필요함:

```
> nchar(fmd$Address)
Error in nchar(fmd$Address) : 'nchar()' requires a
character vector
```

- 일부 함수는 형 변환을 통해 요인 벡터를 문자열 벡터로 변환하는 경우도 있음:

```
> substring(fmd$Address, 5, 10)[1:10]
 [1] "an-ri," "a-ri, " "byeong" "gyang-" "ae-ri,"
 [6] "u-ri, " "ong-ri" "ri, Da" "ngwol-" "eon-ri"
```


요인 벡터와 문자열 벡터

- 문자열을 결합하는 용도로 많이 사용하는 `paste()` 함수도 그 중 하나로 아래와 같이 실행될 수 있음

```
> paste("The result is:", fmd$Result)
 [1] "The result is: Negative"
 [2] "The result is: Negative"
...
```

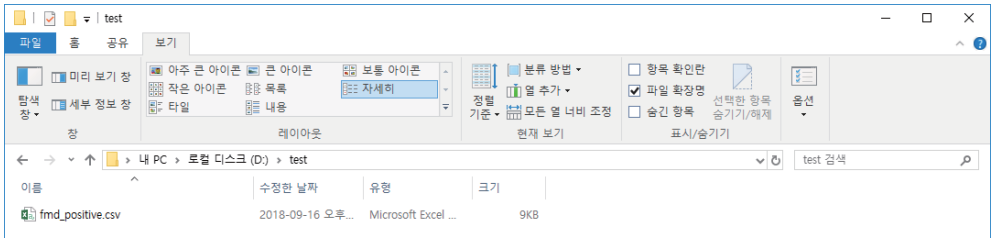
  – 주어진 문자열을 공백(space)으로 구분하여 결합하는 함수로, `sep` 인자를 사용하면 공백 외의 다른 기호로 문자열을 구분할 수 있음

```
> paste("Hello", "World", sep = ", ")
[1] "Hello, World"
```

데이터 저장하기

- R에서 수정, 가공된 객체는 다시 CSV 파일로 저장할 수 있음
  - 아래와 같이 전체 구제역 발생 신고 데이터에서 양성으로 판정된 농장만 별도로 저장하고자 한다면 아래와 같이 `write.csv()` 함수를 사용할 수 있음

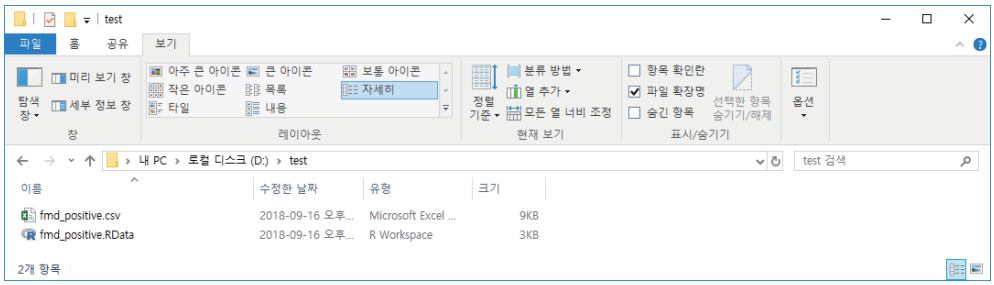
```
> fmd.pos <- fmd[fmd$Result == "Positive",]
> nrow(fmd.pos)
[1] 65
> write.csv(fmd.pos, file = "d:/test/fmd_positive.csv")
```



데이터 저장하기

- 객체는 CSV 파일 외에 다른 포맷으로도 저장될 수 있음
  - R에서 객체를 저장하는 기본 파일 포맷은 RData 포맷으로 `save()` 함수를 사용해 저장하고, `load()`를 통해 불러옴

```
> save(fmd.pos, file = "d:/test/fmd_positive.RData")
> rm(fmd.pos)
> load(file = "d:/test/fmd_positive.RData")
```



데이터 저장하기

- 일반적으로 CSV 파일에 비해 RData 파일이 좀 더 가볍고, 읽고 쓰는 속도가 빠를 수 있으나 다른 프로그램과의 호환성이 부족함
  - 대부분의 스프레드시트, 통계 프로그램 등이 CSV 파일을 지원하는 반면 RData 파일은 주로 R에서만 사용하게 됨
- 저장 목적에 따라 적절한 포맷을 사용하는 것이 중요함
  - 여러 사람과의 데이터 공유가 목적이라면 CSV 파일이 적합할 수 있음
  - 시간이 오래 소요되는 작업 중 백업 파일을 만드는 목적이라면 RData로 저장하는 것이 좀 더 효율적일 수 있음
  - 하나의 객체가 아니라 여러 개의 객체가 있는 전체 작업공간을 한 번에 저장하는 경우에도 RData 포맷을 사용하는 것이 좋음