BigGIS

Nutzung von Big-Data-Technologien für den Umgang mit invasiven Spezies

Hannes Müller
Gesellschaft für Angewandte Hydrologie und Kartographie mbH
Rehlingstr. 9
79100 Freiburg im Breisgau

Andreas Abecker; Johannes Kutterer
Disy Informationssysteme GmbH
Ludwig-Erhard-Allee 6
76131 Karlsruhe

Daniel Seebacher
Universität Konstanz
Universitätsstr. 10
78464 Konstanz

Wolfgang Schillinger
Landesanstalt für Umwelt Baden-Württemberg
Griesbachstr. 1
76185 Karlsruhe

Kapitelübersicht

1. Überblick zum Forschungsprojekt BigGIS

Das vom Bundesministerium für Bildung und Forschung (BMBF) geförderte Forschungsprojekt BigGIS hatte zum Ziel, Big-Data-Technologien mit der Welt der Geoinformationssysteme (GIS) zu verbinden. Innerhalb seiner 3-jährigen Laufzeit (2015–2018) wurde ein prototypisches System entwickelt, welches in verschiedenen Anwendungsfällen Entscheidungen auf Basis von großen Mengen an heterogenen, geo-temporalen Daten besser und schneller unterstützt, als es mit herkömmlichen GIS-Produkten möglich ist. Das BigGIS-Projekt hat die verschiedenen Anwendungsfälle in drei übergeordnete Szenarien zusammengefasst, welche die Flexibilität und Relevanz des Zielsystems demonstrieren sollen:

1) Smart City Szenario
2) BOS Szenario (Behörden und Organisationen mit Sicherheitsaufgaben)
3) Umwelt Szenario

Von diesen drei Szenarien wollen wir in den nachfolgenden Kapiteln vor allem auf die Datenpipeline und Ergebnisdarstellung des Umweltszenarios eingehen, weil hier der Schwerpunkt bei den Arbeiten der BigGIS-Projektpartner Landesanstalt für Umwelt Baden-Württemberg (LUBW) und Disy gelegt wurde. Für weitere Informationen zum BOS und zum Smart City Szenario siehe die BigGIS Projekt-Webseite:

http://biggis-project.eu/biggis-docs/scenarios/environment/.

Ein wichtiges Anwendungsfeld für GIS ist das Management von (oftmals in der jüngsten Vergangenheit eingewanderten) Tieren und Pflanzen mit negativen Auswirkungen auf die menschliche Gesundheit, das ökologische Gleichgewicht oder ökonomische Interessen (wie landwirtschaftlicher Ertrag). Ein Beispiel für eine solche „invasive Spezies“ ist die Kirschessigfliege *(Drosophila suzukii),* im Folgenden auch KEF abgekürzt.

Im Gegensatz zu anderen Drosophilae befällt sie gesunde Früchte und stellt daher für Obstbauern und für Winzer eine große wirtschaftliche Bedrohung dar. Die wichtigsten kommerziellen Auswirkungen betreffen Sommerfrüchte wie Kirschen, Blaubeeren, Trauben, Nektarinen, Birnen, Pflaumen, Pfirsiche, Himbeeren und Erdbeeren. In den Vereinigten Staaten schwanken die Schätzungen der Auswirkungen durch KEF-Befall stark, erreichten jedoch in einigen Gebieten und Kulturen einen Ertragsverlust von 80 % /1/. Auch in Deutschland hat sich die Fliege seit 2011 stark verbreitet und bereits zu großen Ertragseinbußen in der Rotweinernte geführt /2/. Für die Bekämpfung und den Schutz vor Schädlingen wie der Kirschessigfliege lassen sich verschiedene Informationsquellen in Echtzeit kombinieren und analysieren, um das Ausbreitungsverhalten besser beobachten, verstehen und vorhersagen zu können. Das BigGIS-Projekt adressiert diese Herausforderung durch die folgenden Verarbeitungsschritte:

- Sammeln von Daten aus verschiedenen Quellen wie aktuelle Verteilung der Fliege, Wetterbedingungen, Elevation, Landnutzung usw.
- Datenanalyse zur Vorhersage von Risikobereichen
- Interaktive Visualisierung von Risikobereichen und Infektionswahrscheinlichkeiten

2. Infrastrukturkomponenten und Vektordatenpipeline

Konkrete BigGIS-Anwendungslösungen werden aus einem Werkzeugkasten aufeinander abgestimmter Softwarekomponenten zusammengestellt. Die wichtigsten Infrastruktur- und Lösungskomponenten sind in der BigGIS-Lösungsarchitektur in Abb. 1 eingeordnet.

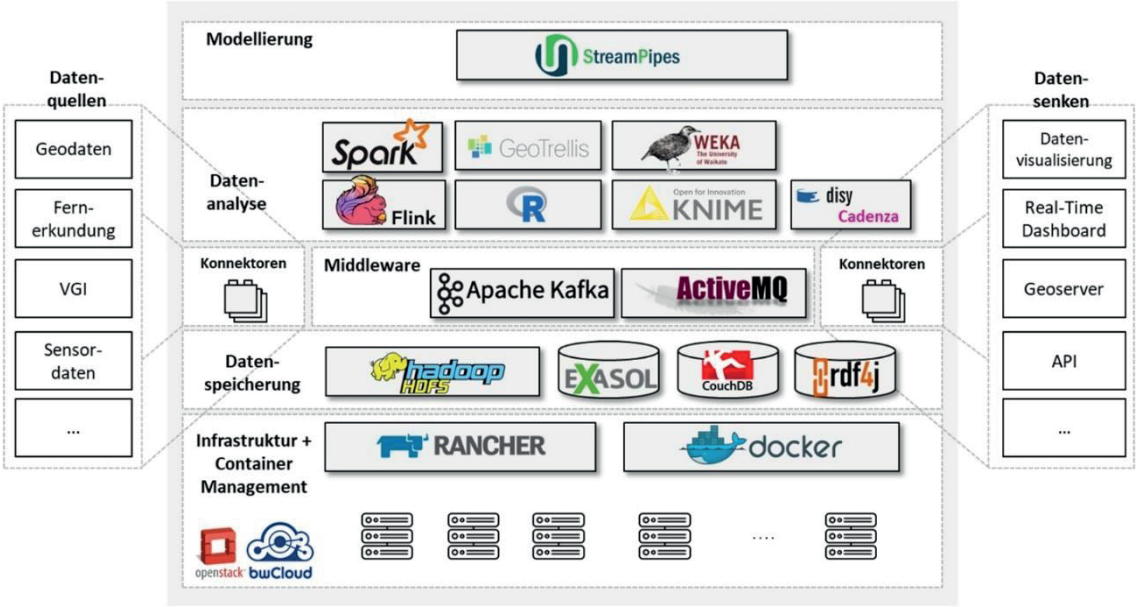

Abbildung 1: Ausgewählte Softwarekomponenten für BigGIS-Anwendungslösungen

Auf der Infrastrukturebene stehen Werkzeuge wie Docker und Rancher für die Bereitstellung der erforderlichen Speicher- und Rechenressourcen und für das automatisierte Management dieser Virtualisierungsschicht. Dadurch kann eine BigGIS-Lösung container-basiert verhältnismäßig einfach installiert und im Fall starker Schwankungen im Datenanfall und Rechenanforderungen leicht automatisch skaliert werden.

Auf der Ebene der Datenspeicherung kommen verschiedene moderne Datenbanktechnologien im Projekt zum Einsatz. Disy führte Benchmarks von In-Memory-Datenbanken wie SAP HANA im Vergleich mit etablierten Geodatenbanken wie Oracle Spatial und Postgres/ PostGIS durch. Der BigGIS-Projektpartner Exasol erweiterte die Geodatenfähigkeiten seiner bereits führenden In-Memory Datenbank erheblich. Dokumentorientierte NoSQL-Ansätze für die hochverfügbare Datenspeicherung (Hadoop HDFS, CouchDB) werden insbesondere für sehr große Datenvolumina wie Rohdaten aus der Fernerkundung verwendet. RDF4J (früher Sesame) dient zur Speicherung semantischer Metadaten im Projekt.

Auf der Middleware-Ebene geht es primär um das Message Brokering, also die Entkopplung der direkten synchronen Kommunikation von eingehenden Datenströmen mit Weiterverarbeitungsprozessen. Hier werden Werkzeuge wie Kafka und ActiveMQ genutzt.

Die Ebene der Data Analytics war der Projektschwerpunkt verschiedener Partner. Vielfältige Tools und Frameworks stehen zur Verfügung, wie Flink, R und Spark als sehr generische

Werkzeuge für (unter anderem) Datenanalyse und Maschinelles Lernen, oder auch GeoTrellis als Spezialwerkzeug für große Mengen von Raster-Geodaten. Die Universität Konstanz nutzte Maschinelle-Lern-Algorithmen aus KNIME und WEKA und entwickelte verschiedene neue Visualisierungsmöglichkeiten zur interaktiven Datenanalyse (siehe Kapitel 3 in diesem Beitrag). Disy integrierte die Teillösungen prototypisch und koppelte sie mit Cadenza, der eigenen Plattform für Data Analytics, Reporting und GIS. Die Ebene der Modellierung unterstützt die Konfiguration komplexer Big-Data-Verarbeitungsworkflows mithilfe des FZI-Werkzeugs StreamPipes.

Im konkreten Beispielszenario der Untersuchung der Kirschessigfliege wurden Lernalgorithmen aus den Data-Mining-Systemen KNIME und WEKA genutzt, um räumlich-zeitliche Befallsprognosen zu erstellen, die dann in von der Universität Konstanz eigens entwickelten Visualisierungen für weitere Untersuchungen dargestellt wurden (siehe Kapitel 3). Die Datengrundlage umfasste neben Geobasisdaten (Landnutzung, Geländemodell) und Wetterdaten des Deutschen Wetterdienstes (DWD) insbesondere georeferenzierte Beobachtungsdaten von VitiMeteo, dem Informationsdienst des Staatlichen Weinbauinstituts Freiburg, nämlich: (i) Daten zu Fallenfängen, (ii) Daten zu Eifunden von Mitarbeitern des WBI Freiburg sowie der LVWO Weinsberg und (iii) Befallsbeobachtungen, die vor Ort von Rebschutzwarten, Weinbauberatern und Mitarbeitern des WBI Freiburg sowie der LVWO Weinsberg gemacht wurden.

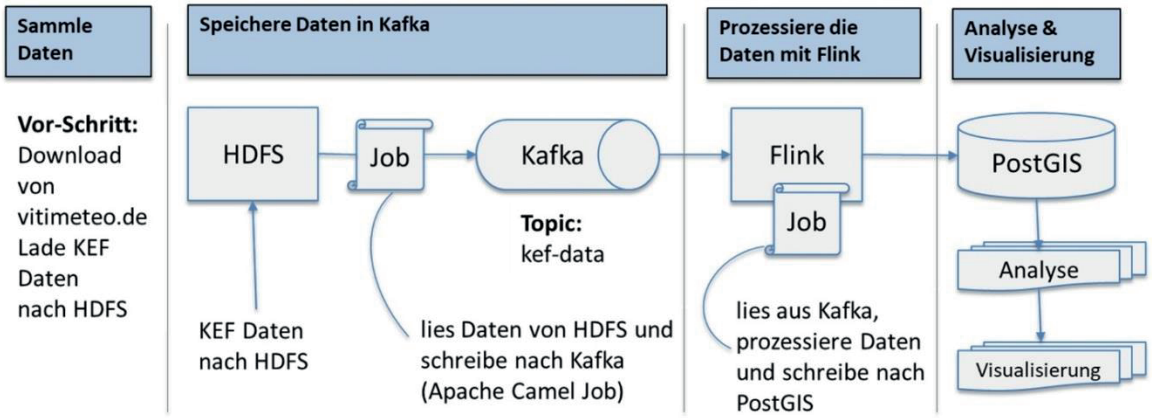

Abbildung 2: Einfache Datenpipeline zum Einlesen der Beobachtungsdaten von VitiMeteo

Abb. 2 zeigt den Prozess der Datenintegration von VitiMeteo ins BigGIS-System. Auch wenn hier keine riesigen Datenmengen oder -ströme auftreten, zeigt das Beispiel doch schon das Zusammenspiel der meisten Softwarekomponenten: Um die Webseite nicht übermäßig zu belasten, wurden die VitiMeteo-Daten zu Versuchszwecken einmal heruntergeladen und lokal im HDFS abgelegt. Natürlich würden die nachfolgenden Schritte genauso funktionieren, wenn sie online mit einem Datenstrom von VitiMeteo verbunden wären. Aus HDFS werden die Daten dann als GeoJSON-Datei in Kafka eingespeist. Kafka entkoppelt Datenlieferanten und Datenkonsumenten, so dass auch im Fall eines zeitweiligen Ausfalls der nachgelagerten Verarbeitungsprozesse oder bei extrem großem Datenaufkommen seitens der Datenlieferanten keine Daten verlorengehen können. Die Prozessierung der Daten erfolgt dann in einem

Flink-Job, der die Daten aus Kafka liest und in die PostGIS-Datenbank schreibt. Der Flink-Job validiert die Daten und erzeugt georeferenzierte Beobachtungen aus ursprünglich zwei Datensätzen (Fallen-Standorte einerseits und Beobachtungszeitreihe pro Falle andererseits). Die PostGIS-DB ist dann die Ausgangsbasis für die Lernverfahren und die darauf aufsetzenden Visualisierungen der Ergebnisse in Kapitel 3.

3. Analyse und Visualisierung Umweltszenario

Um die Ausbreitung der Kirschessigfliege anhand bestimmter Faktoren vorherzusagen, wurden die von VitiMeteo zur Verfügung gestellten datumsbezogenen Fallenfänge mit Datensätzen zur Landnutzung (ATKIS/ALKIS) und Geländetopographie (ASTER Geländemodell) angereichert. Zusätzlich wurden die Klimadaten des DWD verwendet, da die Kirschessigfliege anfällig für Temperaturschwankungen ist. Als nächster Schritt wurde ein Ensemble von Klassifikatoren trainiert, um auf Grundlage der Landnutzung, Geländetopographie sowie der Klimadaten des DWD die Risikogebiete der Kirschessigfliege räumlich und zeitlich explizit vorherzusagen (detaillierte Beschreibung siehe /3/). Aufgrund der zeitlichen Variabilität der Fallenfänge wurde diese Vorhersage für alle 12 Monate eines Jahres durchgeführt, was insgesamt über 20.000 Vorhersagen über alle Monate und Weingebiete Baden-Württembergs ergab.

Für die räumliche und zeitliche Darstellung der über 20.000 Vorhersagen und deren Modellunsicherheit wurde eine interaktive Kartendarstellung entwickelt, im Folgenden „Drosophigator“ genannt (siehe Abb. 3). Der Drosophigator besteht aus folgenden vier Darstellungsebenen, die eng miteinander verzahnt sind:

- Kartenkomponente
- Glyphendarstellung
- Gereihte Koordinatenachsen
- Liniendiagramm

Innerhalb der Kartendarstellung erlaubt der Drosophigator stufenloses Zoomen und damit eine dynamische Aggregierung der Modellergebnisse auf unterschiedlichen räumlichen Skalen für alle vier Darstellungseben. Diese stufenlose dynamische Aggregation erlaubt es, Gebiete unterschiedlicher Größen zu analysieren, was es Experten ermöglicht, die Auswirkung mikro- und makroökologischer Faktoren zu untersuchen. Die Glyphendarstellung ist wie eine Uhr aufgebaut, mit 12 Segmenten für die 12 Monate (siehe Abb. 4). Innerhalb eines Segments wird über den Füllstand der blauen und roten Einfärbung die Befallswahrscheinlichkeit für die Weinbaugebiete visualisiert (rot= Befall, blau= kein Befall).

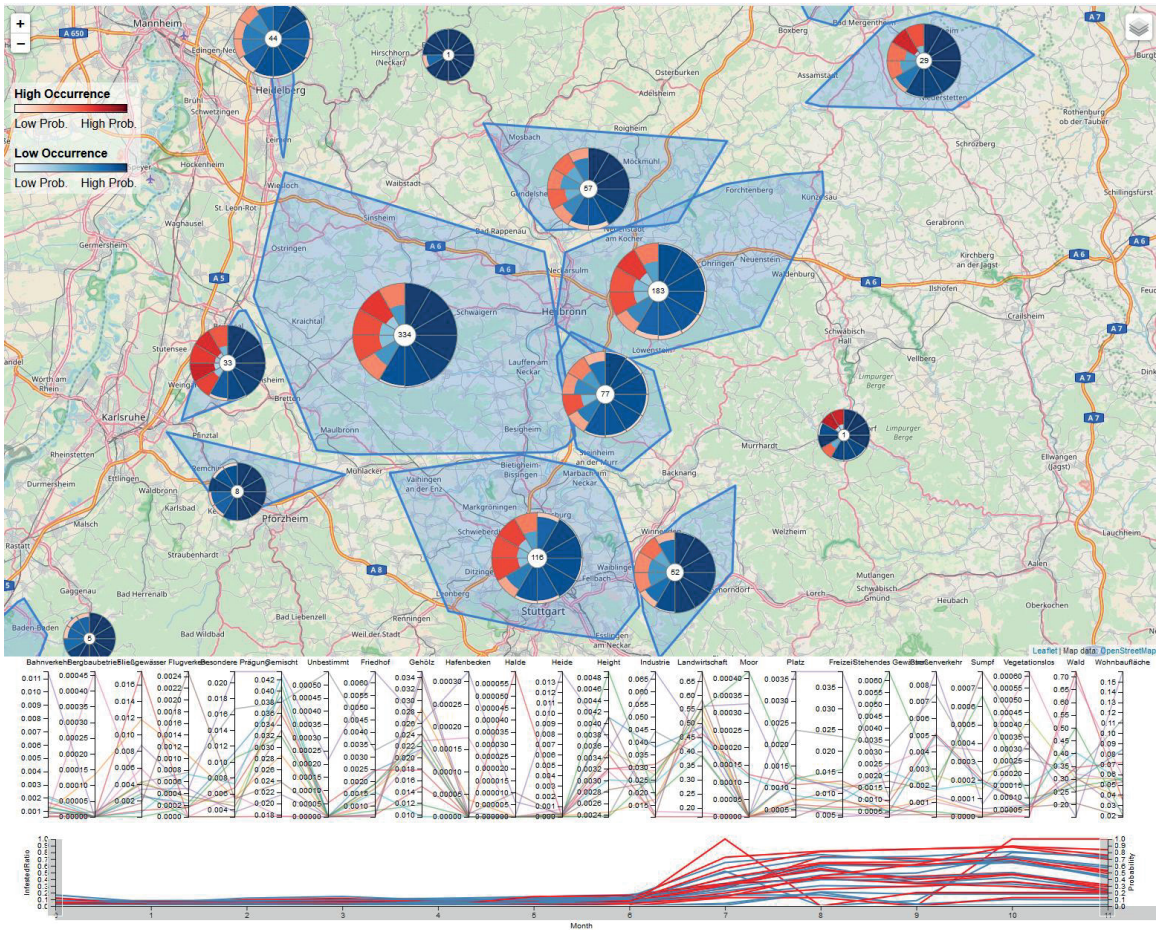

Abbildung 3: Interaktives Tool zur Analyse der Auftrittswahrscheinlichkeit der Kirschessigfliege (Drosophigator) mit Kartenkomponente, Glyphendarstellung, gereihten Koordinatenachsen und Liniendiagramm. Quelle: /4/.

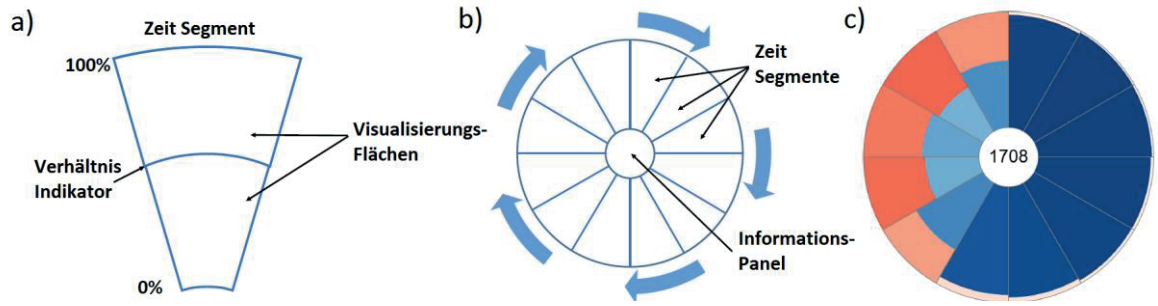

Abbildung 4: Aufbau einer Glyphendarstellung mit a) Beispiel eines Zeitsegmentes, b) zeitlicher Ereignisvorhersage und c) einer tatsächlichen Glyphe. Geändert von /4/.

Über die gereihte Achsendarstellung lassen sich die Informationen zur Landnutzung und Topographie der jeweiligen Weinbaugebiete einblenden. Die räumliche Einheit wird außerdem mit einem Polygon hinterlegt (siehe Abb. 5). Kenner der Region können auf diese Weise die Modellvorhersage besser evaluieren und zu einer eigenen Einschätzung gelangen. Auf diese Weise unterstützt das System die Nutzung von Expertenwissen.

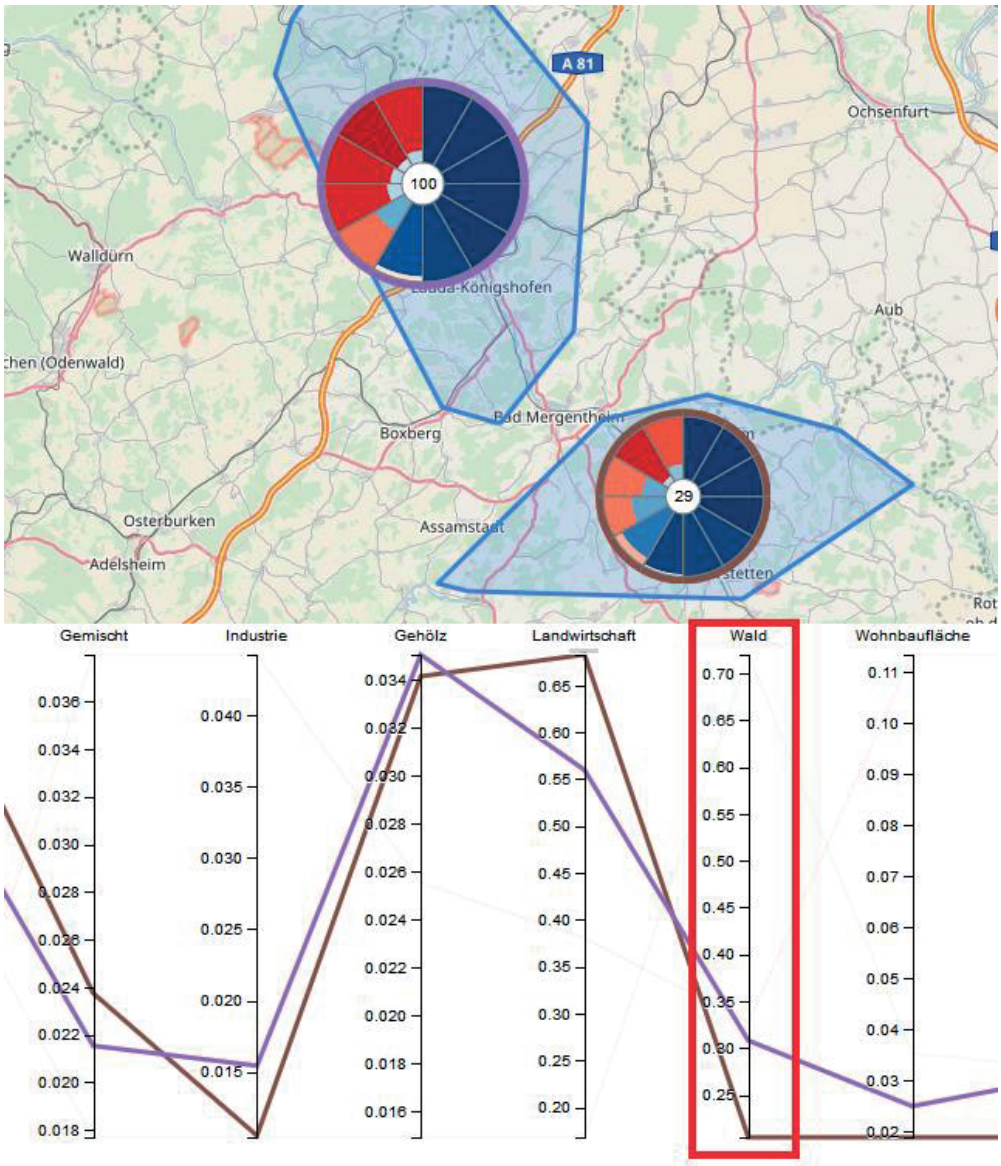

Abbildung 5: Auswahl von zwei verschiedenen Weinbaugebieten (lila und brauner Glyphenrand). Die gereihte Achsendarstellung erlaubt den direkten Vergleich der jeweils hinterlegten Landnutzung. Für Vergleichszwecke können beliebig viele Gebiete gleichzeitig ausgewählt werden. Quelle: /4/.

4. Fazit

Im vorliegenden Beitrag haben wir einerseits die Software-Technologien für Big-Data aus dem BigGIS-Projekt skizziert und andererseits einige Lern- und Visualisierungsverfahren anhand des Beispiels der Kirschessigfliege präsentiert. Auch wenn dieses konkrete Beispiel aus Big-Data-Sicht noch keine extremen Anforderungen hinsichtlich zu verarbeitender Datenvolumina oder Echtzeitverarbeitung stellt, kann es doch als Proof-of-concept dienen, und unsere Benchmarks zeigen, dass die verwendete Basisinfrastruktur auch hochskalierbar arbeitet. Besonders überzeugend für den praktischen Einsatz waren beispielsweise die container-basierten Virtualisierungsansätze, das Message Brokering mit Kafka und die verteilte Daten-prozessierung mit Flink-Jobs. Schon wenn man in zukünftigen Erweiterungen feinkörnigere

Wetterdaten bzw. lokale Wetterprognosen oder Echtzeitbeobachtungsdaten für die Online-Analyse verwendet, werden solche Aspekte wie die leistungsfähige Datenstromverarbeitung sofort relevant. Wenn man den Beobachtungsrahmen von den amtlichen Daten eines einzigen Bundeslandes ausweitet auf grenzüberschreitende Daten und ggf. nutzergenerierte Daten aus der Öffentlichkeit einbindet, werden auch verfeinerte Methoden für die Datenvalidierung und Datenharmonisierung erforderlich, wie sie in der BigGIS-Architektur bereits vorbereitet sind.

Die Nutzung fortgeschrittener Maschinelle-Lern-Methoden zur Offline-Analyse für die Befallsprognose auf Basis von historischen Daten, Landnutzungsdaten und Wetter wird in /3/ ausführlich diskutiert. Es gibt hier noch vielfältige mögliche Ansatzpunkte (zusätzliche Input-Daten, andere zeitliche und räumliche Datenstrukturierung) für weitere Untersuchungen. Die Darstellung und interaktive visuelle Inspektion von Analyseergebnissen mithilfe des Drosophigators wird ebenfalls in /3/ ausführlicher diskutiert und wurde im Dezember 2017 beim 6. Workshop der Arbeitsgruppe "D. suzukii" in Bad Kreuznach demonstriert und evaluiert.

Die Ergebnisse der Evaluation machen deutlich, dass ein starker Bedarf an intuitiven und interaktiven Systemen besteht, die die Experten bei ihren täglichen Analyseaufgaben unterstützen. Die Experten sind größtenteils sehr zufrieden mit den Inhalten des Drosophigators, insbesondere mit der Glyphendarstellung und den zoombaren Aggregationsstufen. Allerdings ist der Nutzen der Anwendung zurzeit noch eher zur Erweiterung des allgemeinen Verständnisses geeignet, als um direkt Ursachen für das Auftreten von *D. suzukii* abzuleiten. Langfristig sind natürlich sowohl die Nutzbarkeit für die Auftrittsanalyse anderer invasiver Spezies bzw. allgemein schädlicher Spezies von Interesse als auch die Nutzung der verwendeten Methoden für das bessere Verständnis anderer biologisch-ökologischer lokaler oder regionaler Phänomene wie z. B. regionales Bienensterben. Langfristig könnten sich aus der Kombination solider Offline-Analysen über die Verbreitungsprozesse von Schädlingen mit aktuellen Beobachtungen auch Ansätze für proaktive Umweltmanagementsysteme in Nah-Echtzeit ergeben, die dabei helfen, effektiver und effizienter mit auftretenden Schädlingspopulationen umzugehen.

Die LUBW und Disy haben im Rahmen des BigGIS-Projektes zusammen mit den akademischen Projektpartnern Erfahrungen mit Big-Data-Technologien gesammelt, die in zukünftige Digitalisierungsprojekte in der Umweltverwaltung einfließen können. Die in diesem Beitrag skizzierten Visualisierungsmethoden bilden einen wertvollen Werkzeugkasten, der auf andere komplexe Datensätze (z. B. in der hydrologischen Modellierung) angewandt werden kann. Dank der ausführlichen Dokumentation des BigGIS-Projekts inklusive freiem Zugriff auf GitHub Repositories /4/ stehen auch für die INOVUM-Partner viele der technologischen Entwicklungen aus BigGIS zur Verfügung.

Danksagung: Die vorgestellten Arbeiten wurden mit Unterstützung des Bundesministeriums für Bildung und Forschung (BMBF) im Rahmen des Big-Data Forschungsprojektes „BigGIS: Prädiktive und präskriptive Geoinformationssysteme basierend auf hochdimensionalen geotemporalen Datenstrukturen“ (FKZ: 01IS14012A-G) durchgeführt.

5. Literatur

/1/ Bolda, M.P., Goodhue, R.E., Zalom, F.G. (2009): Spotted Wing Drosophila: Potential Economic Impact of Newly Established Pest (PDF). Giannini Foundation of Agricultural Economics, University of California, https://s.giannini.ucop.edu/uploads/giannini_public/81/fe/81feb5c9-f722-4018-85ec-64519d1bbc95/v13n3_2.pdf, abgerufen am 15.05.2018.

/2/ Julius Kühn Institut (2018): Wissensportal *Drosophila suzukii*, http://drosophila.jki.bund.de/index.php?menuid=3, abgerufen am 15.05.2018.

/3/ Seebacher, D. et al. (2017): Visual Analysis of Spatio-Temporal Event Predictions: Investigating the Spread Dynamics of Invasive Species. Symposium on Visualization in Data Science (VDS) at IEEE VIS 2017, Phoenix, https://bib.dbvis.de/uploadedFiles/seebacher.pdf, abgerufen am 15.05.2018.

/4/ BigGIS Demo Dokumentation (2018): http://biggis-project.eu/biggis-docs/demos/invasive-species/, abgerufen am 15.05.2018.