Cambridge University Press
978-0-521-89715-0 - Geoinformatics: Cyberinfrastructure for the Solid Earth Sciences
Edited by G. Randy Keller and Chaitanya Baru
Excerpt
More information

Introduction

1

Science needs and challenges for geoinformatics

G. RANDY KELLER

1.1 What is geoinformatics?

Before we can begin to discuss geoscience informatics needs and challenges, we must first explain our use of the term geoinformatics for the purposes of this book. Over the past decade geoinformatics has become a term that has been independently employed by groups in several geospatial and geoscience fields around the world. In addition, this word appears in the title of several periodical publications. For example, there is an online magazine named *GeoInformatics* (www.geoinformatics.com) and an *International Journal of Geoinformatics* (www.j-geoinfo.net) that primarily focus on geospatial data and analysis within a geographic information system (GIS) framework. However, our emphasis in this book is on the data, software tools, and computational infrastructure that are needed to facilitate studies of the structure, dynamics, and evolution of the solid Earth through time, as well as the processes that act upon and within it from the near surface to the core. To approach such challenges, we must not only think and work in 3-D spatially, but we must include a 4th dimension, time. Time in this case ranges from seconds, such as in an earthquake, to millions of years, such as in plate movements over the Earth. *Here we have used **geoinformatics** to describe a variety of efforts to promote collaboration between computer scientists and geoscientists to solve complex scientific questions*. This book builds on the foundation of a book entitled *Geoinformatics: Data to Knowledge* (Sinha, 2006) that emphasized databases and their analysis, but here we emphasize topics such as web services, modeling of earth processes, visualization, and international developments.

At the U.S. National Science Foundation (NSF), geoinformatics has emerged as an initiative within the Earth Sciences Division to address the growing recognition that Earth functions as a complex system, and that existing information science infrastructure and practice within the geoscience community are inadequate to address the many difficult problems that must be overcome to understand this system (e.g., Allison *et al.*, 2002). In addition, there is now widespread recognition that successfully addressing these problems requires integrative and innovative approaches to analyzing, modeling, and developing extensive and diverse datasets.

Currently, the geoscience community is awash in data due to many new satellite observing systems that provide data to study phenomena such as changes in the Earth's surface via multi-band remote sensing (e.g., ASTER), the Earth's gravity field and small changes in it (e.g., GRACE), vertical movements of the Earth's surface (e.g., inSAR), the topography of the Earth (SRTM: Shuttle Radar Topography Mission), and the Earth's magnetic field (Maus *et al*., 2010). Also, massive amounts of seismological data are being archived in databases around the world. However, a lack of easy-to-use access to modeling and analysis codes are major obstacles for scientists and educators alike who attempt to use these data to their full potential, especially in a highly integrated fashion. However, recent advances in fields such as computational methods, visualization, and database interoperability provide practical means to overcome such problems and some examples are presented in this book. Thus, in *addition to the statement above*, ***geoinformatics*** *can be thought of as the field in which geoscientists and computer scientists are working together to provide the means to address a variety of complex scientific questions using advanced information technologies and integrated analysis*. This type of activity is also being called cyberinfrastructure.

1.2 Geoinformatics as a scientific tool is data driven

Open access to data from satellites is very common but spatial resolution is a limitation for many applications. In many cases, access to land-based or low-altitude measurements and even maps remains an issue in many countries due to government policies, but progress is being made on many fronts (e.g., gravity data, Aldouri *et al.*; seismic data, Casey and Ahern, this volume). Even though many useful datasets are emerging, discovering and accessing them is difficult if scientists wish to find the very best data for their particular application or research project. However, a very promising example of the development of an advanced data discovery and access system is the Global Earth Observation System of Systems (GEOSS) whose 10-Year Implementation Plan states that the purpose of GEOSS is "to realize a future wherein decisions and actions for the benefit of humankind are informed via coordinated, comprehensive and sustained Earth observations and information." GEOSS is seen by its participants as an important contribution to meeting United Nations Millennium Development Goals and to furthering the implementation of international treaty obligations (www.earthobservations.org).

In an ideal world, geospatial data developed by governmental agencies or by researchers using governmental support would be freely and openly available. However, crafting high-quality, easily accessible databases is expensive, especially if legacy data are to be converted to digital form. Thus in many cases, it is not possible for data to be accessible free of charge, but costs need to be low enough to make them available to a broad cross-section of users (e.g., Jackson and Hughes, this volume).

In order to understand the subsurface of the Earth, one has to be able to correlate known geological features with geophysical data and models. OneGeology (Jackson, this volume) is an example of international geological organizations banding together to produce a geological map of Earth's surface. Such a product would be invaluable to countless researchers, governmental agencies, environmental protection efforts, and planning efforts to name a few.

1.3 Geoinformatics as a scientific tool seeks to foster the development of community-based software

A guiding principle in geoinformatics is fostering community-based development of software that is open source and highly usable (e.g., Gurnis *et al.*, this volume). In the following chapter, Baru discusses the technical issues and developments that affect this and other technical challenges that affect geoinformatics, but below I discuss an example of major scientific need.

1.3.1 Building 3-D models

Today, a major research goal in the geosciences is the construction of geologically realistic (i.e., as complex as in nature) 3-D models of earth structure and variations in physical properties such as seismic velocity (P-wave and S-wave), density, and electrical resistivity. The physical basis of many geophysical techniques is inherently scale-independent, so it is realistic to aspire to build models that range in scale from the near surface (environmental and groundwater studies), to geologic studies of features such as basins and fault zones, to studies of tectonic plates and their boundaries (e.g., Boyden *et al.* and Liu *et al.*, this volume), to mantle dynamics, to studies of the core and its boundaries. In order to construct such models, software that enables the integration of a wide range of geological and geophysical data is required. This software should also facilitate the application of empirical and theoretical relationships that provide constraints for integrated modeling via estimations of relationships between various physical properties (e.g., P-wave velocity, S-wave velocity, and density; Brocher, 2005), the effects of porosity (e.g., Mavko *et al.*, 1998), and the effects of pressure and temperature (e.g., Perry *et al.*, 2006).

One way to conceive of an ideal model would be for it to consist of geological structures and major discontinuities in physical properties that are represented by surfaces that bound layers and within which variations in multiple physical properties are associated with voxels, which need not be cubical in form. Since the resolution of geophysical techniques decreases with depth, it would make sense that the size of the

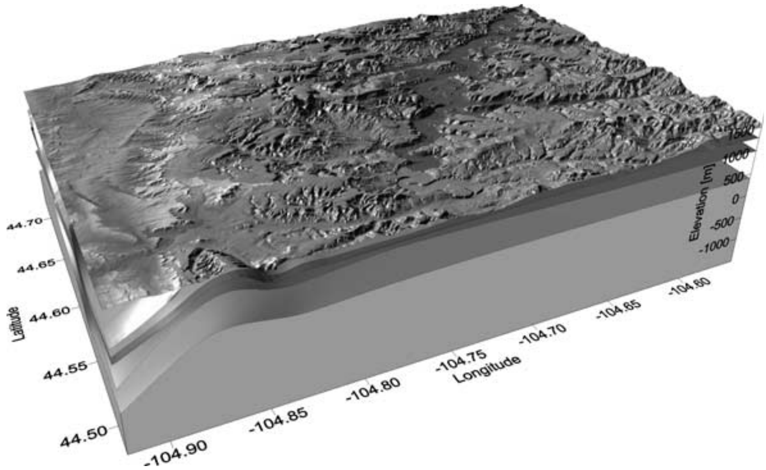

Figure 1.1. Example of a 3-D geological/geophysical model consisting of layers that are bounded by geologic interfaces that have been extracted from surface and subsurface geologic data. The lowest layer is the Precambrian basement. The interfaces are georeferenced and provide a framework for assigning physical properties to the layers between them. Image provided by Kevin Crain. See color plates section.

voxels would increase with depth. This type of model is shown in Figure 1.1, where the topographic relief and surfaces that represent the tops of a series of stratigraphic units are shown above the last surface, which is the top of the Precambrian basement. The concept is that these surfaces bound the stratigraphic layers and Precambrian basement that form the model. These layers can then be populated with voxels with associated physical properties based on studies of samples collected from exposures, data from drill holes, and geophysical surveys. In this ideal case, the resulting model would be structured in a form that would facilitate calculations such as various geophysical responses, fluid flow in the layers, and response to stress. Modeling a response to stress would be an example of adding the dimension of time to the analysis.

In most cases, seismic data have the highest spatial resolution (and cost) of subsurface imaging techniques, and many diverse techniques are available to process and analyze these data at various spatial and depth scales. Each type of seismic data has its own sensitivities and resolution and can constrain important aspects of earth structure. For example, tomographic modeling is based on voxels, seismic refraction/wide-angle reflection data produce models with interfaces and velocity values measured directly, and seismic reflection data produce images of earth structures from which surfaces and discontinuities such as faults can be extracted. It is intuitively obvious that, when a variety of seismic data are used together in a quantitative manner, the resulting earth model should be better resolved than in the typical approach of simply comparing results qualitatively. However, proving this inference mathematically is not easy. As constraints from geological and drilling data and other geophysical techniques are added, the resolution will improve further, which is also hard to prove mathematically. These extra data also make it possible to add non-seismic physical properties (e.g., density, electrical conductivity, magnetic susceptibility) to the model.

Tools for modeling seismic data and honoring independent constraints exist for 2-D approaches, and an example of some preliminary results from a large experiment in Central Europe (Figure 1.2) are shown in Figure 1.3. The final scientific results of the analysis of the long profile (CEL05, Fig. 2) are presented in Grad *et al.* (2006).

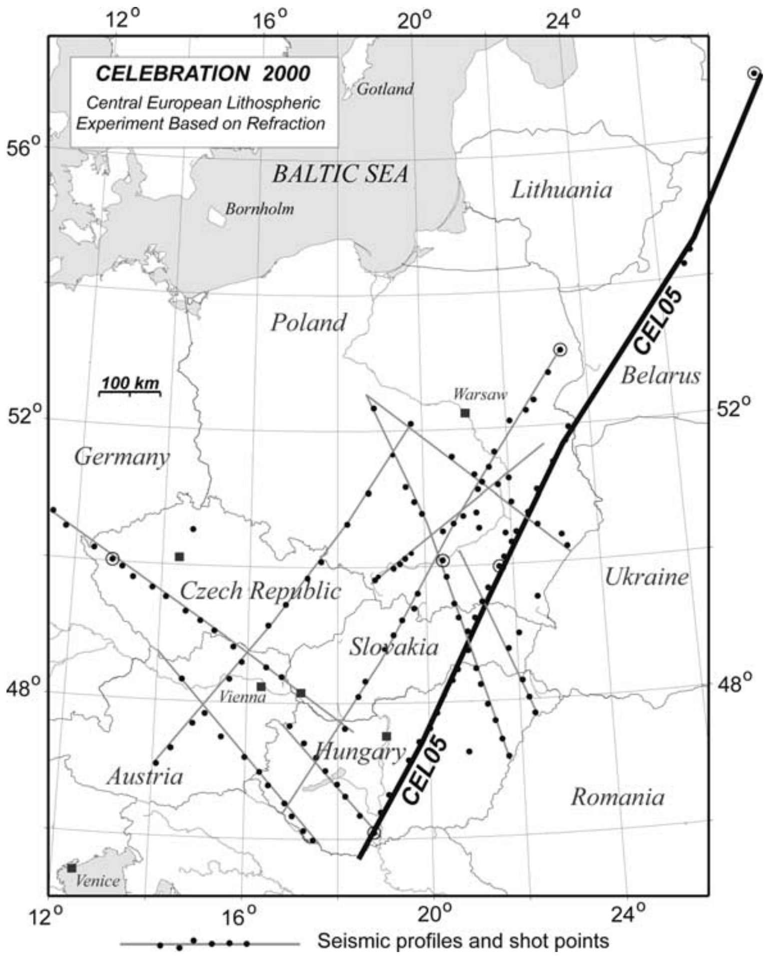

Figure 1.2. Index map of the CELEBRATION 2000 seismic experiment showing the location of the 1400 km long CEL05 profile (heavy black line). The gray lines indicate the location of other profiles that were recorded. The seismic velocity models shown in Figure 1.3 are for this profile.

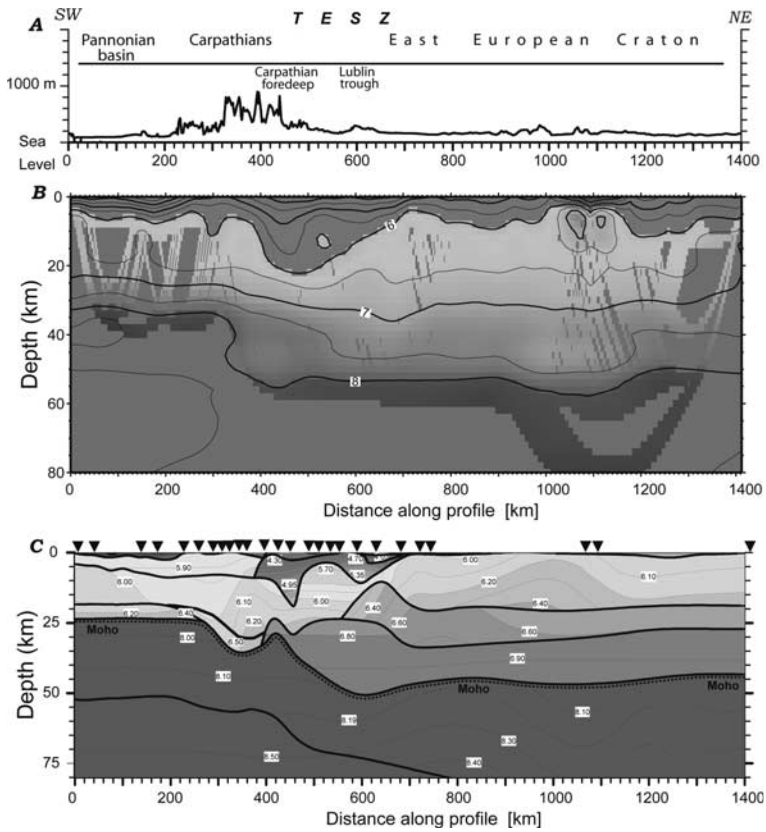

Figure 1.3. (*A*) Topographic profile showing the main geologic features present; (*B*) Preliminary seismic velocity model derived by tomographic inversion of the arrival times of the first seismic wave observed. The model is smooth and lacks the detail that is needed to make a suitable geological interpretation. The numbers in the model are P-wave velocities in km/s; (*C*) Seismic velocity model derived by ray trace forward (trail-and-error) modeling of all observed seismic arrivals. This approach has the advantage of providing more detail, but the formal analysis of certainty is difficult. The numbers in the model are P-wave velocities in km/s. Inverted triangles indicate the locations' shot points that produced the observed seismograms. See color plates section.

The tomographic result (*B*) shows the broad variations in seismic velocity based on voxels. Using the tomographic result as a starting point, modeling of waves reflected and refracted at interfaces from within the Earth add structural detail (*C*) that can be interpreted geologically. In turn, the upper few kilometers of the model could be further refined using geological, drilling, and other types of geophysical data. Presently, expanding this example of an analysis scheme to 3-D, quantitatively assessing resolution, and moving smoothly between modeling approaches are at best very challenging. The software tools that do exist for 3-D modeling (e.g., Hole, 1992) need further development, need to be interoperable, and need to facilitate integrated analysis.

In summary, scientific advances on many fronts face technical barriers that require a geoinformatics approach if they are to be overcome. In a lot of cases, there are large volumes of data to examine and mine, and in others, interoperability between analysis and modeling software is needed. Obviously, providing the "best" integrated model of earth structure possible with existing data is a goal that we are far from achieving, except in very special circumstances. Thus, geoscientists and computer scientists have many interesting and important problems that they can attack together in the future.

References

Allison, M. L., Snyder, W. S., and Walker, J. D. (2002). Geoinformatics: A nascent revolution in the earth sciences. *GSA Today*, **12**: 17.

Brocher, T. M. (2005). Empirical relations between elastic wavespeeds and density in the Earth's crust. *Bulletin of the Seismological Society of America*, **95**: 2081–2092.

Grad, M., Guterch, A., Keller, G. R. *et al*. (2006). Lithospheric structure beneath trans-Carpathian transect from Precambrian platform to Pannonian basin: CELEBRATION 2000 seismic profile CEL05. *Journal of Geophysical Research*, **111**, B03301, doi:10.1029/2005JB003647.

Hole, J. A. (1992). Nonlinear high-resolution three-dimensional seismic travel time tomography. *Journal of Geophysical Research*, **97**: 6553–6562.

Maus, S., Barckhausen, U., Berkenbosch, H. *et al*. (2010). EMAG2: A 2-arc min resolution Earth Magnetic Anomaly Grid compiled from satellite, airborne, and marine magnetic measurements. *Geochemistry Geophysics Geosystems*, **10**, Q08005, doi:10.1029/2009GC002471.

Mavko, G., Mukerji, T., and Dvorkin, J. (1998). *The Rock Physics Handbook: Tools for Seismic Analysis in Porous Media*. Cambridge: Cambridge University Press, 329pp.

Perry, H. K. C., Jaupart, C., Mareschal, J.-C., and Shapiro, N. M. (2006). Upper mantle velocity-temperature conversion and composition determined from seismic refraction and heat flow. *Journal of Geophysical Research*, **111**, B07301, doi:10.1029/2005JB003921.

Sinha, A. K., ed. (2006). *Geoinformatics: Data to Knowledge*. Geological Society of America, Special Paper 397, 283pp.

2

Introduction to IT concepts and challenges

CHAITANYA BARU

Scientific applications have been at the forefront of driving computer and information technology since the early days: from the development of early computers for numerical computing, to the introduction in the USA of the NSFNET (which helped launch the Internet), and the subsequent invention of the World Wide Web. The geosciences, in particular, have been a long-standing user of such technologies, given the importance of applications related to weather, natural resources, natural hazards, and environmental monitoring. Scientific computing was focused initially on the need for fast computers to perform larger numbers of complex numerical calculations. The concerns more recently have turned towards the ability to manage the very large amounts of data that are being generated by a wide range of sensors and instruments, sophisticated observing systems, and large-scale simulations on large computer systems. Data rates of terabytes per day and petabytes per year are not uncommon (1 petabyte = terabytes) (Hey *et al.*, 2009, p. 9). Yet, computer science and information technology solutions must deal not only with the size and scale of data, but also the inherent richness and complexity of scientific data – especially when data are combined across multiple projects, institutions, and even multiple science disciplines and subdisciplines. The need to understand complex, interdependent, natural as well as anthropogenic phenomena has made science a *team sport*, requiring collaborations among multidisciplinary teams of scientists to process, analyze, and integrate extremely heterogeneous data.

The *e-Science* initiative in Europe and the *cyberinfrastructure* initiative in the United States were launched in the early 2000s to tackle these issues, by harnessing the power of advanced information technologies for scientific research and education. Scientific research, it has been suggested, has entered the *fourth paradigm* (Hey *et al.*, 2009). The first three being *empirical*: focused on observations and descriptions of natural phenomena; *theoretical*: focused on the development and use of models and generalization of scientific principles; and, *computational*: focused on simulations of complex phenomena using computers. This fourth paradigm is *data intensive*, focused on building unified theories of complex phenomena, but based on data exploration and integration using software tools and computer platforms capable of dealing with complex data and large data (Hey *et al.*, 2009, p. 177).

2.1 Cyberinfrastructure and geoinformatics

The study of complex phenomena in earth, ocean, and atmospheric sciences all require integration of heterogeneous data from a wide variety of sources and disciplines. As in every area of science, discovery in the geosciences is also driven by the ease and efficiency with which one is able to do this integration by manipulating and assimilating large, heterogeneous datasets. Remote sensing instrument and observing systems are able to generate rapidly large amounts of data, while large-scale computational models are able to generate increasingly large outputs that require post-processing, visualization, and eventually integration with other simulation, observational, and contextual data. A range of cyberinfrastructure capabilities is needed to provide such capabilities and to support scientific research and discovery at the frontiers of the earth sciences.

NSF's *Cyberinfrastructure Vision for 21st Century Discovery* describes the set of challenges and opportunities in computing systems, data, information resources, networking, digitally enabled sensors, instruments, virtual organizations, and observatories, along with an interoperable suite of software services and tools (NSF, 2007). As described in the report, this technology is complemented by the interdisciplinary teams of professionals who are responsible for its development, deployment, and its use in transformative approaches to scientific and engineering discovery and learning. The vision also includes attention to the educational and workforce initiatives necessary for both the creation and effective use of cyberinfrastructure. Figure 2.1 depicts the set of cyberinfrastructure components, from hardware platforms, systems software, middleware services, user services/functions, and a portal providing access to this environment.

As mentioned in Chapter 1, geoinformatics is the term used to describe the set of activities related to the development and use of cyberinfrastructure for the earth sciences. The area has been making rapid progress since the early 2000s, with the introduction by NSF of its cyberinfrastructure initiative and, subsequently, the *geoinformatics* program in the Earth Sciences Division (NSF EAR, 2010). Since then, major geosciences professional organizations have also recognized geoinformatics as a special area. Both the American Geophysical Union (AGU) and the European Geophysical Union have an Earth and Space Science Informatics focus area (AGU, 2010; EGU, 2010). The Geological Society of America created a Geoinformatics division, which defined geoinformatics as "*the science discipline that utilizes cyber-products, tools and discovery of data and*