Академия наук УССР
Морской гидрофизический институт

На правах рукописи
УДК 551.463:51

Мамуду Кейта

Контроль качества и полноты баз океанографических данных в интегрированных системах обработки информации на примере Гвинейского Научно-исследовательского центра

Специальность 04.00.22 - геофизика

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Севастополь - 1990

Работа выполнена в Морском Гидрофизическом Институте АН УССР и в Симферопольском государственном университете им. М.В.Фрунзе

Научные руководители: доктор технических наук В.А.Гайский
кандидат техническик наук С.Ф.Толкачев

Официальные оппоненты: доктор ф.-м. наук В.А.Иванов
кандидат технических наук А.П.Уриков

Ведущая организация: Всесоюзный научно-исследовательский институт гидрометеологической информации - Мировой центр данных.

Защита диссертации состоится "19" октября 1990 г. в "15" часов на заседании специализированного Совета Д 016.01.01 при Морском гидрофизическом институте АН УССР (335000, г.Севастополь, ул.Капитанская, 2)

С диссертацией можно ознакомиться в читальном зале НТБ МГИ АН УССР.

Автореферат раз

Ученый секретар
Специализирован
кандидат физико

Общая характеристика работы

Актуальность темы.

Банки океанографических данных существуют и развиваются уже более 20 лет. В последние годы, с появлением персональных ЭВМ появилась возможность создавать специализированные банки для многих пользователей на основе ограниченных по объему выборок из универсальных баз данных. В этой связи особую остроту приобрел вопрос: каким образом оценивать качество и полноту баз океанографических данных? Без ответа на него у пользователя нет уверенности в том, что формируемая ограниченная по объему база данных позволит ему удовлетворительно решить какую-либо конкретную задачу. С другой стороны, ответ на такой вопрос позволяет оценить пригодность для тех или иных целей полных баз данных национальных океанографических центров и спланировать рациональную стратегию их пополнения за счет обмена данными с другими центрами или организации экспедиций. До настоящей работы достаточно удовлетворительного ответа на этот вопрос не было.

Целью работы является разработка метода и программных средств для контроля качества и полноты баз океанографических данных и совершенствования системы управления базами данных в распределенных вычислительных системах с разнородными персональными ЭВМ.

Основные задачи исследования:

1. Провести анализ потоков и объемов экспериментальной информации об окружающей среде различных направлений исследований и разработать концепцию построения банка данных Гвинейского научно-исследовательского центра (ГНИЦ).

2. Используя теорию дискретизации случайных процессов и полей, исследовать возможность оценки качества и полноты баз океанографических данных через оценку точности представления дискретной пространственно-временной решеткой в зависимости от параметров решетки, типовых статистических характеристик процессов и полей.

3. Разработать методику и программное обеспечение оценки качества и полноты баз океанографических данных.

4. Провести анализ базы океанографических данных гвинейского научно-

исследовательского центра, оценить её качество и полноту, разработать рациональную модель банка данных в интегрированных вычислительных системах.

5. Создать специальное программное обеспечение для хранения, обработки и визуализации информации в распределенной системе персональных ЭВМ типа IBM/PC и Macintosh.

Научная новизна.

На основании впервые выполненного анализа текущих и перспективных информационных потоков различных направлений исследований Гвинейского научно-исследовательского центра разработана концепция построения регионального интегрированного банка данных ГНИЦ.

На основе применения теории дискретизации случайных процессов и полей впервые проведена математическая формализация задачи оценки качества и полноты баз данных о процессах и полях окружащей среды. Задача сведена к оценке точности представления случайного процесса или поля в заданном пространственно-временном окне (что является формализацией цели потребителя) дискретной пространственно-временной решетки, значения поля в узлах которой составляют базу данных, в зависимости от размеров ячейки и объема решетки и априорных статистических характеристик процесса или поля.

Получены новые формулы для оценки погрешностей непосредственной дискретизации случайных процессов со степенными спектрами, характерными для многих временных процессов и сечений полей окружащей среды.

Проведен количественный анализ погрешностей дискретизации по этим формулам. Аналогичные оценки впервые получены для степенных процессов, прошедших инерционное звено, что соответствует практическим измерениям инерционными приборами.

Впервые проведен анализ на качество и полноту различных баз данных ГНИЦ и получены количественные оценки, позволяющие сделать выводы об ограниченных возможностях исследований на базах данных, формируемых по съемкам типового Гвинейского полигона и необходимости расширения полигона в 5-10 раз.

С целью эффективного использования различных возможностей разнотипных микро-ЭВМ в распределенной системе обработки данных впервые создано специальное программное обеспечение, включенное в систему управление базами данных.

Практическая ценность работы.

Результаты работы непосредственно касаются совершенствования банка

данных об окружающей среде ГНИЦ, его программного и аппаратного обеспечения. Вместе с тем они могут использоваться для оценки качества и полноты других баз данных о процессах и полях, рационального планирования полигонных съемок, а также при организации вычислительных сетей из персональных ЭВМ типа IBM/PC и Macintosh.

Апробация работы.

Основные результаты работы докладывались на семинаре "Автоматизированные системы сбора и обработки гидрофизической информации" (Севастополь, 1987), республиканском семинаре "Интерфейсные средства систем автоматизации научных исследований" (Севастополь,1988), на семинарах отдела автоматизации океанографических исследований морского гидрофизического института АН УССР (Севастополь,1988-1990 гг) на семинаре кафедры прикладной математики Симферопольского Государственного Университета им.М.В.Фрунзе (Симферополь,1990г).

Публикации. По теме диссертации опубликована одна работа.

Личный вклад автора. Автор участвовал в разработке концепции построения банка данных ГНИЦ, получены теоретические результаты, и самостоятельно проведен анализ погрешностей дискретизации случайных процессов, прошедших инерционное звено, созданы вычислительные и управляющие программы, выполнены вычислительные эксперименты.

Структура и объем работы. Диссертация состоит из введения, четырех глав, заключения, приложения, списка использованной литературы из 73 наименований. Общий объем диссертации 147 страниц машинописного текста, в том числе, 27 иллюстрации и 13 таблиц.

Содержание работы.

Во введении обоснована актуальность темы, сформулированы цель и задачи исследования, кратко изложено содержание диссертации, приведены основные результаты.

В первой главе рассматриваются вопросы построения банков данных об окружающей среде на примере банка данных ГНИЦ. Приведены требования к таким банкам и выполнен анализ принципов их функционирования. Показано, что наибольшей отдачи при существенных затратах в получении и использовании информации об окружающей среде можно достичь лишь в том случае, если все виды работ по сбору, постоянному хранению, обработке и доведению до пользователей океонографических данных осуществляются в рамках единой системы. Проанализированы основные показатели, в зависимости от которых обеспечивается эффективность функционирования такой системы. Показано, что показатели эффективности в научно-

технических системах имеют противоречивую природу. Например, выполнение условия наибольшей полноты собираемых материалов наблюдений и разнообразия баз данных требует увеличения объема и сложности программных средств их обработки, что ведет к снижению оперативности решения запросов пользователей. С целью преодоления таких трудностей, рассматривается возможность использования многомашинных средств.

Рассматривается роль вычислительного эксперимента при проведении научных исследований, пути решения вопросов повышения уровня управления информацией, хранимой и обрабатываемой в вычислительной системе. Одним из таких путей является создание систем управления данными, основанных на наиболее эффективных способах организации, идентификации, классификации, запоминания и выборки данных, интеграция на основе системы управления данными (СУД) информационных баз вычислительных центров, внедрение единой технологии организации хранения данных и обращения к ним. Все в совокупности предъявляет к системам управления данными ряд требований, без учёта которых немыслимо удовлетворительное решение указанной выше проблемы.

Приведены оценки объемов баз данных и потоков данных, изложены источники их получения и представлена их структура (Рис.1,2).

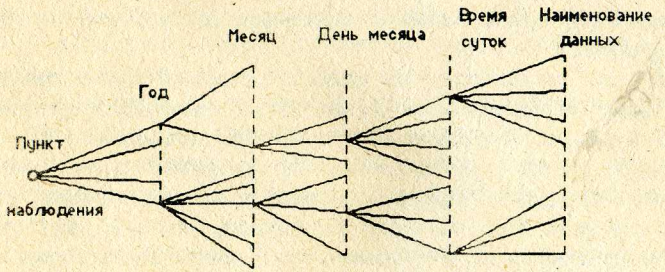

Рис. 1 Структура метеоданных

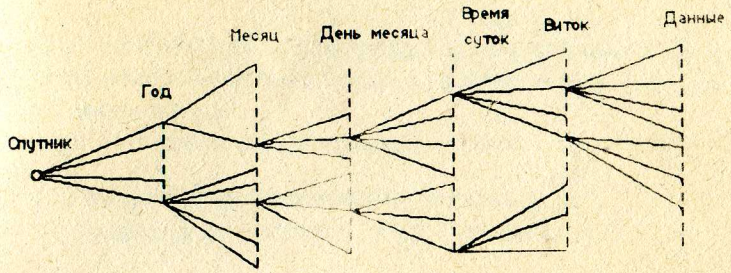

Рис. 2 Структура спутниковых данных

Вторая глава посвящена проблеме дискретизации гидрофизических процессов и полей, которые являюся объектами исследований. Пространственно-временные физические поля, непрерывны как в пространстве, так и во времени и возможности реализации непрерывных n-размерных сечений поля (n>3) весма ограничены. Поэтому в основе современных методов гидрофизических исследований лежат дискретные представления. Это обусловлено как тем, что теоретически непрерывный объект обычно можно отобразить с определенной прогрешностью дискретным представлением, так и технической и (или) экономической реализуемостью систем и модели.

Оценивается потенциальная точность дискретного представления случайных процессов со степенными спектрами вида

$$S_m(f) = c|f|^{-m} \qquad (1)$$

где $c = const$, $m = [\,1,\ 3/2,\ 5/3,\ 2,\ 3,\ 4,\ 5\,]$

Оценка относительной погрешности δ_m дискретизации производится по выражению:

$$\delta_m^2 = \frac{\varepsilon_m^2}{\sigma_m^2} \qquad (2)$$

где ε_m^2 – оценка погрешности, которая вычисляется по формуле:

$$\varepsilon_m^2 = \frac{2}{\tau_0^2} \sum_{\substack{i=-\infty \\ i\neq 0}}^{\infty} \int_{f_1}^{f_2} S_m\left(f - \frac{i}{\tau_0}\right) df \qquad (3)$$

где τ_0 – интервал дискретизации в полосе $[f_1, f_2]$;

σ_m^2 – дисперсия дискретизированного процесса в полосе $[f_1, f_2]$ вычисляется по выражению:

$$\sigma_m^2 = \frac{2}{\tau_0^2} \int_{f_1}^{f_2} S_m(f)\, df \qquad (4)$$

например, при $m = 1,2$

$$\sigma_1^2 = \frac{2c}{\tau_0^2} \ln \frac{f_2}{f_1} \qquad (5)$$

$$\sigma_2^2 = \frac{2c}{\tau_0^2} \frac{(f_2 - f_1)}{f_2 f_1} \qquad (6)$$

используя (2), (3) и (4) получены оценки δ_m ($m = 1 \ldots 5$).
Например,

$$\delta_1^2 \geq \frac{\sum_{i=1}^{\infty} \ln \frac{(i + f_2\tau_0)(i - f_1\tau_0)}{(i + f_1\tau_0)(i - f_2\tau_0)}}{\ln \frac{f_2}{f_1}} \qquad (7)$$

$$\delta_2^2 \geq \frac{\tau_0 f_2 f_1}{f_2 - f_1} \ln \frac{(1 - f_1\tau_0)(1 + f_2\tau_0)}{(1 + f_1\tau_0)(1 - f_2\tau_0)} \qquad (8)$$

Далее, с целью получения удобных для анализа и графического представления зависимостей δ_m, введены две новые переменные α и β, принимающие значения от 0 до 1.

α связана с частотой дискретизации τ_0^{-1} и верхней частотой f_2 в полосе восстанавливаемого процесса, а β – с нижней f_1 и верхней f_2 частотами как

$$\alpha = 2 f_2 / \tau_0^{-1} = 2 f_2 \tau_0 \qquad (9)$$

$$\beta = f_1 f_2^{-1} \qquad (10)$$

Таким образом, получены оценки погрешности δ_m в зависимости от двух переменных α и β, принимающих значения в диапазоне от 0 до 1 для любого сочетания частот дискретизации и границ полосы восстановления исследуемого процесса так, как это показано на рис. 3-7. Используя полученные выше соотношения, полученные для δ_m ($m = 1 \ldots 5$)

$$\delta_1^2 \geq \sum_{i=1}^{\infty} \ln \frac{(2i + \alpha\beta)(2i - \alpha)}{(2i + \alpha)(2i - \alpha\beta)} \Big/ \ln \beta \qquad (11)$$

$$\delta_2^2 = \frac{\alpha\beta}{2(1 - \beta)} \ln \frac{(2 - \alpha\beta)(2 + \alpha)}{(2 + \alpha\beta)(2 - \alpha)} \qquad (12)$$

$$\delta_3^2 = \frac{\alpha^3\beta^2}{1-\beta^2}\left(\frac{1}{4-\alpha^2} - \frac{\beta}{4-\alpha^2\beta^2}\right) \qquad (13)$$

$$\delta_4^2 = \frac{2\beta^3\alpha^4}{1-\beta^3}\left[\frac{1}{(\alpha^2-4)^2} - \frac{\beta}{(\alpha^2\beta^2-4)^2}\right] \qquad (14)$$

$$\delta_5^2 = \frac{\alpha^5\beta^4}{3(1-\beta^4)}\left[\frac{12+\alpha^2}{(4-\alpha^2)^3} - \frac{\beta(12+\alpha^2\beta^2)}{(4-\alpha^2\beta^2)^3}\right] \qquad (15)$$

Выражения (11) - (15) можно использовать для оценки потенциальной точности реальных баз данных. Показано, что если для конкретного эксперимента с базой данных некоторого случайного процесса с функцией спектральной плотности $S_m(f)$ вида (1), соотношения частот дискретизации τ_0^{-1} верхней f_2 и нижней f_1 частот выделяемого процесса соотвествуют согласно (9), (10) значениям $\alpha = 0.01$ и $\beta = 0.01$, то соотвествующая точка Э на графиках рис. 5-7 дает следующие оценки для δ_m

$$\delta_1 > 10^{-1},\ \delta_2 > 10^{-2},\ \delta_3 > 10^{-3},\ \delta_4 > 10^{-4},\ \delta_5 < 10^{-4} \qquad (16)$$

Реальная точность данного эксперимента с этой базой данных не может быть выше этих оценок.

В реальных данных часто представляются отсчёты процессов, прошедших инерционное звено (данные измерительного прибора), например с постоянной инерции $T_и$.

Функция спектральной плотности таких процессов описывается выражением вида:

$$S_{1m} = \frac{c|f|^{-m}}{1 + 4\pi^2 T_и f^2} \qquad (17)$$

В работе получены выражения для оценки погрешности δ_m по формуле (2) при $S_m(f) = S_{1m}(f)$ по формуле (17) для $m = 1...5$.

Проведенный выше анализ устанавливает связь погрешности восстановления процесса (поля) с интервалом дискретизации или качества базы данных с размером ячейки пространственно-временной решетки.

Связь погрешности δ_m восстановления процесса (поля) с длиной реализации (качества базы данных с её объемом N) для $f_1 = 1/T$, $f_2 = 1/2\tau_0$ и следовательно, $N = T\tau_0^{-1}$, $\alpha = 1$, $\beta = 2\tau_0 T^{-1} = 2N^{-1}$ устанавливается выражениями

$$\delta_1 \simeq (\ln N)^{-1}$$

$$\delta_2 \simeq N^{-1/2} \qquad (18)$$

$$\delta_3 \simeq N^{-1}$$

$$\delta_4 \simeq N^{-3/2}$$

графики которых показаны на рис. 8.

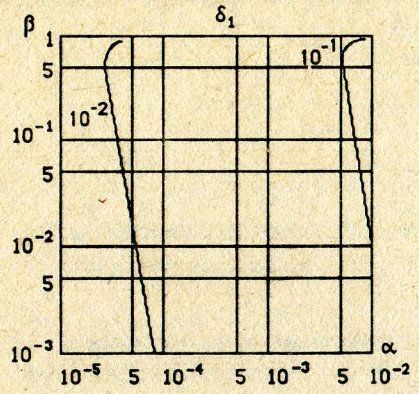

Рис. 3 Изолинии нижних оценок погрешности дискретизации процессов с показателями степени спадания спектра m = 1 в зависимости от α и β.

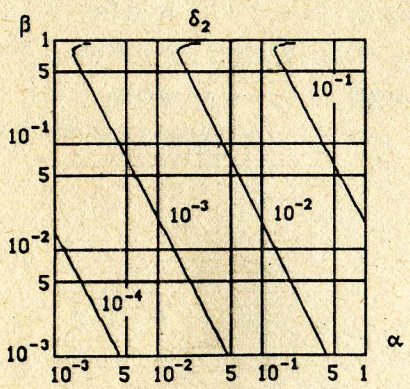

Рис. 4 Изолинии нижних оценок погрешности дискретизации процессов с показателями степени спадания спектра m = 2 в зависимости от α и β.

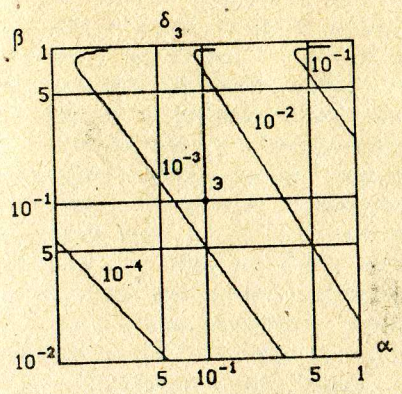

Рис. 5 Изолинии нижних оценок погрешности дискретизации процессов с показателями степени спадания спектра m = 3 в зависимости от α и β.

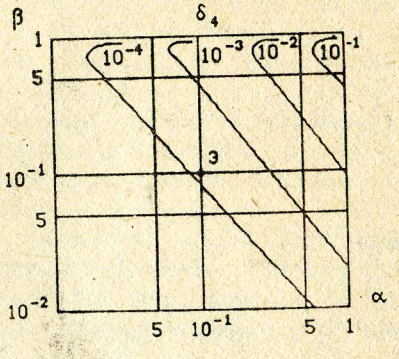

Рис. 6 Изолинии нижних оценок погрешности дискретизации процессов с показателями степени спадания спектра m = 4 в зависимости от α и β.

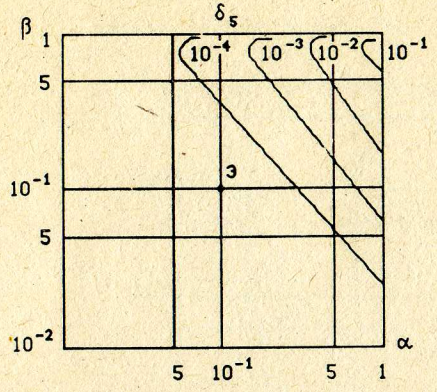

Рис. 7 Изолинии нижних оценок погрешности дискретизации процессов с показателями степени спадания спектра m = 5 в зависимости от α и β.

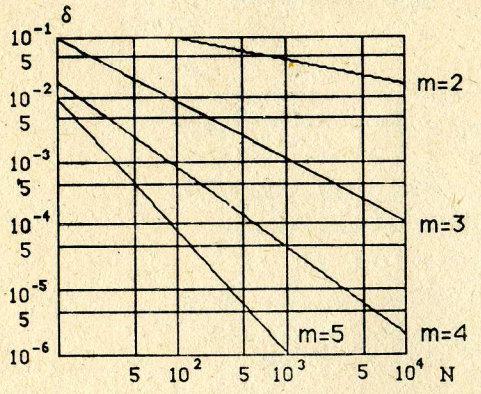

Рис.8 Зависимость нижней оценки δ погрешности дискретизации процессов с показателем степени спадания спектра m = 2 ... 5 от числа N членов ряда.

В третьей главе работы проведены анализ статистических характеристик океанографических процессов и полей Гвинейского региона и оценка качества и полноты баз данных ГНИЦ. При этом использованы данные экспедиционных исследований научно-исследовательских судов МГИ АН УССР, результаты экспедиций, результаты исследований в ГНИЦ.

Установлено, что спектральные характеристики временных процессов и пространственных сечений гидрометеорологических полей на гвинейском полигоне в большинстве случаев удовлетворительно оценивается степенными фукциями вида (1). Это позволяет использовать для оценки качества и полноты баз данных ГНИЦ полученные выше теоритические оценки погрешностей дискретного представления.

Банк океанографических данных ГНИЦ в основном сформирован из наблюдений на гвинейском полигоне, стандартная сетка которого содержит 50 - 70 станций через 15 миль, выполняемых в среднем 2 раза в год на протяжении последних 10 - 20 лет.

Анализ показывает, что база данных этого полигона позволяет выделять с погрешностью 20 - 100% случайную составляющую междугодичную и более продолжительной изменчивости при накоплении времени наблюдений 30-50 лет. Все другие составляющие большей изменчивости (сезонная, синоптическая, мезомасштабная), имеющие убывающие спектры менее чем степенной с показателем минус два, не могут быть выделены с погрешностью меньшей 100%. При пространственном интервале дискретизации 15 миль, верхняя граница полосы волновых чисел выделяемых явлений лежит ниже области существования таких волновых процессов как гироскопические волны и, в основном, внутренние гравитационные волны, но захватывает область существования волн Россби. Однако, для обеспечения выделения волн Россби с погрешностью менее 100% необходимо расширение нижней границы полосы волновых чисел или увеличение размеров полигона примерно в 5 - 10 раз.

Четвёртая глава диссертации посвящена вопросам создания распределённой системы обработки океанографических данных в ВЦ ГНИЦ. В данной главе рассмотрены две системы обработки данных: централизованная и распределенная системы. В централизованной системе обработки данных, все имеющиеся базы данных: гидрологических, гидрооптических, гидрохимических данных и др. объединяют в одной общей базе, содержащей общие признаки, такие например, как номер рейса, номер станции, координаты, число и время выполнения станции и т.д. (рис.9).

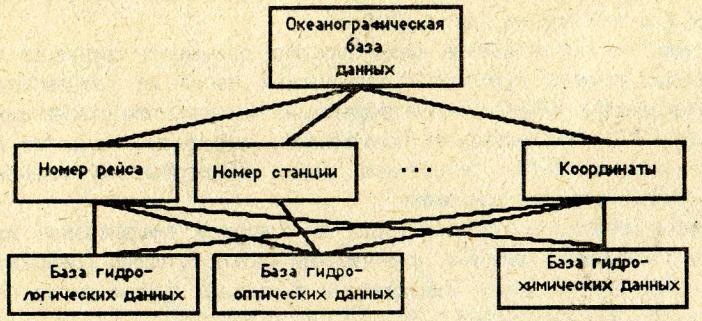

Рис. 9 Структура централизованной базы океанографических данных

Такая структура позволяет избежать избыточности информации и в самом деле использует меньший объём информации. Однако эта централизованная система используется в научном утверждении, которое характеризуется децентрализованной деятельностью. Например метеорологические и гидрохимические данные используются разными потребителями. Такое расхождение между управленческой и функциональной системой вызывает большие трудности в организации оперативного сбора данных, контроля обработки и выдачи информации пользователям. С целью устранения таких трудностей был рассмотрен другой вариант системы обработки данных: распределенную систему обработки. В этой системе, вышеизложенные базы считаются подбазами — глобальной базы океанографических данных (рис.10).

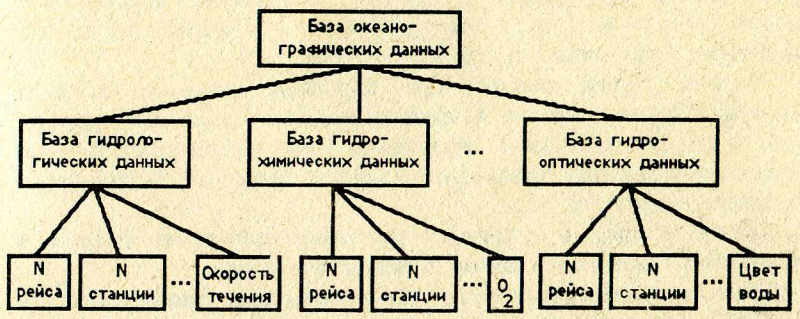

Рис. 10 Структура распределенной базы данных

В такой структуре, каждая подбаза содержает в себя все информационные признаки. Это приводит к определенному увеличению избыточности информации, однако, с другой стороны, позволяет применять проблемно-ориентированые рабочие места, использующие как разнородные технические средства так и программное обеспечение. В главе показано, что этот подход обработки данных в настоящее время обладает рядом преимуществ по сравнению с обычными централизованными вычислительными системами:

- модульность архитектуры и возможность расширения конфигурации;
- повышение надежности функционирования системы в целом;
- сокращение времени обработки независимых запросов за счёт обеспечения высокого уровня параллельной обработки.

В приложении приведены листинги и краткое описание разработанных программ контроля качества и полноты баз данных, а также специальное программное обеспечение СУБД ГНИЦ.

Заключение.

В диссертации формализована, теоретически и практически, с применением современных математических, программных и технических средств, решена актуальная научно-техническая задача по созданию методики и программных средств для контроля качества баз данных об окружающей среде и специального программного обеспечения СУБД в интегрированных системах.

Основные результаты работы состоят в следующем:

1. На основании проведенного анализа потоков и объемов экспериментальной информации различных направлений исследований окружающей среды в ГНИЦ сформулирована концепция построения интегрированного банка данных и требования к СУБД.

2. Впервые теория дискретизации случайных процессов и полей привлечена для оценки качества и полноты баз океанографических данных, получены новые аналитические выражения для оценки погрешностей дискретного представления степенных случайных процессов, в том числе для прошедших инерционное звено.

3. Создана методика и программное обеспечение для оценки качества и полноты баз океанографических данных в диалоговом режиме.

4. Проведен анализ и оценки качества полноты базы океанографических данных ГНИЦ, разработана распределенная модель банка данных с интегрированными интерфейсами между пакетами прикладных программ, СУБД и средствами визуализации.

5. Создано программное обеспечение для распределенной сети ЭВМ (типа IBM/PC, Macintosh и другие), обеспечивающее обработку, хранение и визуализацию информации в языковом среде (C++, Pascal, HyperCard и др.).

Основные научные результаты диссертации опубликованы в работе:

Гайский В.А., М. Кейта, Трубчиков П.Б.
Потенциальная точность дискретного представления случайных процессов со степенными спектрами. В сб.: Экспериментальные исследования тропической Атлантики. (МГИ АН УССР), Севастополь, 1987, Деп. ВИНИТИ 23.12.87, N9035-B87, с. 114-125.

Академия наук Украинской ССР
Морской гидрофизический институт

Мамуду Кейта

КОНТРОЛЬ КАЧЕСТВА И ПОЛНОТЫ БАЗ ОКЕАНОГРАФИЧЕСКИХ ДАННЫХ В ИНТЕГРИРОВАННЫХ СИСТЕМАХ ОБРАБОТКИ ИНФОРМАЦИИ НА ПРИМЕРЕ ГВИНЕЙСКОГО НАУЧНО-ИССЛЕДОВАТЕЛЬСКОГО ЦЕНТРА

А в т о р е ф е р а т

диссертации на соискание учёной степени
кандидата технических наук

Подписано в печать 07.09.90 г.
Формат бумаги 60x90 I/I6 Объем
Заказ 385 Тираж I00 экз.

Отпечатано на ротапринте Морского гидрофизического института АН УССР
335005, Севастополь-5, ул. Ленина, 28