



Аграрен Университет – Пловдив

Факултет „Икономика“

Катедра „Математика и информатика“

НИКОЛА ЙОРДАНОВ ЦАЙКИН

**Анализ на многомерна база данни с
динамични редове за
растениевъдството в България**

АВТОРЕФЕРАТ

на дисертационен труд

за присъждане на образователна и научна степен „доктор“

област на висшето образование: **4.0 Природни науки, математика и
информатика**

професионално направление: **4.6 Информатика и компютърни науки**
научна специалност: „Информатика“

Пловдив 2019 г.

Дисертационният труд е структуриран в увод, три глави, заключения и препоръки, литература. Общият обем е 125 страници, от които 116 основен текст. Представени са 14 таблици и 37 фигури.

Дисертационният труд е обсъден и предложен за защита на разширено заседание на катедра „Математика и информатика“ при Аграрен университет – Пловдив, с Протокол № 12/19.04.2019 г.

Списъкът на авторските публикации се състои от 3 заглавия.

Защитата на дисертационния труд ще се състои на 12 юли 2019 г. в Катедра „Математика и информатика“, Аграрен университет – Пловдив на заседание на Специализирано научно жури, назначено от Ректора на Аграрен университет – Пловдив със Заповед в състав:

Проф. д-р Аврам Ескенази - рецензент;

Проф. д-р.Асен Рахнев;

Проф. д-р Никола Колев;

Проф. д-р Кольо Онков;

Доц. д-р.Тодорка Мокрева - рецензент.

Материалите по защитата са на разположение на интересуващите се в библиотеката на Аграрен Университет – Пловдив, бул. Менделеев № 12.

Съдържание

I. ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД.....	4
II. КРАТКО ИЗЛОЖЕНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД.....	7
Първа глава Съвременни тенденции при анализа на динамични редове за растениевъдна продукция и източници на данни.....	7
Съвременни тенденции в анализа на динамични редове за растениевъдна продукция.....	7
Източници на статистически данни за растениевъдната продукция в България и Европейския Съюз.	9
Втора глава Многомерна БД, методи и софтуер за управление и анализ.....	12
База данни Crop_ProductionBGEC с динамични редове, за производството и реколтирани площи от земеделски култури в страните от Европейския съюз .	12
Интегрирана софтуерна платформа за извличане и математико-статистически анализ на динамични редове.....	15
Обновяване на базата от данни Crop_ProductionBGEC	18
Софтуерно приложение SAVATAP.....	19
Трета глава Статистически анализи на данни от базата „Crop_ProductionBGEC”	23
Анализ на добива от слънчоглед за България по райони с използване на език за статистическа обработка R.	23
Анализ на данни с използване на пакета за статистическа обработка SPSS.	26
Анализ на динамични данни за суровина (грозде) и краен хранителен продукт (вино)	26
Сравнителен анализ на ключови култури пшеница, соя, ориз и царевица в България	28
III. ИЗВОДИ И ПРЕПОРЪКИ.....	31
IV. ПРИНОСИ НА ДИСЕРТАЦИОННИЯ ТРУД.....	33
V. НАУЧНИ ПУБЛИКАЦИИ.....	34
Използвана литература:.....	34
Abstract	Грешка! Показалецът не е дефиниран.

I. ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

В своето развитие съвременните бази данни са намерили широко приложение при съхранение и анализ на различни типове данни в множество области на човешката дейност. Интегрирането на данни и аналитични инструменти е мощен ресурс за всеки сегмент на управление, изследвания и икономическа дейност (Hearst, 1998). Базите данни се използват главно за съхранение и свързване на данни с оглед на две цели – информационно осигуряване на процеси и предоставяне на данни като основа при анализ за вземане на решения. Софтуерния пакет Microsoft Office осигурява на нискобюджетния потребител и на предприемача от малкия бизнес достатъчно достъпен инструмент за СУБД като Microsoft Access. В аграрната сфера такива потребители са научните работници, фермерите и търговците на селскостопанска продукция. Достъпът до статистически данни и инструменти за анализ (Alwan, 2000) предлага обективен поглед върху реалните процеси и тенденции и подпомага вземането на решения (Chaudhuri, et al. 2001).

България има капацитета да развие конкурентно земеделско производство и да заеме достойни позиции както на европейския, така и на международния пазар. Чрез осигуряването на прехрана и препитание, селското стопанство в България определя стратегическото си значение и допринася за националната сигурност и социална стабилност на страната. Растениевъдството е основен отрасъл на селското стопанство решавайки проблема с осигуряването на продукти с растителен произход. В световен мащаб са проведени множество научни изследвания за растениевъдното производство, които показват необходимостта от неговото увеличаване. Освен демографския фактор, други важни фактори за развитие на растениевъдството са специфичните промени в храненето на населението (напр. диетично хранене) и нарастващото използване на биогоривата. Растениевъдното производство има важно значение за социално-икономическото развитие на България. Растителните продукти представляват голяма част от храната на населението в страната. Затова е от значение да бъде

направена прецизна оценка на заетите площи с растителни култури и възможността за увеличаване площите на някои от тях.

Международни и правителствени институции съхраняват в бази данни (БД) огромни обеми от данни (Jensen, at al. 2010) от различни области – икономика, демография, промишленост, селско стопанство (вкл. за растениевъдното производство). Обикновено такива бази данни са представени в Web среда със сравнително удобен интерфейс за достъп до отделни съвкупности от данни. Само в редки случаи се предоставя online обработка (Javier, 2004), която от своя страна е елементарна и недостатъчна за по-задълбочени анализи и изводи. С нарастването на размерността и обема на тези бази данни, нарастват трудностите на крайния потребител при работа с нея.

Математико-статистическата обработка на данни е неотменна част от научната и мениджърска дейност в аграрния сектор. Един от обектите на анализ са динамични редове (time series) на показатели по години за относително дълги периоди от време. Източници на такива данни са базите данни, поддържани от институциите на отделните страни, Европейския съюз, FAO (Food and Agriculture Organization of the United Nations – Световна организация за храни и земеделие към ООН) и др. FAO събира и съхранява данни за култури, площи, количество продукция, търговия и други в многомерна бази данни както за отделни страни, така и обобщена информация за групи страни на континентално, регионално и политическо ниво на годишна база, а някои пазарни показатели и на месечна основа.

В България се отглежда асортимент от земеделски култури, съответстващ на климатичните условия, изисквания свързани с прехраната и националната сигурност, пазарни условия и др. Националният статистически институт (НСИ) и Министерството на земеделието и храните (МЗХ) събират, съхраняват и предоставят данни на годишна основа за реколтирани площи, производство, добив и други икономико-статистически показатели. Структурирането на данните в динамични редове и съхраняването им в електронни таблици е най-удобно за автоматизираната им обработка с оглед проследяването на тенденции, сравнителни анализи, сходство и различия в отглеждането на различни култури.

Публикуването в pdf и други неявно структурирани формати затруднява привеждането на данните във вид, съвместим със софтуер за анализ.

Обекта на дисертационната работа е да предложи интегрирано решение от инструменти за представяне и анализ на данни и информация, събрани в приложно програмен пакет, чиято цел е получаване на аналитични резултати от математико-статистическа обработка с прилежащото интерактивно графично и таблично представяне. Потребителите на такова решение са изследователи, учени и експерти, заинтересовани от закономерностите, заложи в динамичните редове (Glymour, 1997).

За създаване на приложно-програмния пакет са използвани следните **методи**: интегриране на софтуерни инструменти и многомерни бази данни, обектно-ориентирано програмиране за приложение с потребителски ориентиран интерфейс и използване на функции от езика R. За математико-статистически анализи върху динамични редове са приложени дисперсионен и клъстерен анализ и алгоритъм за изчисляване на екологична валентност.

Целта на дисертационната работа е да се създаде информационна среда, подпомагаща специалисти и експерти при изготвяне на аналитични оценки и вземане на адекватни решения за отрасъла растениевъдство. Тази информационна среда трябва да съдържа многомерна база данни с динамични редове за различни аспекти на растениевъдното производство в България и страни от Европейския съюз: териториални (пространствени), биологични и статистически. Към базата данни да има разработен софтуер за нейното обновяване и анализиране.

Класически подход за изграждане на информационна среда е създаване на интеграция между база данни и множество софтуерни приложения, отнасящи се до поддържане на базата, както и проектиране на удобен интерфейс за работа на потребителите с нея и вграждане на програми за математико-статистически анализ.

За постигането на целта следва да се решат следните **задачи**:

1. Проектиране и създаване на многомерна база данни с динамични редове за растениевъдното производство в България и страни от Европейския

съюз, като се обхванат пространствените, времеви, биологични и статистически аспекти на данните от информационните източници. Целесъобразно е да се оптимизира структурата на базата данни с оглед на бързи и леки като изчислително натоварване заявки. Освен това, да се разработи "потребителски ориентиран" софтуер за обновяване на данните на годишна основа, чрез интегриране на данни от големите многомерни институционални бази данни.

2. Създаване на софтуер за извличане на динамични редове от базата данни и тяхното управление чрез графичен потребителски интерфейс. Необходимо е тези софтуерни инструменти да са подходящи за по-широк кръг от потребители.

3. Разработване на софтуер за математико-статистически анализ на данните от динамичните редове, включително чрез използването на пакети приложни програми и модули от програмния език R.

4. Анализиране на динамични редове за растениевъдното производство на отделни култури с оглед на изследване на закономерности и откриване на тенденции.

II. КРАТКО ИЗЛОЖЕНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

Първа глава

Съвременни тенденции при анализа на динамични редове за растениевъдна продукция и източници на данни

Съвременни тенденции в анализа на динамични редове за растениевъдна продукция.

Основен източник на статистическа информация, организирана в динамични редове са институции на глобално и национално ниво, защото имат необходимите административни механизми и финансови ресурси да събират данни и да ги съхраняват в големи бази данни с online или друг електронен достъп. Най-голямата и достъпна база данни в областта на аграрната сфера и по-специално растениевъдното производство е FAOSTAT. В (Trnka, 2016) е проучен добива на пшеница и ечемик за 12 страни от Европа в зависимост от климатични показатели за периода от 1901г. до 2012г. резултатите показват, че

факторите на зависимост между климата и добивите са се променили значително. Някои от показателите са имали по-голям ефект върху добива след 1991г.

Изграждането на РС база данни за съхранение на определен профил на данни в информационен модел дава на потребителя мощен инструмент за анализ и вземане на решения (Onkov, 2016). Една база данни в прецизирана предметна област осигурява информационен фокус върху данни и софтуер за обработка и представяне на резултати (Gandhi, 2016)

Създаден е информационен модел – многомерна база данни с динамични редове за рибните ресурси на Гърция за периода от 1990 – 2015 г. и аналитичен софтуер (Tegos & Onkov, 2015). Анализът на данни показва устойчивост на общия улов след 2000г. В (Tegos, 2007) са представени SQL базирани инструменти за заявки към базата данни на гръцките рибни ресурси. Чрез експоненциални методи за изглаждане и прогнозиране са изчислени скорости на намаляване на общия улов на риба за изследвани периоди от време (Tegos, 2016). Получената информация за рибите в риск е важна за експлоатацията и защитата на рибните ресурси в Гърция.

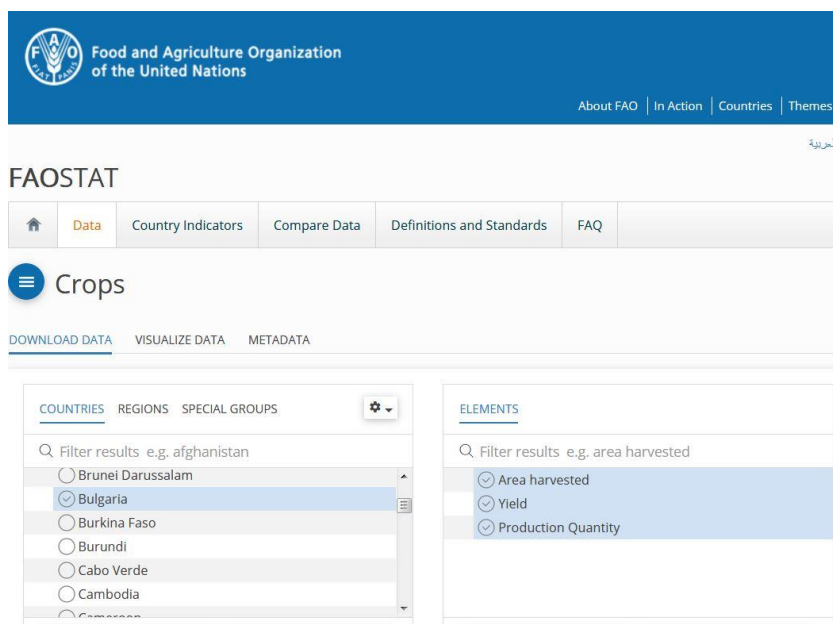
Онков (2011) разработва информационна среда и „data mining” алгоритми за йерархично структуриране и анализ на многомерни динамични редове с приложение в морския риболов в Гърция и растениевъдната продукция на балканските страни членки на Европейският Съюз.

Софтуерната система R е мощна среда за статистически изчисления, графична визуализация и програмиране, работеща под различни операционни системи: Unix, Linux, Windows и MacOS. Понастоящем системата се използва от десетки хиляди хора ежедневно за сериозен статистически анализ и научно моделиране. (Shimoda, 2018) изследва добива на картофи в Япония за периода от 1986г. до 2014г. В резултат на повишените температури през лятото от края на 2000-те и непроменените пролетни температури, както и промяната в сортовете са причинили добивите на картофи в източната част на Хокайдо да спаднат под тези на други страни, където добивите от картофи са високи.

Източници на статистически данни за растениевъдната продукция в България и Европейския Съюз.

Статистическите данни за растениевъдната продукция в България, Балканския регион и Европейския Съюз се извличат от Световната организация по храни и земеделие на обединените нации (FAO), Националния статистически институт на България (НСИ) и Министерство на земеделието и храните (МЗХ). Те се съхраняват в статистически годишници и в бази данни (БД) с голям обем. Данните съдържат информация за земеделска култура, страна, област и регион, като региона представлява съвкупност от страни. Съдържат още код на показателя на същността или обекта на данните и време. Тъй като в източниците се съдържа предимно годишна статистическа информация отчетният интервал за време е фиксиран на година.

Формирането на заявка се реализира чрез избор на страна, статистическа единица (площ, количество продукция и др.), земеделска култура и период от време (Фиг. 1.1.).



The screenshot displays the FAOSTAT website interface. At the top, the FAO logo and 'Food and Agriculture Organization of the United Nations' are visible. Below the navigation bar, the 'FAOSTAT' title is present. The 'Data' tab is selected, and the 'Crops' section is active. Under 'Crops', the 'DOWNLOAD DATA' link is highlighted. The 'COUNTRIES' tab is selected, showing a list of countries with 'Bulgaria' chosen. The 'ELEMENTS' tab is also selected, showing a list of elements with 'Area harvested', 'Yield', and 'Production Quantity' chosen.

Фиг. 1.1. Формиране на заявка за достъп и извличане на данни.

Базата данни на FAO съдържа статистически динамични (времеви) данни за производството на растителни култури на национално ниво, но липсват данни за райони и области на отделните страни. Това обикновено е приоритет на националните източници на информация. Такива данни за България се

предоставят от Националния Статистически Институт (НСИ) и Министерство на земеделието и храните на България.

Ключово значение за достъпа до динамични данни за растениевъдното производство по географски райони на България има Агростатистическия справочник на сайта на МЗХ. (<http://www.mzh.government.bg/>). Фигура 1.2 представя пример за достъп и извличане на данни от справочника във формати pdf и xls (a) и фрагмент от данни (b).

За потребителите на агростатистика

Анкетни

Анкета за потребителите на земеделска статистика - анкетна карта за изследване удовлетвореността на потребителите от предлаганите статистически информационни продукти и услуги от отдел Агостатистика към Главна дирекция Земеделие и регионална политика.

Документи

Справочник, календар и резултати

Агостатистически справочник 2000-2016

RAR файл, 6,1 MB, качен на 16.02.2016



(a)

Статистически райони	2016			
	Винени сортове Wine-grape varieties			Десертни сортове Table grapes varieties
	Червени сортове Red varieties	Бели сортове White varieties	Общо Total	
Реколтирани площи - хектари / Area under cultivation - hectares				
Общо за страната	22 911	11 631	34 542	2 009
Северозападен	2 413	727	3 140	56
Северен централен	862	700	1 562	85
Североизточен	683	1 890	2 573	131
Югоизточен	6 261	6 394	12 655	551
Югозападен	3 140	231	3 371	121
Южен централен	9 552	1 689	11 241	1 065

(б)

Фиг. 1.2. Статистически справочник (a), данни по райони за България (б)

Отчитането на статистическите данни за растениевъдното производство на България е в 6 планови икономически райони, определени по географски принцип.

Настоящият дисертационен труд е необходим защото предлага цялостно решение по отношение на структурирането и обработката на динамични редове за растениевъдството в България. Предимствата на това решение (БД и приложен софтуер) могат да се обобщят по следния начин:

1) **Интеграция на данни.** Многомерната БД с динамични редове интегрира данни, извлечени от **3 различни източници**: FAO, НСИ, Министерство на земеделието на България. Факт е, че FAO предоставя данни единствено на национално ниво, докато НСИ и Министерство на земеделието осигуряват данните на ниво райони на България. Добавянето на динамични редове за растениевъдството в Гърция и Румъния на национално ниво в същата база данни ще даде възможност за сравнителни анализи;

2) **Управление и работа с данни.** При извличане на данни от Web БД потребителят често е принуден да прави последователност от маркирания и избор на статистически показатели, на обекти – култури, площи, количество производство, на период (и) от време и др., след което тяхното съхраняване във формат, подходящ за пряко прилагане на приложен софтуер с графичен потребителски интерфейс (SPSS, Excel, Statistica ...). PC базата данни е значително по-малка като обем, по-лесно обозрима и достъпна. Нещо повече, към нея ще се изгради интерфейс, който да направи достъпа до данните и избора на метод за обработка по-лесен и по-бърз.

3) **Математико статистическа обработка на данни.** Система Excel като част от Microsoft Office пакета е с много ограничени възможности. SPSS (Statistical Package for the Social Sciences) е **скъп софтуер**, който много изследователи, учени и преподаватели не могат да си позволят. В същото време интегрираната софтуерна среда на езика R е **безплатна (open source software)** и предлага множество модули и процедури

Използване на езика R. Необходимо е да се спомене това, че програмирането на R може да осигури предимство във възможността да се постигне по-висока степен на автоматизираност на процеса на обработка на данните, тъй като с една програма могат да се извикват последователно желани

процедури за обработка на избрани съвкупности от данни, включително вариации за периода на изследване. Това предимство, разбира се, касае предимно работата на опитен софтуерен специалист при обработка на по-мощни обеми от данни. Все пак, заслужава да се подчертае, че нерядко в магистърските и докторантските програми в областта на природните науки в САЩ, Англия, Швейцария, Дания, в някои програми и в България, се изучава езика R заради специфичните функционални възможности и за реализирането на симулационни експерименти. Това означава, че дори и в редки случаи изследователи от България: биолози, физици, еколози, биха се възползвали от софтуерни техники в настоящия дисертационен труд.

Втора глава

Многомерна БД, методи и софтуер за управление и анализ.

База данни Crop_ProductionBGEC с динамични редове, за производството и реколтирани площи от земеделски култури в страните от Европейския съюз

Базата е реализирана чрез Microsoft Access и съдържа информация за продукцията и засетите площи в България и ЕС за 82 земеделски култури за периода 1961-2017 година. Проектирана е като многомерна (Jensen, at al. 2010) и е развита в научни проекти в обучаващата катедра с участието на докторанта (Научен проект 09.09, 2009), (Научен проект 07.11, 2011). Първичните данни, необходими за изграждане на такава база данни, са взети от FAOSTAT, а данните по статистически райони за България са извлечени от статистическите годишници на НСИ. Извлечените съвкупности са многомерни динамични редове, чийто формат е показан на фиг. 2.1. Затова базата е проектирана като многомерен хиперкуб (Agrawal, at al. 1997).

Този хиперкуб от данни е с 4 дименсии: регион, показател, земеделска култура, време.

Земеделските култури, съхранявани в **Crop_ProductionBGEC**, са класифицирани в четири йерархични нива според Indicative Crop Classification на FAO за 2010 година:

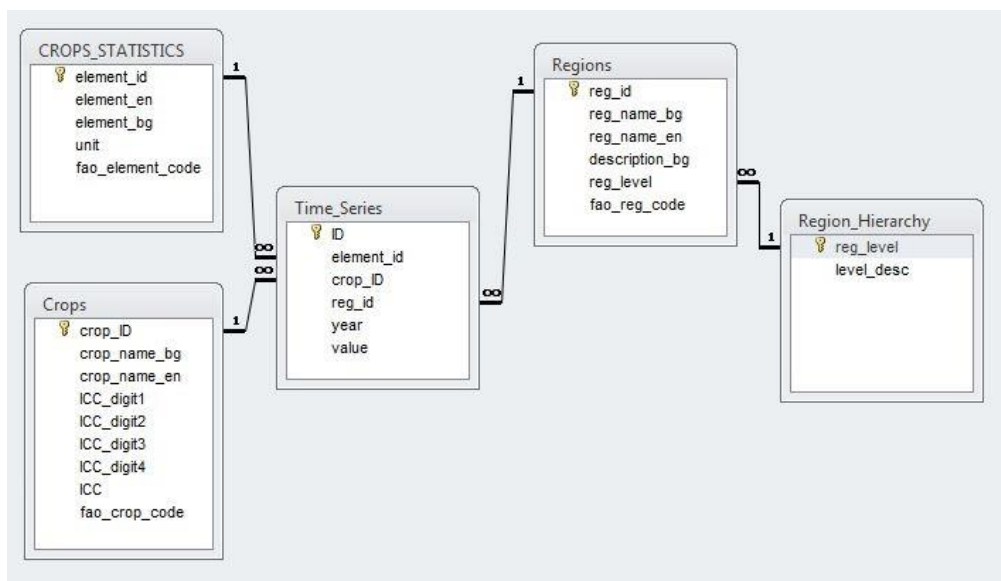
- група (group)- обозначава се с еднозначен цифров код
- клас (class)- обозначава се с двузначен цифров код

- подклас (subclass)- обозначава се с тризначен цифров код
- ред (order)- обозначава се с четиризначен цифров код

Всяка земеделска култура има ICC код, който съвпада с кода на йерархичното ниво, към което принадлежи културата. Това осигурява на потребителите допълнителни данни за статистически изследвания и анализи, които пряко не са достъпни от базите на FAO .

Базата данни е проектирана като релационна на основата на ER-модел на данни (Entity-Relation Model). Обектите от многомерния куб и отношенията между тях са систематизирани по този модел. Същността на организацията на данните е динамичен ред. Фокусът на информационното съдържание на динамичния ред е стойност, изменяща се във времето.

С оглед на това архитектурата на релационните връзки е тип звезда (Фиг. 2.1.). Състои се от основна, централна таблица и три разклонения от таблици, съхраняващи характеристиките на дименсиите на данните. Централната таблица *time_series* съхранява основните данни от времеви редове. Всеки запис от тези редове съдържа стойността, която е фокусът на информационната система и вторичните ключове на четирите дименсии. Тъй като една от дименсиите е време тя няма нужда от по-подробно охарактеризиране освен стойността на годината и тя се представя само с полето “year” на основната централна таблица.

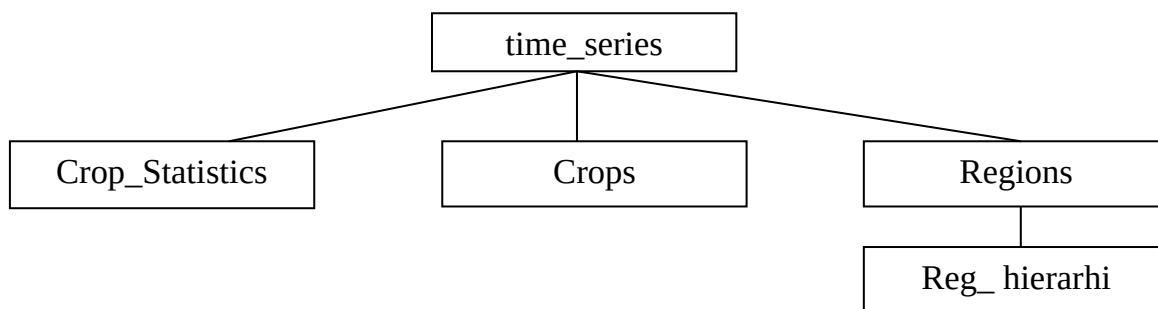


Фиг. 2.1. Релационен модел на Crop_ProductionBGEC

Природата на описанието на атрибутите на трите дименсии за култура, елемент и страна е текстов етикет. Този текст се съхранява в атрибутните таблици и чрез първичен ключ се свързва с централната таблица чрез релация „едно към много”. В базата от данни Crop_ProductionBGEC всички таблици са в трета нормална форма, което гарантира, че схемата на БД е в трета нормална форма и е свободна от аномалии (Elmasri & Navathe, 2010).

Използван е ICC – код на културата според класификацията на ICC (Indicative Crop Classification). Това е четирицифрен код, където отделните разреди представят не само идентификация на отделни култури, но и чрез по висшите разреди се кодират групи от култури. Това позволява да се правят извадки от данни по групиращ критерий.

Концептуалното ниво описва организацията на данните, включваща състава, типовете данни и логическите връзки между тях. То може да се представи чрез йерархичен модел на данните (Feng et al. 2010). В основата си той има корен, където е фокусът на информационното съдържание и най-голяма динамика на входни и изходни информационни потоци. В случая коренът е обекта time_series от фиг. 2.2, с прилежащите му атрибути. Възлите на второ ниво са обектите Crops, Crop_Statistics и Regions. На трето ниво е взела Reg_hierarhi, като разклонение на възела Regions.



Фиг. 2.2. Йерархичен модел на данните.

Свързващите дъги между възлите показват логическите връзки между обектите и следват релационните връзки от релационния модел (Фиг. 2.1).

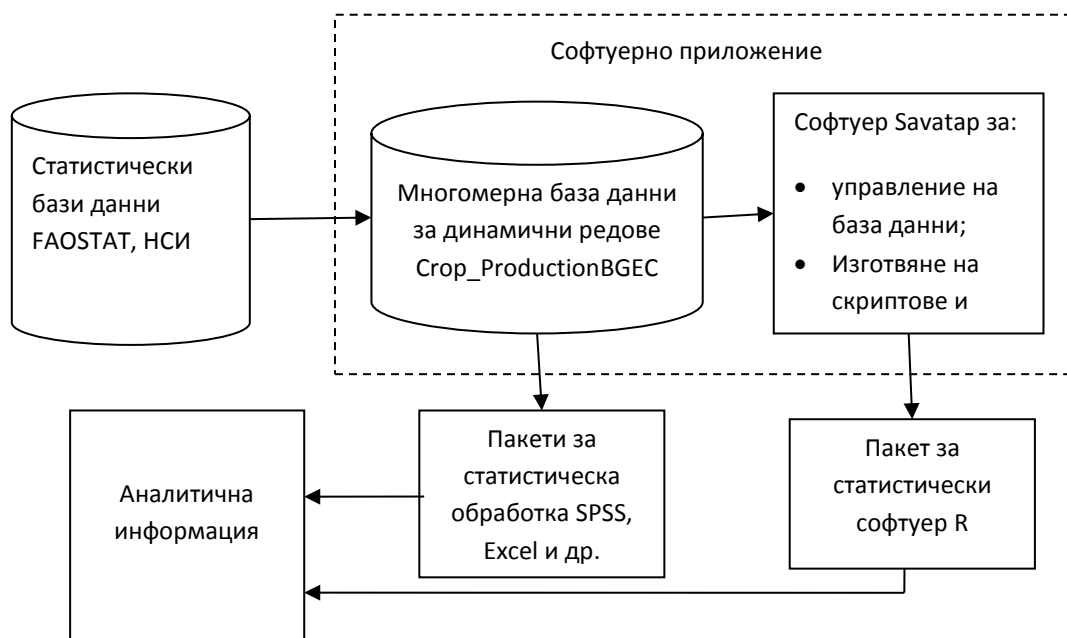
Описаният модел с четирите дименсии е отворен и има структурен капацитет да съхранява данни и от други отрасли, свързани с аграрната сфера. Дименсиите идентификация, статистически показател, пространствена

характеристика и време формират четири-векторен ортогонален модел на данни с потенциал да съхрани информация за обекти от широк кръг предметни области. Данните от голямата статистическа база за другите отрасли се избират от web-интерфейса на страницата от четири списъка с атрибути по същия модел като на фигура 1.1. и които съответстват на хипердимензионния куб на базата данни Crop_ProductionBGEC.

Интегрирана софтуерна платформа за извличане и математико-статистически анализ на динамични редове.

Интегрирането на решения винаги има мултиплицираща роля за ефекта от отделните решения, използвани самостоятелно (Lenzerini, 2002). Създаването на интегрирана платформа от свързани решения (Илиева et al. 2010) дава качествено нови инструменти и понякога единствено възможни такива за работа в предметната област на потребителя – изследовател и експерт

В настоящата работа е изградена софтуерна платформа (Онков, 2011), интегрираща данни от големи бази данни, статистически софтуер и софтуерно решение за автоматизиране на информационния поток от институционалните източници до резултати от статистически анализ.



Фиг. 2.3. Принципна схема на софтуерна платформа за статистически анализ.

На Фиг. 2.3 е представена принципна схема на софтуерна платформа за статистически анализ. Едната възможност е извлечени динамични редове, през Clipboard или чрез експорт в .csv-файл, да се обработват от програмни статистически пакети като SPSS или чрез функции на Excel от пакета Microsoft Office. На фиг. 2.4 е показан интерфейса на СУБД Access за йерархично преглеждане на данните в базата по страни и региони.

reg_level	level_desc
1 View data by country	
reg_id	reg_name_bg
BG	България
GR	Гърция
RO	Румъния
crop_ID	crop_name_bg
1	Бадеми с черупка
2	Анасон, звезден анасон, коп
3	Ябълки
4	Кайсии и зарзали
5	Аспержи
6	Ечемик
year	value
#1961	284000
#1962	250475
#1963	224100
#1964	195500
#1965	232771
#1966	245700

Фиг. 2.4. Йерархично преглеждане на данни от базата

Crop_ProductionBGEC

Другият способ на обработка е чрез функциите на езика R. В този случай потребителят е подпомогнат от софтуерното приложение Savatar, което извлича динамичните редове по избрани в потребителския интерфейс критерии и съставя програмен скрипт за извикване на програмната среда на R. Основните функции на приложението са управление на многомерна оперативна база данни и изготвяне на входни файлове за данни и програмен код за езика R.

На таблица 2.1 са обобщени модулите на софтуерната платформа и тяхната принципна функционалност.

- Чрез софтуерното приложение Savatar потребителят работи с компактна платформа с висока степен на автоматизация и в среда на приятелски интерфейс. Достъпът до динамичните редове е улеснен и базата данни е скрита зад

интерфейса. Изборът на статистически метод също е улеснен, като алгоритъмът на метода е автоматизиран чрез функциите на езика R. Чрез софтуерното приложение Savatar потребителят използва интегрираните функции на езика R безплатно заради отворения характер на R софтуера. Това е важно предимство с оглед избягването на платени приложни програмни пакети.

Таблица 2.1. Модули на софтуерна платформа за статистически анализ

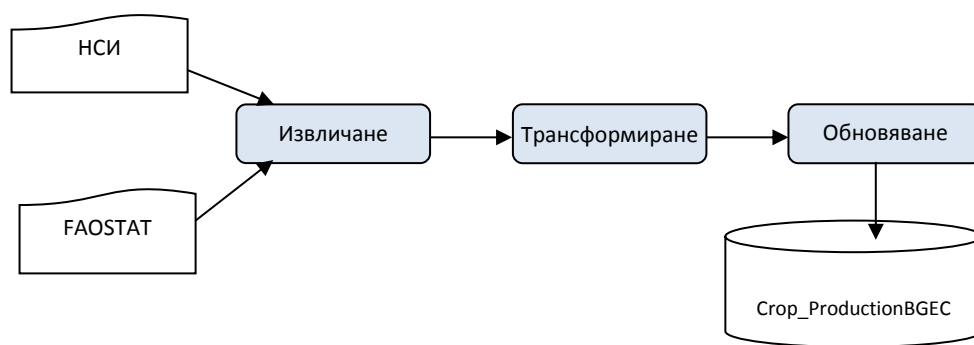
Софтуерни модули	Наименование на програмните среди	Описание
Статистическа обработка на извлечени динамични редове	SPSS, Statistika, Access, Excel	Статистически анализи като дисперсионен, клъстерен, тренд и други методи за анализ на специализиран софтуер.
Автоматизирана статистическа обработка чрез функции на езика R	R – софтуер	Програмиране на R-функции и зареждане на входни данни от .csv файлове.
База данни за динамични редове	Crop_Production BGEC	Съхранение, обновяване, разглеждане и извличане на динамични редове. Експортиране на данни през Clipboard и в буферни .csv файлове.
Софтуерно приложение с потребителски интерфейс	Savatar	Автоматизирано разглеждане и извличане на динамични редове и статистическа обработка Автоматизирано създаване и опериране с .csv и .r файлове.

•Работата с Access, SPSS и други статистически програми изисква от потребителя определени познания и умения за тях. Необходимо е той да познава организацията на базата данни и детайлите на статистическите методи, за да планира етапите и настройките на избрания анализ.

Базата данни се нуждае от обновяване чрез частично автоматизирани операции на извличане на данни от базите данни на FAOSTAT и национални институции, които изискват знания и умения. Тези операции се извършват на периоди и са част от поддръжката на базата данни Crop_ProductionBGEC. Чрез потребителският интерфейс на софтуерното приложение Savatar за избрани динамични редове се избягват повтарящи се операции на потребителя в самата база данни.

Обновяване на базата от данни Crop_ProductionBGEC

Изграден е алгоритъм за автоматично обновяване на базата, протичащ в три основни стъпки, показани на фиг. 2.5. С участието на ас. д-р Дияна Стоянова е разработено специално приложение, което се изпълнява веднъж годишно, след обновяването на базата на FAOSTAT и дава възможност за автоматична актуализация на базата.



Фиг. 2.5 Основни стъпки при обновяването на базата от данни Crop_ProductionBGEC

1. **Извличане на данни за нов период от време.** Данните, необходими за обновяването на **Crop_ProductionBGEC** се извличат от FAOSTAT, а данните по статистически райони – от статистическите годишници на НСИ. Тези годишници са достъпни на хартиен носител или в pdf формат, което не позволява директното въвеждане на информацията в базата. От сайта на FAO (<http://www.fao.org/faostat/en/#data/QC>) автоматично се сваля и разархивира файла **Production_Crops_E_All_Data_(Normalized).csv**, съдържащ данните за реколтирани площи и производство от всички земеделските култури, за всичките 245 страни, за периода 1961-2016 година. Файлът е в csv формат, което позволява директното му импортиране в таблица на Access.
2. **Трансформиране.** На втората стъпка чрез заявки извлечените данни се преобразуват в подходящ за Crop_ProductionBGEC формат и се изтриват излишните данни.
3. **Обновяване.** Последният етап включва записване на трансформираните данни в таблицата, съхраняваща динамичните редове.

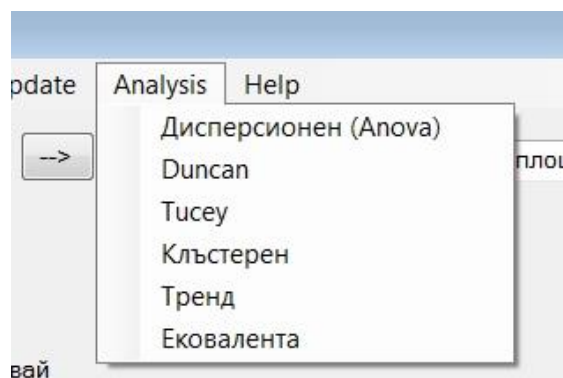
Можем да направим извод, че алгоритъмът осигурява дълъг технологичен живот на базата данни и интегрираната среда.

Софтуерно приложение SAVATAP.

Разработеното софтуерно приложение Savatar е предназначено за извличането и обработката на бази данни и динамични редове, касаещи растениевъдната продукция в България и останалите балкански страни от Европейския съюз. Автоматизира се процеса на статистическа обработка от извличането на данни от web-базираните бази чрез автоматичното обновяване до извеждането на резултати от избран метод. Приложението предоставя удобен потребителски интерфейс,

Приложението за подпомагане на извличането и анализа на динамичните редове от базата данни се състои от два компонента: desktop приложението **SAVATAP** (Software Application for viewing and analysis of Time series in Agricultural Production) и инсталирана програмна среда на езика R. Те са опосредствани от системата за управление на файлове на операционната система. То е предназначено да улесни извличането на динамични редове от неспециалисти по отношение на информационните технологии и без познания за работа с бази данни.

Приложението поддържа опашка от извлечени динамични редове, за да осигури възможност за сравнителни статистически анализи. Има инструменти за управление на последователността и маркиране на заредените редове в опашката.

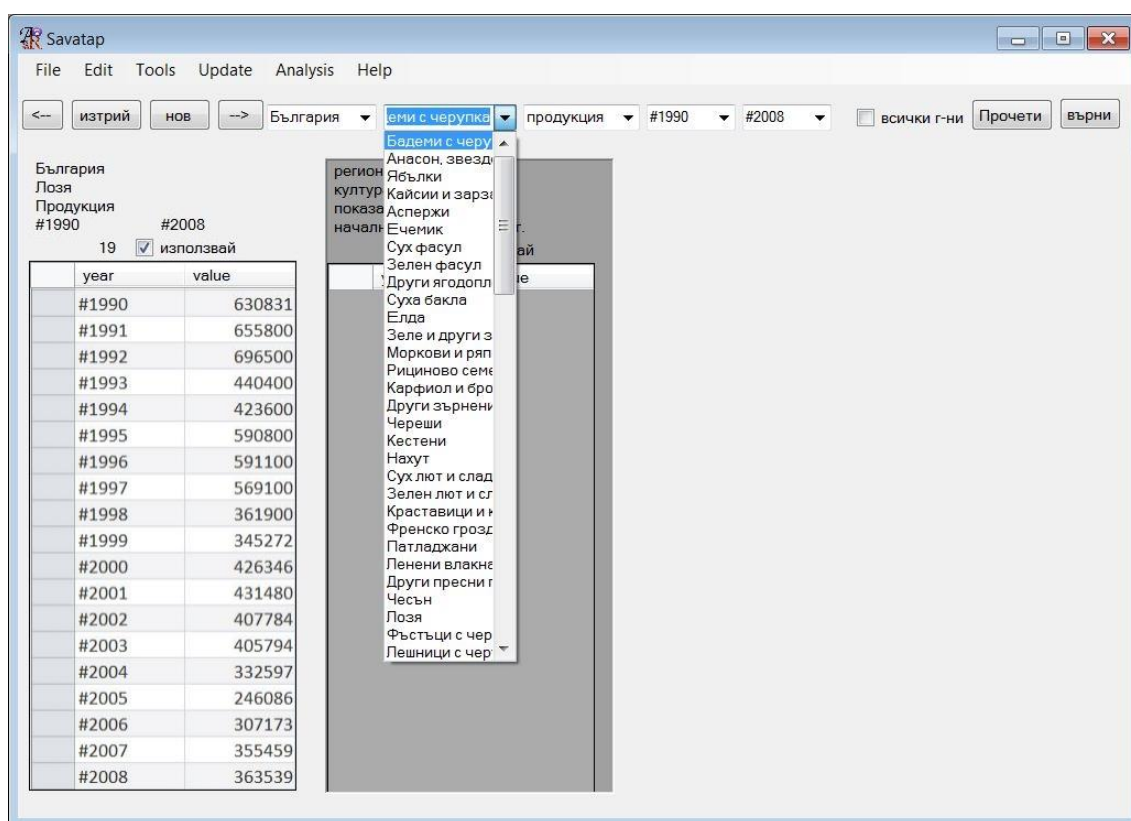


Фиг. 2.6. Избор на статистически метод

Потребителят може да избира статистически метод (Фиг. 2.6) за анализ и въз основа на този избор приложението се създава файл със скрипт и данни като аргумент за програмната среда на R (Braun & Murdoch 2007).

Интерфейса (Фиг. 2.7) на приложението осъществява извличането и разглеждането на динамичните редове от базата данни и избора на статистически метод за анализ. Оформени са няколко обособени процеса на работа на потребителя:

- Избират се критерии от падащи списъци според четирите дименсии на информационния хиперкуб: страна (регион), селскостопанска култура, статистически показател и година. Освен контроли за избор е необходим и команден контрол за потвърждение на избора на критерии и изпълнение на SQL-заявка за извличане на данните.



Фиг. 2.7. Интерфейс на приложението Savatap.

Извлечените динамични редове се разполагат в динамична опашка от съставни контроли. За управлението на тази опашка е добавен команден интерфейс за добавяне, изтриване, разместване и избор на активен контрол.

- Избор на статистически анализ се осъществява от раздел в главното меню като команден контрол. Необходима е проверка за съвместимостта на избраните динамични редове, като коректни аргументи на избрания статистически метод.
- В резултат на изпълнението на скрипт с функции на езика R се извежда резултат, който е част от програмната среда на R.

Потребителският интерфейс има статична и динамична част. Статичната част са контролите извън панела с динамичната опашка от съставни контроли. Това включва: падащите списъци, бутона за извличане на данни, менюто за избор на статистически метод и бутоните за навигиране в опашката. Падащите списъци са директно свързани със съответните таблици и полета в базата данни чрез настройките на техните свойства. Бутоната "Зареждане" стартира SQL-заявка за извличане на данни, чийто резултат се зарежда в активния стандартен табличен контрол от тип `DataGridView`.

Динамичната част се явява панелът с динамична опашка от съставни контроли, визуализиращи извлечените динамични редове. Той е от стандартен за C# тип контейнер (`FlowLayoutPanel`) и при нужда автоматично осигурява контрола скролер по хоризонтала. Съставните контроли се добавят в неограничен брой в опашката. Съставният контрол е проектиран като отделен клас `DinPan` и съдържа табличен контрол от тип `DataGridView` и етикети, показващи избраните атрибути на извлечените данни в четирите дименсии. Този клас е проектиран на основата на стандартен за C# родителски клас `UserControl` и също е контейнер, управляващ динамична опашка от контроли. Класът `UserControl` предоставя интерфейсно пространство от тип `Form`, в което проектантът добавя необходимия набор от контроли.

След избора на статистически метод от менюто и проверката за коректни аргументи се пристъпва към подготовка за извикване на програмната среда на R (Ripley & Murdoch 2009). Тази подготовка се извършва в две части. Едната част е изготвяне на скриптов .R-файл (Фиг. 2.8), който съдържа програмните инструкции за избрания метод и зареждане на данните от създадените .csv-файлове. Другата част е създаване на един или няколко файла в .csv формат, в

които се зареждат данните от заредените в динамичните панели динамични редове. Форматът .csv позволява данни в повече от две полета, където се добавят при необходимост кодове за статистически фактори, отразяващи избраната от потребителя последователност на заредените динамични редове, както и друга необходима информация. В табличния контрол от тип DataGridView, който е част от динамичния панел от тип DinPan, се зареждат само двете полета за година и стойност от извлечения динамичен ред. При създаване на скриптовия .R-файл и .csv-файловете за данни тази информация се използва.

```
require(xlsx)
require(csv)
require(RcmdrMisc)

Dataset <- read.csv("C:/Users/nikola/Desktop/savatempl.csv",1)
library(mvtnorm, pos=14)
library(survival, pos=14)
library(MASS, pos=14)
library(TH.data, pos=14)
library(multcomp, pos=14)
library(abind, pos=19)
sink("C:/Users/nikola/Desktop/analysis-output1.txt")
cat("Start of analysis...\n")

AnovaModel.1 <- aov(value ~ reg_id, data=Dataset)
summary(AnovaModel.1)
with(Dataset, numSummary(value, groups=reg_id, statistics=c("mean", "sd")))
local({
  .Pairs <- glht(AnovaModel.1, linfct = mcp(reg_id = "Tukey"))
  print(summary(.Pairs)) # pairwise tests
  print(confint(.Pairs)) # confidence intervals
  print(cld(.Pairs)) # compact letter display
  old.oma <- par(oma=c(0,5,0,0))
  plot(confint(.Pairs))
  par(old.oma)
})
sink('analysis-output1.txt', append=TRUE)
cat("Some more stuff here...\n")
sink()
```

Фиг. 2.8. Скриптов файл на езика R

Файловете със скрипта и данните се записват във файловата система и се предават като аргументи при извикването от приложението на интерпретатора на програмната среда на R (Илиева et al. 2010). Статистическият анализ и визуализирането на резултатите се извършва чрез библиотечните функции на R, според програмният код в скриптовия .R-файл. Генерирането на скриптовите .R-файлове е разработено с участието на ас. д-р Дияна Стоянова.

Трета глава

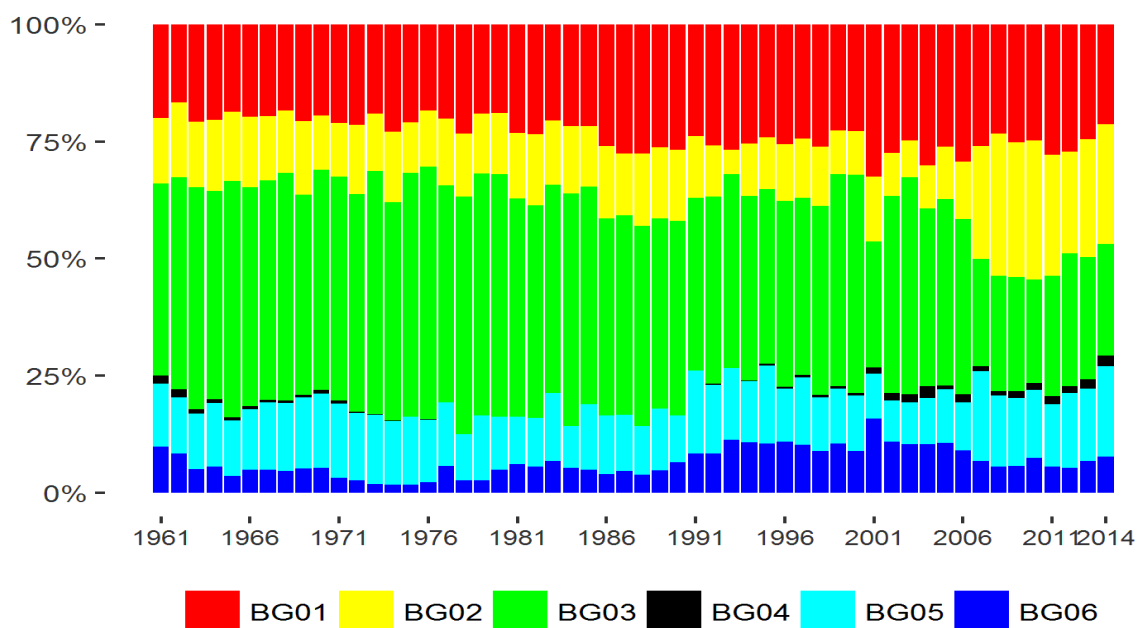
Статистически анализи на данни от базата „Crop_ProductionBGEC”

Анализ на добива от слънчоглед за България по райони с използване на език за статистическа обработка R.

Разработеният софтуер е използван за изследване на продукцията и средните добиви от слънчоглед в България по статистически райони за периода 1961-2014 г. (Stoyanova, 2017). Шестте статистически района от NUTS-2 (Северен централен (BG01), Северозападен (BG02), Североизточен (BG03), Югозападен (BG04), Югоизточен (BG05) и Южен централен (BG06)), са въведени официално през 2000 година.

- За периода 2000-2014 година данните са взети директно от статистическите годишници на НСИ

На фигура 3.1. са представени процентните дялове на всеки район от общата продукция от слънчоглед за страната.

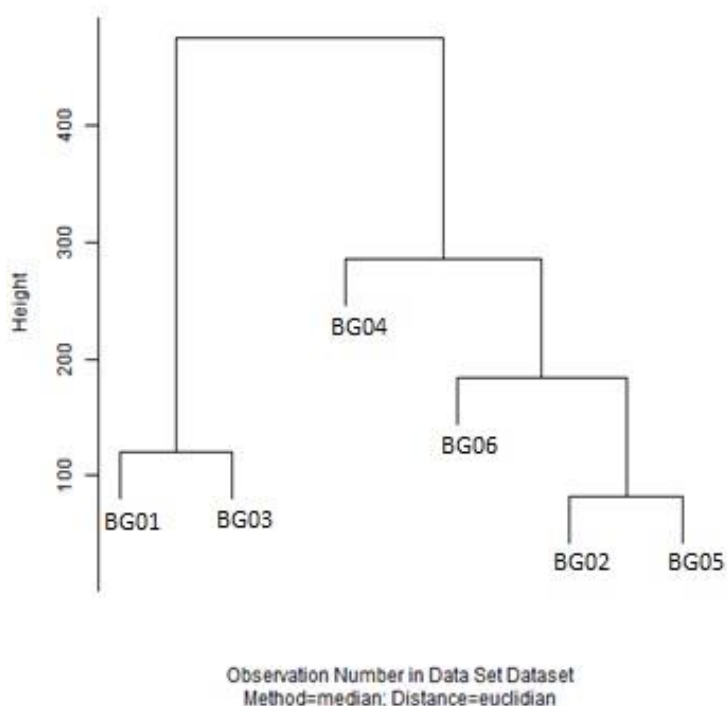


Фиг. 3.1. Разпределение на производството на слънчоглед по райони

За периода 1961 - 2006 година с най-голям дял е Североизточният район (BG03) – средно 41 %. От 2007 до 2014 той дели първото място с още два района - Северен централен (BG01) и Северозападен (BG02), като процентните им

стойности са почти равни. С най-малък относителен дял за целия изследван период от време е Югозападния район (BG04) - средно около 1%.

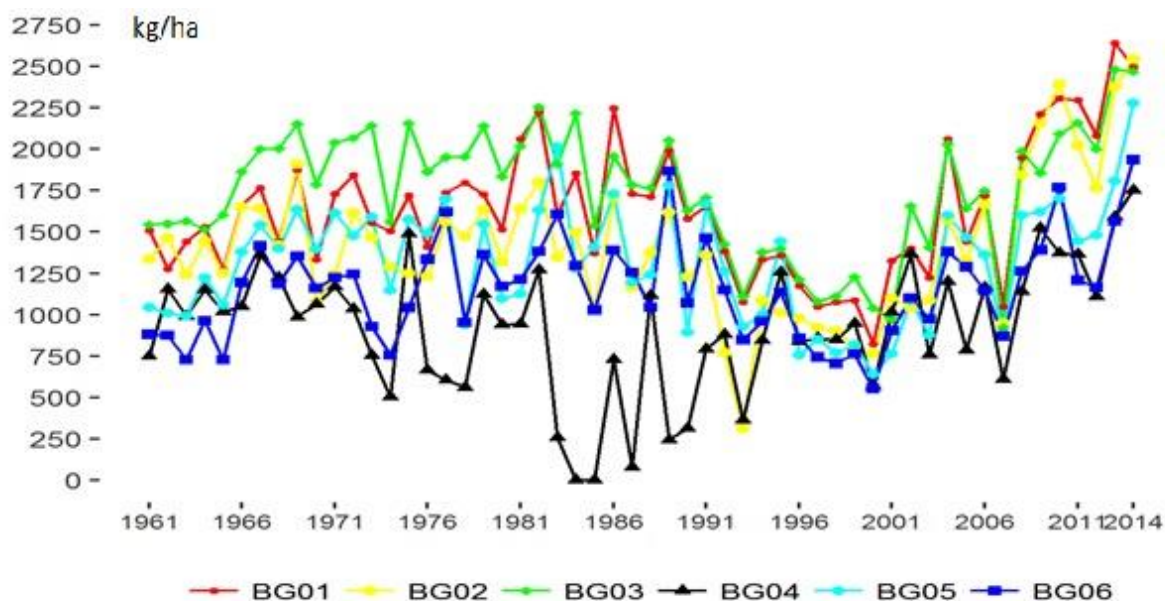
На фиг. 3.2. е отразено групирането в клъстери на статистическите райони в зависимост от степента на близост между стойностите на добива на слънчоглед за периода 1961-2014 г. Според дендрограмата статистическите райони по добив на слънчоглед формират два клъстера. Райони BG02 и BG05 формират един клъстер, които са с висока степен на сходство в стойностите на средните добиви. На разстояние 200 евклидови единици към клъстера се присъединява район BG06 и на приблизително 300 евклидови единици се присъединява район BG04. Райони BG01 и BG03 имат висока степен на сходство и се обединяват в самостоятелен клъстер.



Фиг. 3.2. Визуализация на клъстеризацията на статистическите райони по производство на слънчоглед

На Фиг. 3.3. са отразени емпиричните линии на динамичните редове на средните добиви от слънчоглед по статистически райони. Те показват основното направление в развитието на изследвания процес. Целият период на изследване може да се интерпретира, като се раздели на три периода в зависимост от

тенденцията в изменение на добива от слънчоглед. Трендовият модел за добива от слънчоглед за трите периода ще има различен аналитичен вид.



Фиг. 3.3. Тенденции в добива на слънчоглед по статистически райони

Средните добиви от слънчоглед за Югозападен район (BG04) през последните няколко години от изследвания период рязко намалява, доближавайки се до 0. Средният добив от слънчоглед в България през първия период (1961 - 1989), е 1400 kg/ha. В началото на периода производството на слънчоглед за всички райони, с изключение на Югозападен е умерено стабилно със слаби спадове и пикове.

През втория период (1990 - 2000) във всички райони се наблюдават слаби падове и пикове с близки по стойност средни добиви и с тенденция към намаляване. Средният добив на слънчоглед в България през този период е най-нисък – 1028 kg/ha.

През последния период (2001 - 2014) производството на слънчоглед в България е с тенденция на нарастване. Средният добив на слънчоглед в България през този период е най-висок – 1549 kg/ha.

Комплексната оценка на очертаните тенденции позволява да се направят следните обобщения:

- Най-високи средни добиви от слънчоглед са установени в Северен централен и Североизточен район, но варирането е голямо за целия изследван

период. Най-устойчиво е производството на слънчоглед в Южен централен район.

- След 2001 година средния добив от слънчоглед за всички статистически райони е с нарастваща тенденция и с най-високи отчетени стойности за целия изследван период от време.

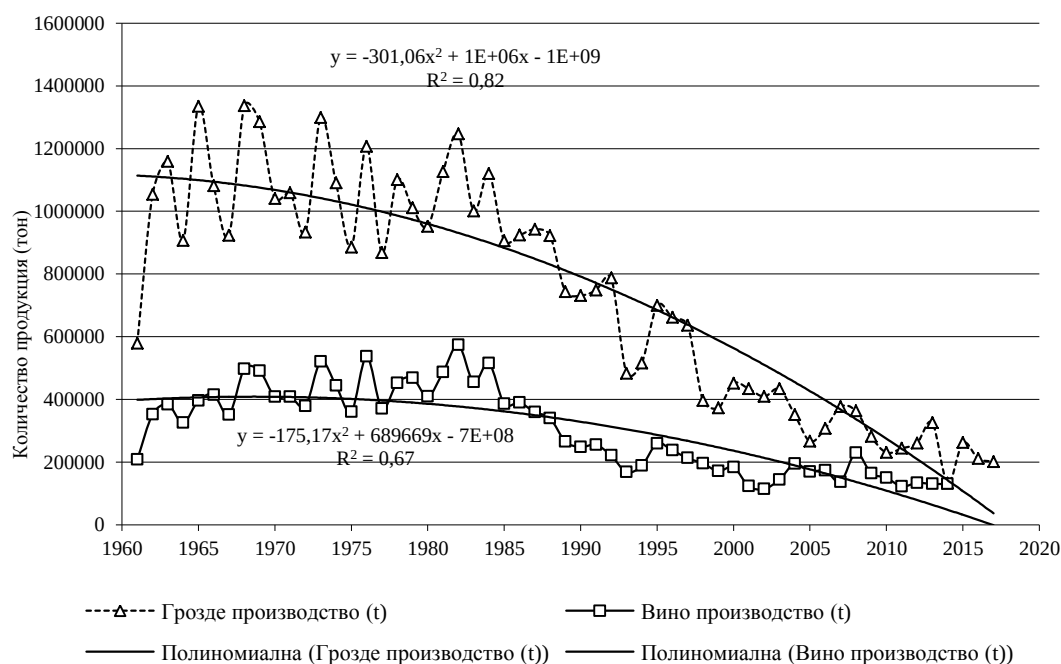
Анализ на данни с използване на пакета за статистическа обработка SPSS.

Анализ на динамични данни за суровина (грозде) и краен хранителен продукт (вино)

Анализ на динамични данни за производството на грозде и вино.

На фигура 3.4. се проследява динамиката на производството на грозде и вино в България. Трендовите функции на производството на грозде и вино са представена чрез уравнение от втора степен. Фигурата съдържа едновременно емпиричните и изгладените стойности на производството на грозде и вино за периода 1961 г.-2015 г. Установен е висок коефициент на корелация ($r = 0,90$) и детерминация ($R^2 = 0,82$) между значенията на производството на грозде и времето. Описаният модел е статистически значим и 82% от изменението на производството на грозде се обяснява с дисперсията на фактор време. Между значенията на производството на вино и времето е отчетена висока корелация ($r = 0,82$) с коефициента на детерминация $R^2 = 0,67$. Трендовата функция описваща производството на вино е статистически значима и 67% от изменението на производството на вино се обяснява с дисперсията на фактор време.

От 60 до 85 година лозарството е структуроопределящ отрасъл и има най-голям дял в селското стопанство.



Фигура 3.4. Емпирични и изгладени стойности на производството на грозде и вино в България

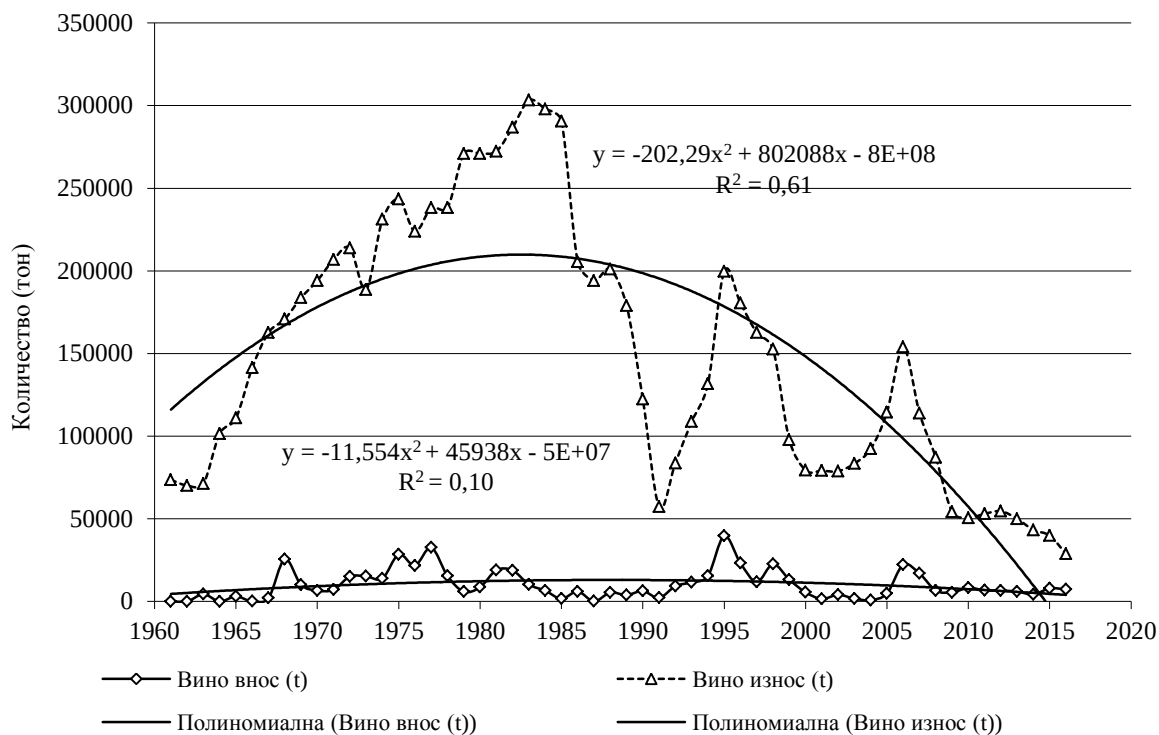
След 1985 година за България са години на преход от централно-планирана към пазарна икономика които поставят сектора, както и цялото ни селско стопанство, пред големи трудности и изпитания. Този период се характеризира със слаби спадове и пикове и с ясно изразена тенденция на намаляване на производството. От 1999 година ситуацията в лозарския сектор се променя. Емпиричните линии се характеризират с много по-слаби спадове и пикове. Причина за това са присъединителните земеделски фондове и европейските програми като САПАРД и ФАР и възникването на добре проектирани и оборудвани нови винарни.

Анализ на динамични данни за износ и внос на вино

Периодът от 1961 г. до 1985 г. се характеризира с много слаби падове и пикове и нарастваща тенденция в износа на вино. През 1966 г. България е шестият най-голям износител на вино в света.

Линията описваща износа на вино след 1985 година се характерно с големи падове и пикове и силно намаляваща тенденцията (фиг. 3.5.). Причина за това е проведената през 1989 година либерализация в този отрасъл, която довежда до намаляване на износа в пъти, а не в проценти. Установена е

значителна корелация ($r = 0,78$) и детерминация ($R^2 = 0,61$) между значенията на износа на вино и времето. Трендовата функция описваща износа на вино е статистически значима и 61% от изменението на износа на вино се обяснява с дисперсията на фактор време.



Фигура 3.5. Емпирични и изгладени стойности на износа и вноса на вино

След 2000 година тенденцията на износа на вино започва на се стабилизира благодарение на предприсъединителните земеделски фондове и европейските програми САПАРТ и ФАР. За периода на изследване около 51% от виното, което се произвежда в България се изнася.

Сравнителен анализ на ключови култури пшеница, соя, ориз и царевица в България

В този раздел е направен сравнителен анализ на производството на ключовите селскостопански култури пшеница, соя, ориз и царевица в България (Onkov, 2017).

Таблица 3.1 представя резултатите от статистическата обработка на данните за произведените количества продукция от изследваните култури. За периода 2009-2015 г. Най-голям относителен прираст на произведеното количество, спрямо предходния период 2002-2008г. се наблюдава при ориза –

119.6%. При соята е установен отрицателен прираст (-30.8%). Високите стойности на вариационния коефициент (136.6%) при соята показват, че добивите от тази култура са по-вариабилни, по-податливи на изменение в сравнение с добивите от другите култури. Най-слабо вариабилни са добивите от пшеница, с отчетен коефициентът на вариация – 24.9% и може да се приеме, че от изследваните култури, пшеницата е с най-устойчиви добиви през периода на изследването.

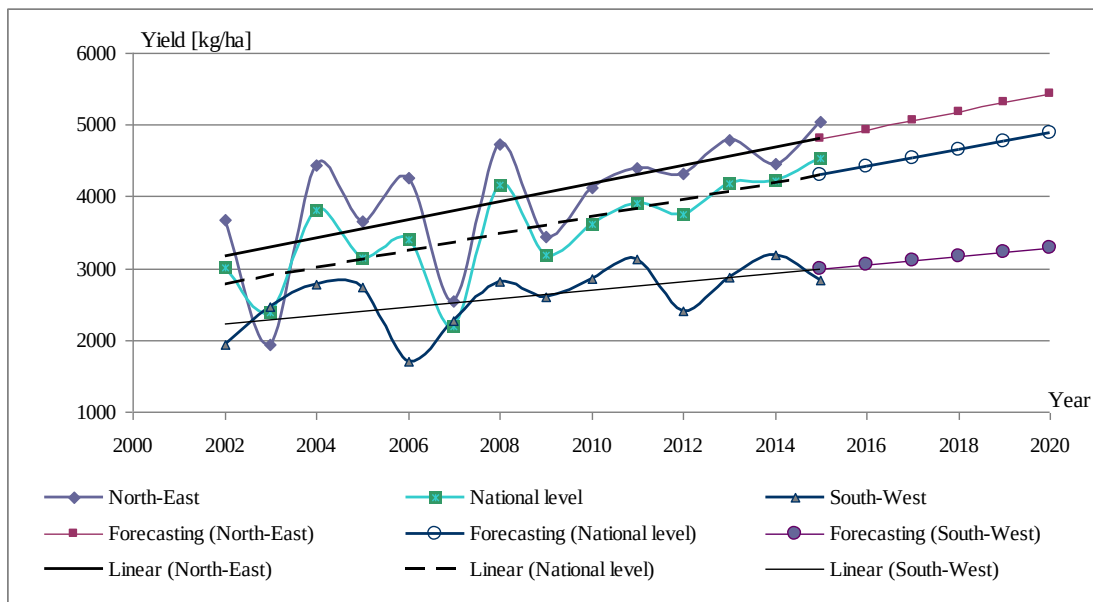
Таблица 3.1. Резултати от статистическата обработка на произведеното количество (t) от изследваните култури

Култура	Средна стойност 2002-2008г	Средна стойност 2009-2015г	Разлика (%)	Средна стойност 2002-2015г	Стандартно отклонение	Коефициент на вариация (%)
Царевица	1346706.7	2262615,4	68,0	1804661,1	745871,4	41,3
Ориз	25588,9	56192,6	119,6	40890,7	17356,3	42,4
Соя	901,4	624,0	-30,8	762,7	1041,9	136,6
Пшеница	3412949,7	4692665,9	37,5	4052807,8	1009227,7	24,9

Линейните модели на трендовете на добивите от пшеница и ориз на национално ниво са с доказана адекватност. Нарастването при пшеницата е 117,4 kg/ha (3,31%), а за ориза е 85 kg/ha (1,75%). За добивите от царевица и соя са изследвани линейни, експоненциални и полиномни от втора и трета степен модели на тренд, но адекватност за нито един от тях не е доказана. Прогнозата на добива от пшеница на национално ниво и на два от географските региони е показана на Фиг. 3.6.

За периода от 2002г. до 2015г. се наблюдават относително големи колебания на средния добив от царевица, което се дължи на влиянието на различни фактори, невключени в изследването (Фиг 3.7). Заради тези големи вариации изследваният модел не е адекватен. В този случай за изглаждане и прогнозиране се използва метода на претеглената пълзяща средна. Претеглената пълзяща средна на всеки 5 години е приложена заради по-малкото отклонение между реалните и усреднените данни. Получените резултати показват

положителна тенденция за добива от царевица в България (Фиг. 3.7).
Усреднените стойности за добива на царевица за периода от 2007г. до 2011г. са $3743 \div 4638$ kg/ha, докато за периода от 2012г. до 2015г. са $4858 \div 6094$ kg/ha.



Фиг. 3.6 Линеен тренд на добив на пшеница

Резултатите от направените статистики на динамичните редове на добива от пшеница, ориз, соя и царевица показват, че на национално и регионално ниво, за периодите 2002÷2008г. и 2009÷2015г. има подобрение през втория период.



Фиг. 3.7. Добив на царевица – 5 и 3 годишни подвижни средни стойности

Положителното нарастване на средния добив от пшеница са страната (3,31%) и по региони (1,46%÷5,11%) за периода 2002÷2015г. е от особена важност за България. Получените резултати от проучването показват положителни тенденции за България в производството на четирите ключови глобални култури, особено при традиционните за страната пшеница и царевица.

Направени са още изследвания и изводи за структурните характеристики на лозарството в България, Гърция, Румъния и ЕС и за производството на ябълки в трите балкански страни, членки на ЕС.

III. ИЗВОДИ И ПРЕПОРЪКИ

Проектирана и създадена е многомерна РС-база данни Crop_ProductionBGEC за растениевъдното производство в България и две страни членки на Европейския съюз – Гърция и Румъния. Базата данни поддържа атрибути за статистически показател, култура, страна и година. Стойностите, определени от тези атрибути са основни и минимално необходими за извършване на статистически анализи със значими за потребителите резултати.

Със своята компактност и специфична тематична насоченост към растениевъдното производство Crop_ProductionBGEC замества на локално ниво големите статистически бази данни и предлага графичен потребителски интерфейс вместо работата с по-комплексните web-базирани статистически бази данни. На основа на тази компактност базата данни е използвана като информационно ядро за разработката на интерфейсни решения за потребителя, които не се предлагат от web-интерфейса на комплексните web бази данни. Разработено е автоматизирано софтуерно решение за обновяване на базата с данни от web статистическите БД и от други информационни източници. По този начин е осигурен дълъг технологичен живот на базата данни Crop_ProductionBGEC.

Създаден е приложен софтуер, който извлича от създадената компактна база данни динамични редове, предоставя интерфейсни инструменти за тяхното управление и ги експортира за статистическа обработка от програмна среда на

езика R. В създаденото приложение са използвани процедури и модули на програмния език R за статистическа обработка.

По този начин е реализирана интегрирана система от три компонента:

- Локална и компактна база данни с тематична насоченост към растениевъдното производство в България и редица други страни;
- Програмната среда на езика R с неговите процедури и модули;
- Софтуерно приложение с потребителски интерфейс за интегриране на другите два компонента.

Направени са анализи на данни както на основата на програмната среда на езика R, така и с директен експорт от базата и обработка със статистическия софтуер SPSS. Изследвани са динамични редове на ниво България, ниво региони в България, ниво страни членки на Европейският съюз от Балканския полуостров и на ниво всички страни членки на Европейският съюз. Анализирани са динамични редове за производството на отделни култури, както и продукти от нея. Направени са изводи за характера на развитие на растениевъдното производство като са открити и връзки със събития, като промяна на нормативната уредба, произтичаща от разрастването на Европейския съюз и събития свързани с политическият преход.

Разработеното решение (многомерна БД и софтуер) може да се приложи в други отрасли на аграрния сектор: животновъдство, рибарство, горско стопанство. Важни аргументи подкрепящи това твърдение са:

1. Информационните източници FAO, НСИ, Министерство на земеделието и начинът на представяне на пространствените и времеви данни в тях са същите, както и за растениевъдство;
2. Целесъобразност в организирането на данните в динамични редове и тяхното йерархично структуриране в локална база данни, което е важна предпоставка за достъп, извличане и разработване на софтуер за обработка на БД.
3. Необходимост от задълбочени математико статистичеки анализи на данните в динамични редове за споменатите отрасли с оглед намирането на тенденции, зависимости, съотношения, различия в показателите и т.н.

IV. ПРИНОСИ НА ДИСЕРТАЦИОННИЯ ТРУД

Приносите и оригиналните елементи на дисертационната работа имат научно-приложен характер и биха могли да се обобщят по следния начин:

1. Разработеният модел на базата данни е на основата на йерархично свързване на съвкупности от динамични редове. Йерархиите съответстват както на представата на потребителите за пространствения характер на изследваните обекти (страна, регион), така и статистическите показатели (площ, количество произведена продукция);

2. Създаден е потребителски ориентиран софтуер за управление на динамични редове от базата данни (достъп, извличане, наблюдаване), както и за математико-статистически анализ чрез използването на пакети и модули от езика R.

3. Предложен е подход и са разработени софтуерни техники за автоматизирано актуализиране на БД, което гарантира нейния технологичен живот. Този софтуер е важна предпоставка за разширяването на базата с данни за растениевъдното производство за други страни.

4. Направен е анализ на динамични редове в за производството на отделни култури, в резултат на който са открити специфични особености и трендове. Като оригинален елемент може да се посочи едновременния анализ на динамични данни за производството на суровина (грозде) и краен хранителен продукт (вино).

V. НАУЧНИ ПУБЛИКАЦИИ

1. Onkov K., D. Dimova, N. Tsaikin, D. Stoyanova (2017), **Estimating multidimensional time series on key global crops production in Bulgaria**, International Multidisciplinary Scientific GeoConference Surveying Geology and Mining Ecology Management, SGEM 17(21), p. 295-302.

Забележка: А. Статията е реферирана в SCOPUS:

<https://www.scopus.com/record/display.uri?eid=2-s2.0-85032472978&origin=resultslist&sort=plf-f&src=s&st1=Onkov&st2=&nlo=1&nlr=20&nls=count-f&sid=9cf58722a942cc44065ceb27d84bc4fe&sot=anl&sdt=aut&sl=30&s=AU-ID%28%22Onkov%2c+K.%22+26434295600%29&relpos=3&citeCnt=0&searchTerm=>

Б. SJR = 0.211 (за 2017г.)

2. Stoyanova D., N. Tsaikin. (2017) **Analysis of the dynamics of sunflower production in Bulgaria by statistical regions**. Agricultural Sciences, v. IX, Issue 21.
3. Tsaikin N., M. Zarkova, D. Stoyanova, K. Onkov (2010). **Analysis of apple production of EU Balkan states based on PC time series database**, Université de Picardie Jules Verne, Centre de Recherche sur les Institutions, l'Industrie et les Systèmes Economiques d'Amiens (CRIISEA), France. Proceedings of the International Conference on International business, Thessaloniki.

Използвана литература:

- Agrawal, R., Gupta, A., & Sarawagi, S. (1997). Modeling multidimensional databases. In IEEE Proceedings of 13th International Conference on Data Engineering (pp. 232-243), Mumbai, India: Indian Institute of Technology Bombay.
- Alwan, L. C. (2000). Statistical process analysis. Book. New York: McGraw-Hill/Irwin.
- Chaudhuri, S., Dayal, U., & Ganti, V. (2001). Database technology for decision support systems. Computer, 34(12), 48-55.
- Elmasri, R., Navathe, S. (2010). Fundamentals of Database Systems (sixth edition), Addison-Wesley, 2010

- Feng, Q., Miao, D., & Cheng, Y. (2010). Hierarchical decision rules mining. *Journal of Expert systems with applications*, 37(3), 2081-2091.
- Gandhi N., L. Armstrong, (2016) Rice crop yield forecasting of tropical wet and dry climatic zone of India using data mining techniques, 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, 2016, pp. 357-363
- Glymour C. (1997) *Statistical Themes and Lessons for Data Mining*, Data Mining and Knowledge Discovery 1, 11–28, Kluwer Academic Publishers.
- Hearst M. (1998) Trends & Controversies: Information integration. *IEEE Intelligent Systems* 13(5), 1998, pp 12-24.
- Javier C. (2004) Fast Nearest Neighbours Search in High-Dimensional Large-Sized Databases, ITI Seminario, 2004. <http://rhlt.iti.es/seminars/SEMINARS/Cano04a.pdf>
- Jensen, C. S., Pedersen, T. B., & Thomsen, C. (2010). Multidimensional databases and data warehousing. *Synthesis Lectures on Data Management*, 2(1), 1-111.
- Lenzerini, M. (2002) Data integration: a theoretical perspective, *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2002, pp 233-246.
- Onkov K, (2016) Classification algorithm for knowledge extraction from multidimensional time series database on crop production, *Sofia SGEM Vol. 1*, 125-131. (2016)
- Ripley B. and D. Murdoch (2009). *R for Windows FAQ*. R Development Core Team
- Shimoda S, Hiromitsu Kanno, Tomoyoshi Hirota, (2018) Time series analysis of temperature and rainfall-based weather aggregation reveals significant correlations between climate turning points and potato (*Solanum tuberosum* L) yield trends in Japan, *Agricultural and Forest Meteorology*, Volume 263, 2018, Pages 147-155
- Tegos G , K Onkov, D Stoyanova, (2016) Estimation and Analysis of Fish Catches by Category Based on Multidimensional Time Series Database on Sea Fishery in Greece

- Tegos, G. and Onkov, K. (2015) Time series database analysis on fishery in Greece. Book chapter 12 in "Progressive engineering practices in marine resource management", IGI Global, USA, 372-398.
- Tegos, G., Onkov, K. (2007) Development and application of queries to PC fishery database of Greece for time series analysis. Proceedings of 3rd International Conference on Agricultural Economics, Rural development and Informatics, vol.1, Hungary, March 20-21 2007, pp 153-159
- Trnka M., J. E. Olesen, K. C. Kersebaum, (2016) Changing regional weather-crop yield relationships across Europe between 1901 and 2012, Inter-Research v70 p195-214
- Илиева, С., Лилов, В. и Манова, И. (2010), “Подходи и методи за реализация на софтуерни системи”, Университетско издателство „Св. Климент Охридски”, София.
- Научен проект 07.11. (2011) на тема: „Алгоритми и софтуер за анализиране на пространствени динамични редове за растениевъдното производство на България и страни от ЕС”, Научно-изследователски център на Аграрен Университет, Пловдив.
- Научен проект 09.09. (2009) на тема: „Структуриране и анализ на динамични статистически редове на растениевъдното производство в България и страни от Европейския съюз”, Научно-изследователски център на Аграрен Университет, Пловдив.
- Онков К. (2011) Информационна среда за йерархично структуриране и анализ на многомерни динамични редове. Академично издателство на Аграрен Университет - Пловдив 2011.

Abstract

International and government institutions store big amount of data in databases from various fields, including crop production. Typically, such databases are presented in a Web environment with a graphical-user interface to access the data, but only rarely provide online processing that is insufficient for in-depth analysis and conclusions.

The subject of the dissertation is to offer an integrated solution consisting of time series database on crop production in Bulgaria and some EU countries and application software for database management and analysis. The use of object oriented programming and R language software procedures are into consideration. A dispersion analysis, cluster analysis and algorithm for calculating the ecological valence have been used in the mathematical and statistical analyses on time series.

The aim of the dissertation is to create an information environment helping specialists to make analytical assessments for the crop-production.

The following tasks have been placed:

- To design and create a multidimensional database with time series for crop production in Bulgaria and European Union countries;
- To create software to extract time series from the database and manage them via a graphical user interface;
- To develop software for mathematical and statistical analysis of time series, including the use of modules and procedures from the programming language R. Application of known software statistical packages with graphical user interface such as SPSS are also in consideration;
- To analyze time series for the crop production of important crops with a view to studying regularities and finding trends.

The data sources for the creation of multidimensional time series database Crop_ProductionBGEC are FAOSTAT (Statistical Institution of the Food and Agriculture Organization), the National Statistical Institute and the Ministry of Agriculture and Food in Bulgaria. Crop_ProductionBGEC contains time series on the production quantity and harvested areas in Bulgaria and two EU-countries (Greece and Romania) for 82 crops for the period 1961-2017. It has been designed as a multi-dimensional hypercube of four dimensions with a hierarchy by countries and regions. A software has been developed that integrates data from large databases, applied statistical software and software for automation the information flow from institutional sources to statistical analysis.

The developed software application Savatap is designed for database management (extraction, manipulation, processing etc.) regarding time series on crop production in Bulgaria and the other Balkan countries of the European Union. The whole process of statistical processing is automated from the updating of time series from web-based databases to output of the results from a selected method. Savatap

provides a graphical user interface. After selecting a statistical method, the R language program is called through a script file.

Data analyses have been made on the basis of the R language software as well as direct export from the base and processing with the SPSS statistical software. The yields of sunflower for Bulgaria by regions, the structural characteristics of viticulture in Bulgaria, Greece, Romania and the EU have been studied; the production of apples in the three Balkan countries; the links between the production, import and export of grapes and wine as relation between raw material and product; the production in Bulgaria of the four key crops: wheat, soybeans, rice and maize. Conclusions have been made on the nature of crop production development as well as links to factors and events such as the impact of pre-accession agricultural funds and the European SAPARD and PHARE programs and changes in the regulatory framework resulting from the expansion of the European Union and events related to the political transitions.

The contributions of the dissertation are as follows:

- The developed database model is based on a hierarchical relations among the sets of time series;
- User-oriented database management software has been developed, as well as mathematical and statistical analysis using packets and modules from the R language;
- An approach has been proposed and software techniques have been developed for automated updating of the database. It guaranties the database maintenance and evolution;
- An analysis of the time series on the production of economically important crops has been made, resulting in specific features and trends.